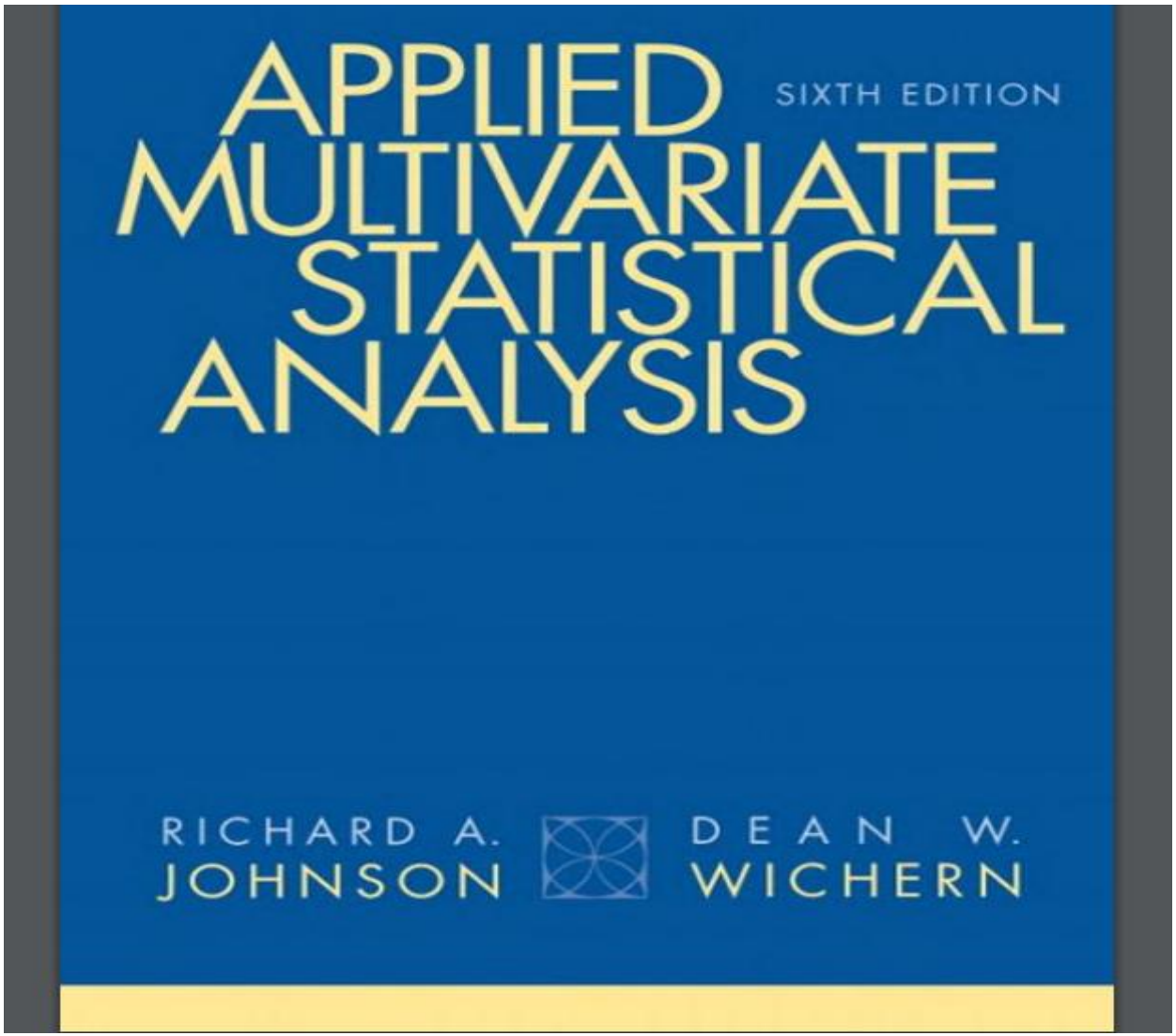
Advanced Multivariate Methods



The Multivariate Normal Distribution or Density Approximation to the "True" Population Distribution and the Central Limit Effect

4.1 Introduction

A generalization of the familiar bell-shaped normal density to several dimensions plays a fundamental role in multivariate analysis. In fact, most of the techniques encountered in this book are based on the assumption that the data were generated from a *multivariate* normal distribution. While real data are never *exactly* multivariate normal, the normal density is often a useful approximation to the "true" population distribution.

One advantage of the multivariate normal distribution stems from the fact that it is mathematically tractable and "nice" results can be obtained. This is frequently not the case for other data-generating distributions. Of course, mathematical attractiveness per se is of little use to the practitioner. It turns out, however, that normal distributions are useful in practice for two reasons: First, the normal distribution serves as a bona fide population model in some instances; second, the sampling distributions of many multivariate statistics are approximately normal, regardless of the form of the parent population, because of a *central limit* effect.

To summarize, many real-world problems fall naturally within the framework of normal theory. The importance of the normal distribution rests on its dual role as both population model for certain natural phenomena and approximate sampling distribution for many statistics.

Probability Density Function i.e. Normal Curve

4.2 The Multivariate Normal Density and Its Properties

The multivariate normal density is a generalization of the univariate normal density to $p \geq 2$ dimensions. Recall that the univariate normal distribution, with mean μ and variance σ^2, has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2} \qquad -\infty < x < \infty \qquad (4\text{-}1)$$

Normal Density with Mean μ and Variance σ^2 : Percentage of Area Under Curve Corresponds to Standard Deviations σ, 2σ & 3σ

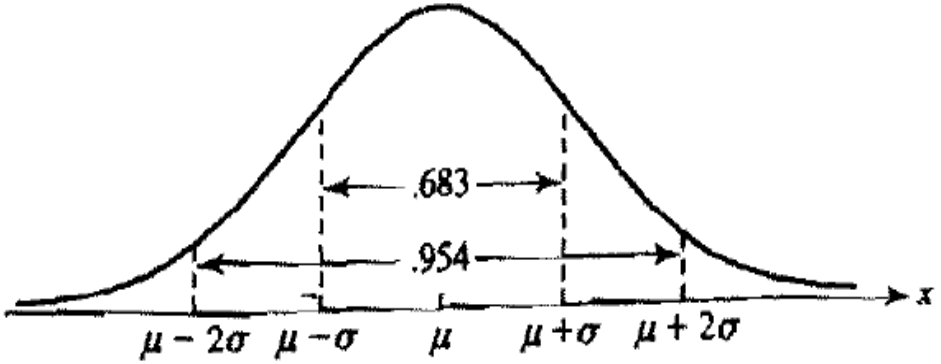
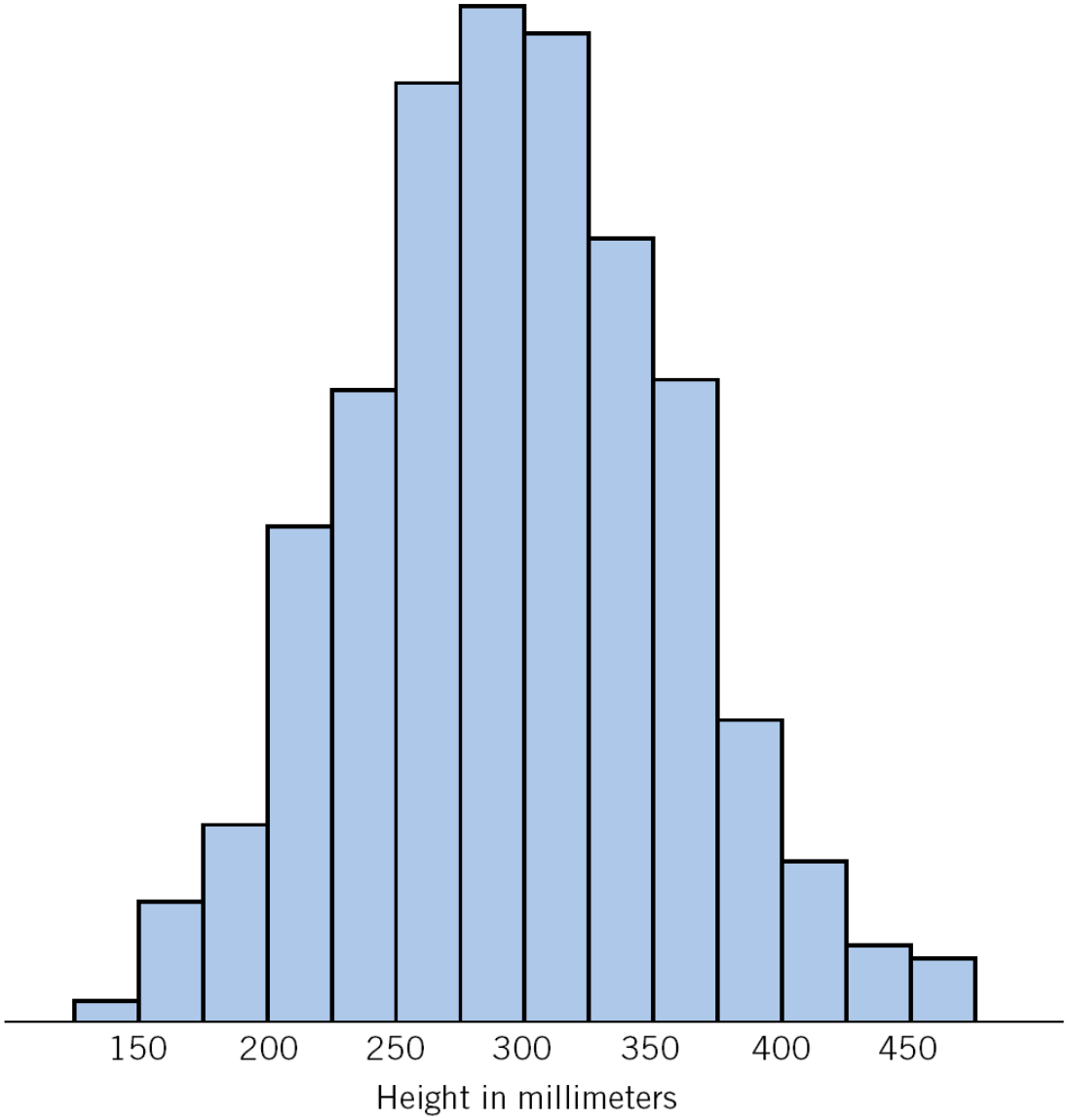


Figure 4.1 A normal density with mean μ and variance σ^2 and selected areas under the curve.

A plot of this function yields the familiar bell-shaped curve shown in Figure 4.1. Also shown in the figure are approximate areas under the curve within ±1 standard deviations and ±2 standard deviations of the mean. These areas represent probabilities, and thus, for the normal random variable X,

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \doteq .68$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \doteq .95$$

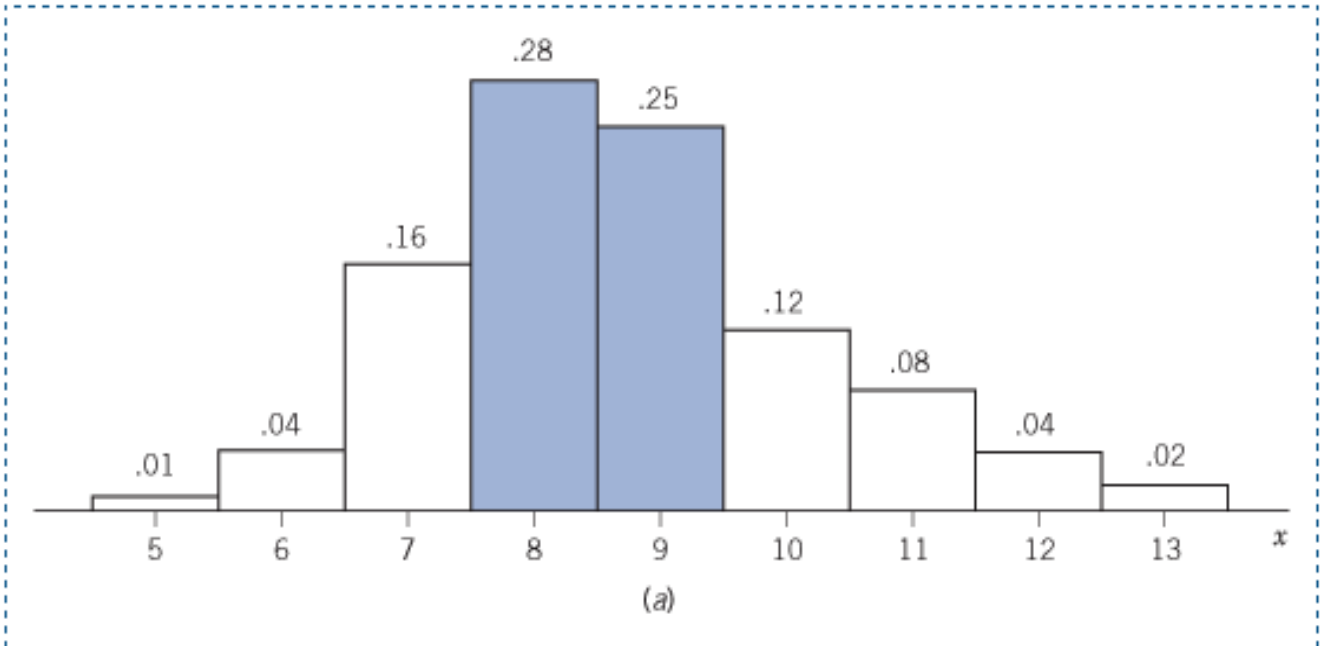


Bell-shaped Distribution of Heights of Red Pine Seedlings, p. 232.

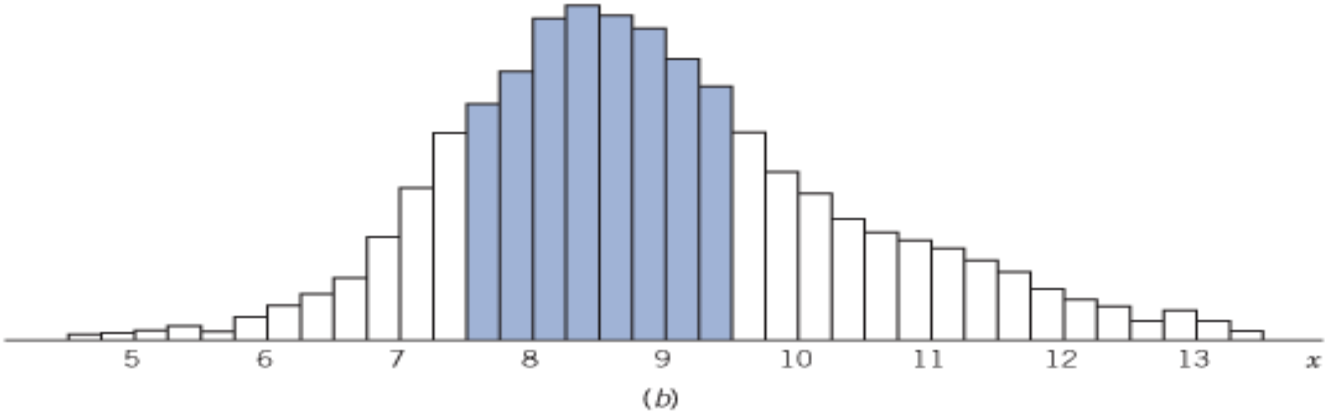
Relative Frequency of n = 100 Birth Weight Newborn Babies

Initially, suppose that the birth weights of 100 babies are recorded, the data grouped in class intervals of 1 pound, and the relative frequency histogram in Figure 1*a* is obtained. Recall that a relative frequency histogram has the following properties:

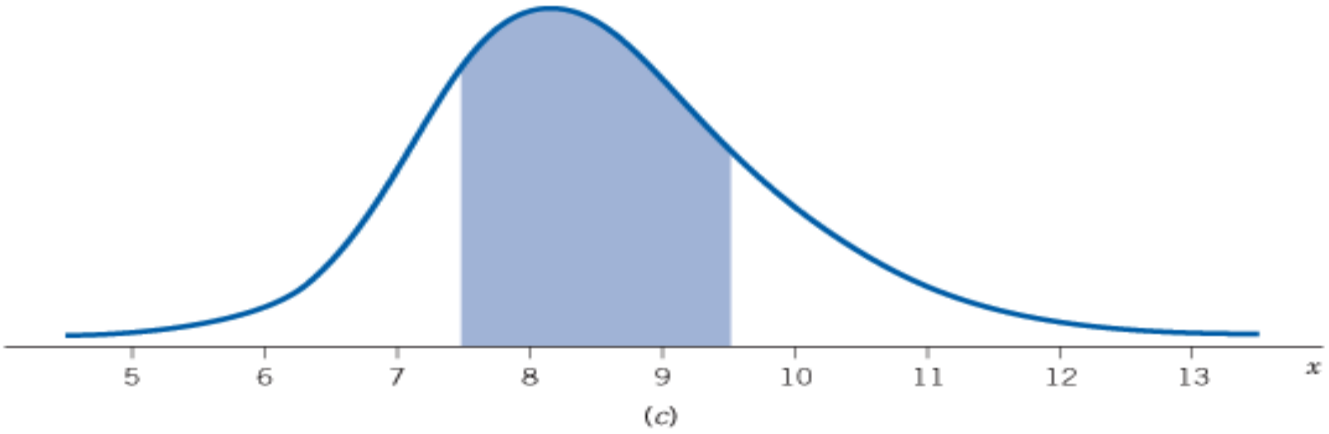
1. The total area under the histogram is 1.
2. For two points a and b such that each is a boundary point of some class, the relative frequency of measurements in the interval a to b is the **area** under the histogram above this interval.



Probability Density Curves – Smoothed with Increasing Sample Size (n)



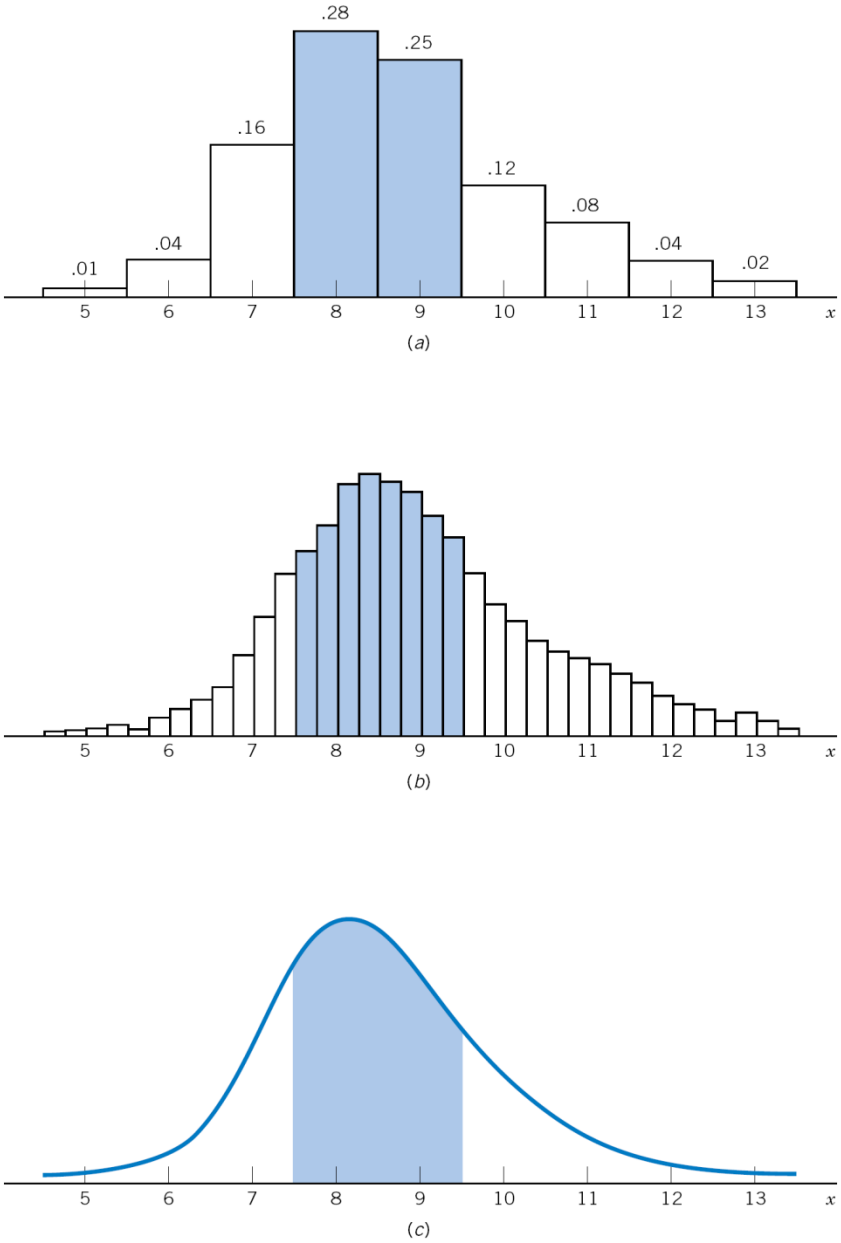
(b)



(c)

Figure 1 Probability density curve viewed as a limiting form of relative

Figure 1,
Probability density curve viewed as a limited form of relative frequency histograms, p. 234



Properties of Probability Density Function or f (x)

The properties 1 and 2 that we stated earlier for a relative frequency histogram are shared by a probability density curve that is, after all, conceived as a limiting smoothed form of a histogram. Also, since a histogram can never protrude below the x axis, we have the further fact that $f(x)$ is nonnegative for all x.

The **probability density function** $f(x)$ describes the distribution of probability for a continuous random variable. It has the properties:

1. The total area under the probability density curve is 1.
2. $P[a \leq X \leq b] =$ area under the probability density curve between a and b.
3. $f(x) \geq 0$ for all x.

Probability Density Function: Interval to the Area or Probability of that Interval

Although the total area is 1, $f(x)$ can be greater than 1.

Unlike the description of a discrete probability distribution, the probability density $f(x)$ for a continuous random variable does not represent the probability that the random variable will exactly equal the value x, or the event $[X = x]$ Instead, a probability density function relates the probability of an interval $[a, b]$ to the area under the curve in a strip over this interval. A single point x, being an interval with a width of 0, supports 0 area, so $P[X = x] = 0$.

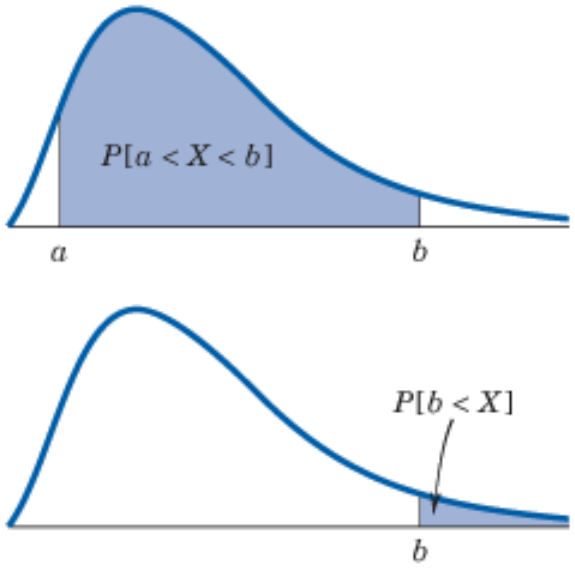
With a continuous random variable, the probability that $X = x$ is **always** 0. It is only meaningful to speak about the probability that X lies in an interval.

Calculating Area Using Probability Distribution Function

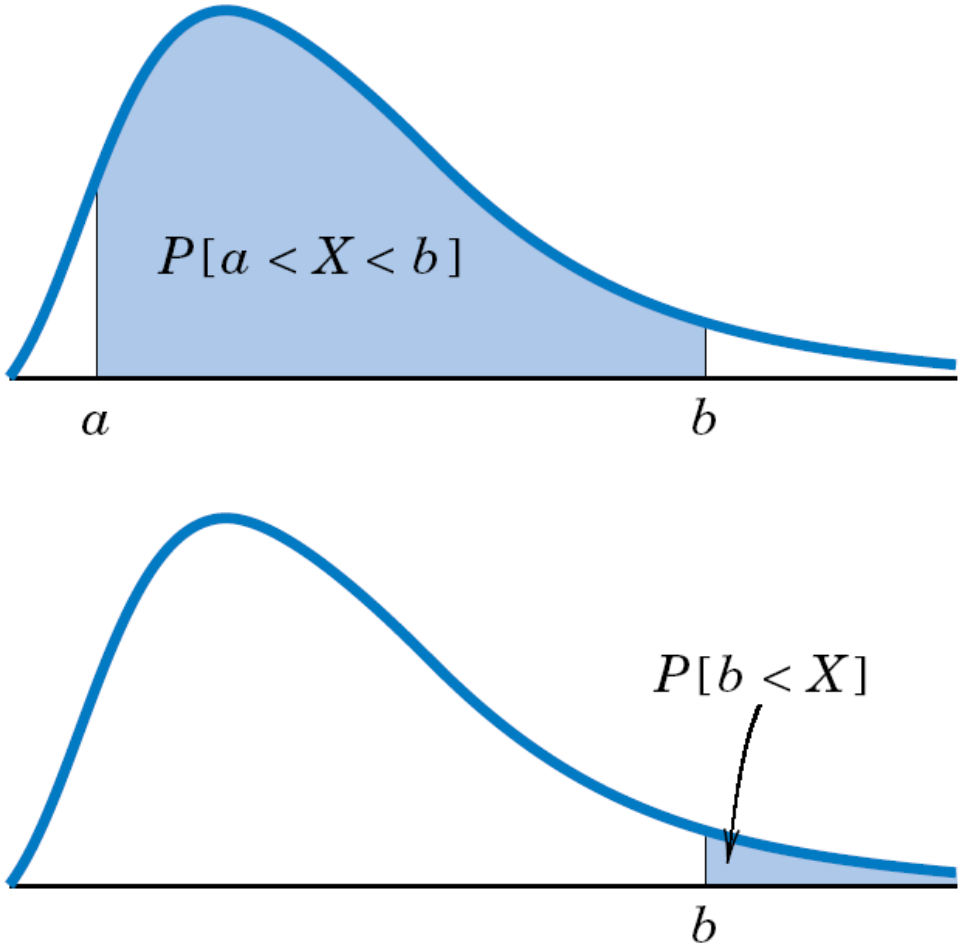
In contrast, these probabilities may not be equal for a discrete distribution.

Fortunately, for important distributions, areas have been extensively tabulated. In most tables, the entire area to the left of each point is tabulated. To obtain the probabilities of other intervals, we must apply the following rules.

$$P[a < X < b] = (\text{Area to left of } b) - (\text{Area to the left of } a)$$



$$P[b < X] = 1 - (\text{Area to left of } b]$$



p. 236.

$$P[b < X] = 1 - (\text{Area to left of } b)$$

Mathematical Form or Curve of Probability Density Function

1.1 SPECIFICATION OF A PROBABILITY MODEL

A probability model for a continuous random variable is specified by giving the mathematical form of the probability density function. If a fairly large number of observations of a continuous random variable are available, we may try to approximate the top of the staircase silhouette of the relative frequency histogram by a mathematical curve.

In the absence of a large data set, we may tentatively assume a reasonable model that may have been suggested by data from a similar source. Of course, any model obtained in this way must be closely scrutinized to verify that it conforms to the data at hand. Section 6 addresses this issue.

Probability Density Function: Continuous Distributions

1.2 FEATURES OF A CONTINUOUS DISTRIBUTION

As is true for relative frequency histograms, the probability density curves of continuous random variables could possess a wide variety of shapes. A few of these are illustrated in Figure 2. Many statisticians use the term **skewed** for a long tail in one direction.

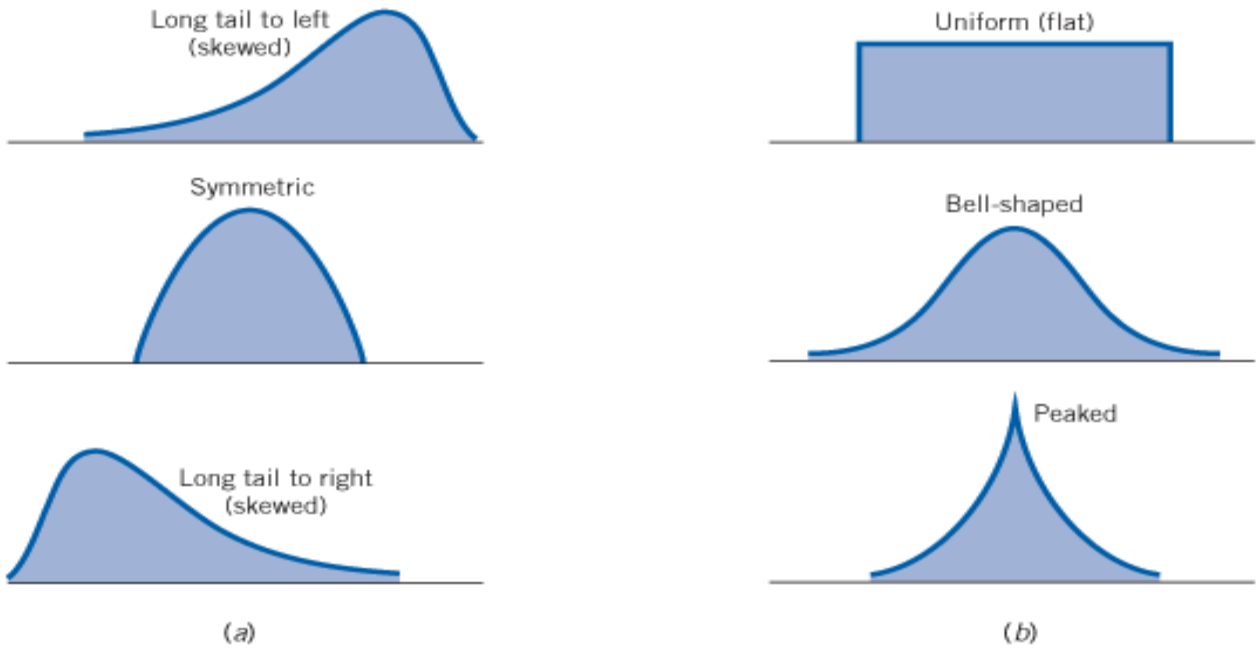


Figure 2 Different shapes of probability density curves. (*a*) Symmetry and deviations from symmetry. (*b*) Different peakedness.

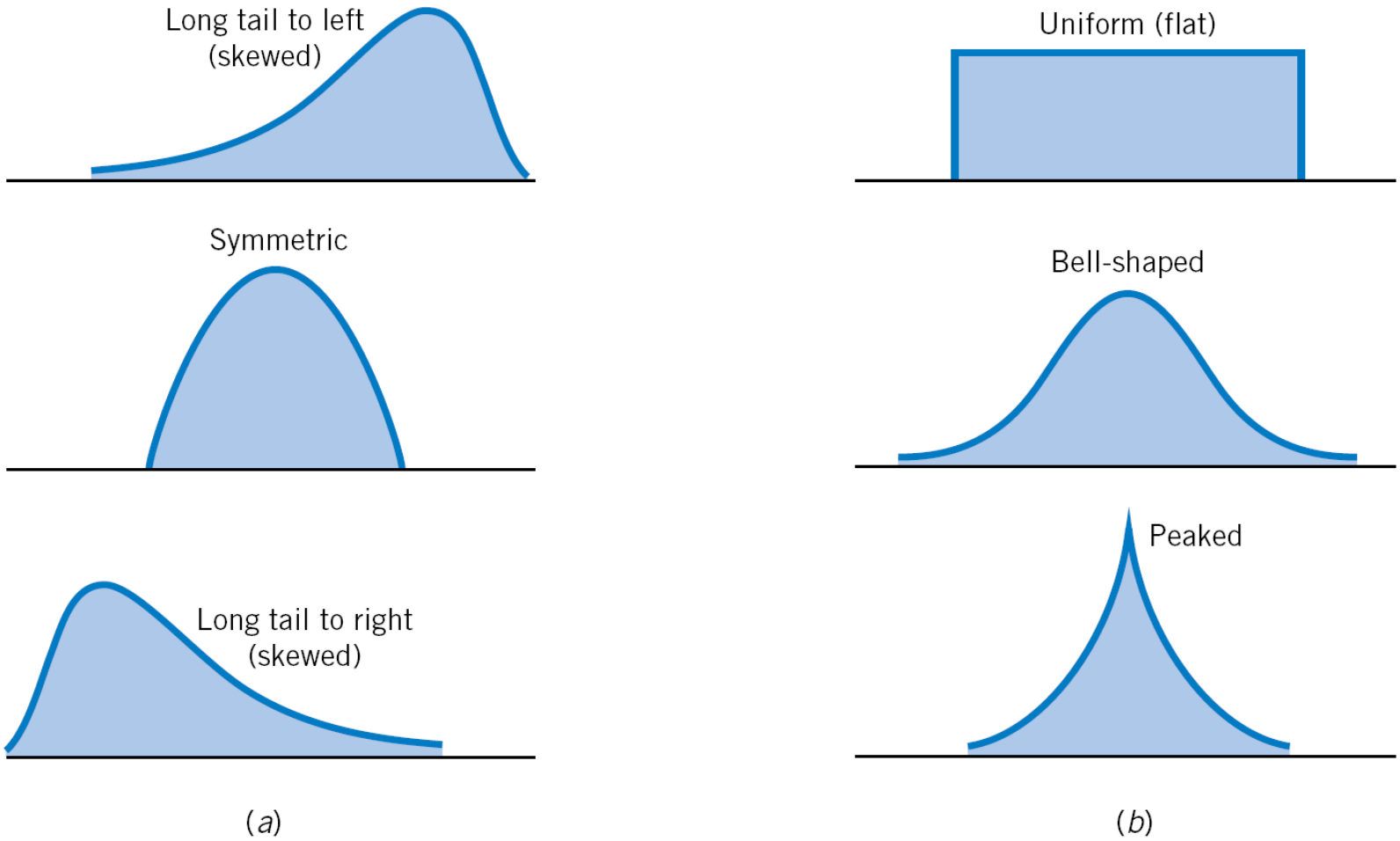


Figure 2, Different shapes of probability density curves. (a) Symmetry and deviations from symmetry; (b) different peakedness, p. 237.

Continuous Random Variable X Has Mean or Expected Value E(X)

A continuous random variable X also has a mean, or expected value $E(X)$, as well as a variance and a standard deviation. Their interpretations are the same as in the case of discrete random variables, but their formal definitions involve integral calculus and are therefore not pursued here. However, it is instructive to see in Figure 3 that the mean $\mu = E(X)$ marks the balance point of the probability mass. The median, another measure of center, is the value of X that divides the area under the curve into halves each with probability 0.5.

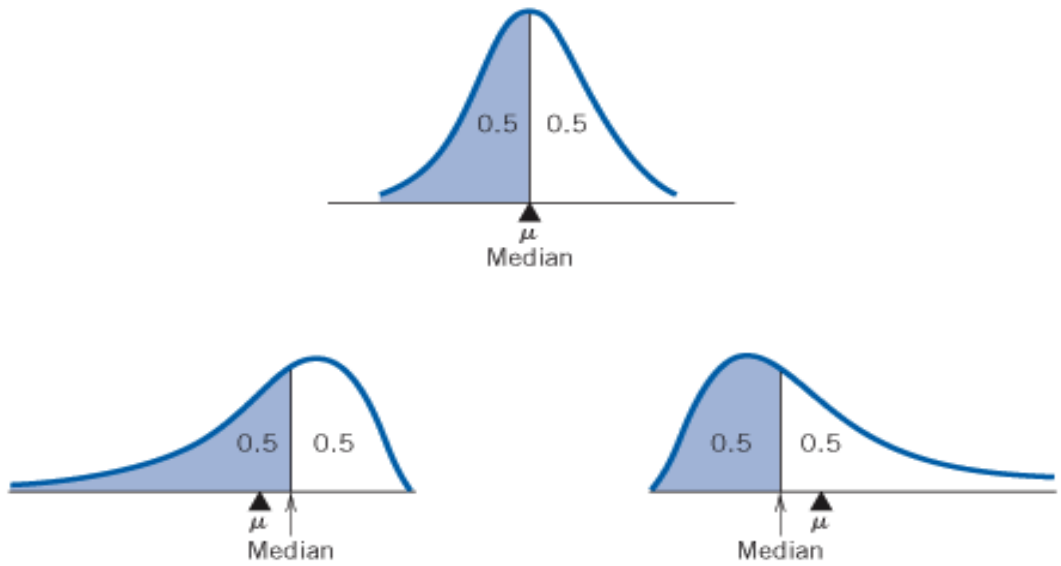


Figure 3 Mean as the balance point and median as the point of equal division of the probability mass.

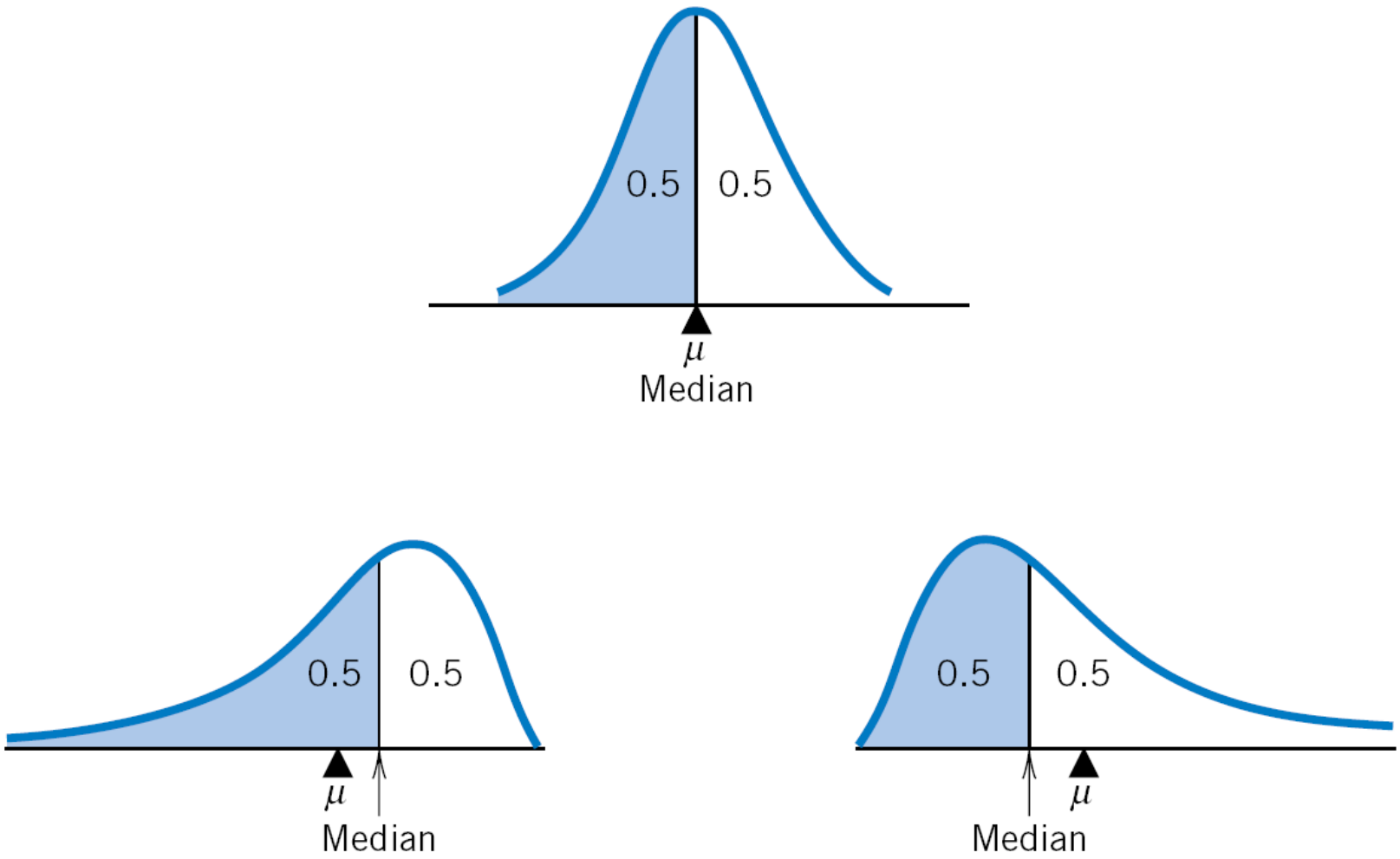


Figure 3, Mean as the balance point and median as the point of equal division of the probability mass, p. 237.

Quartiles and Percentiles of a Probability Distribution

Besides the median, we can also define the quartiles and other percentiles of a probability distribution.

The population **100 p-th percentile** is an x value that supports area p to its left and $1-p$ to its right.

Lower (first) quartile = 25th percentile
Second quartile (or median) = 50th percentile
Upper (third) quartile = 75th percentile

The quartiles for two distributions are shown in Figure 4.

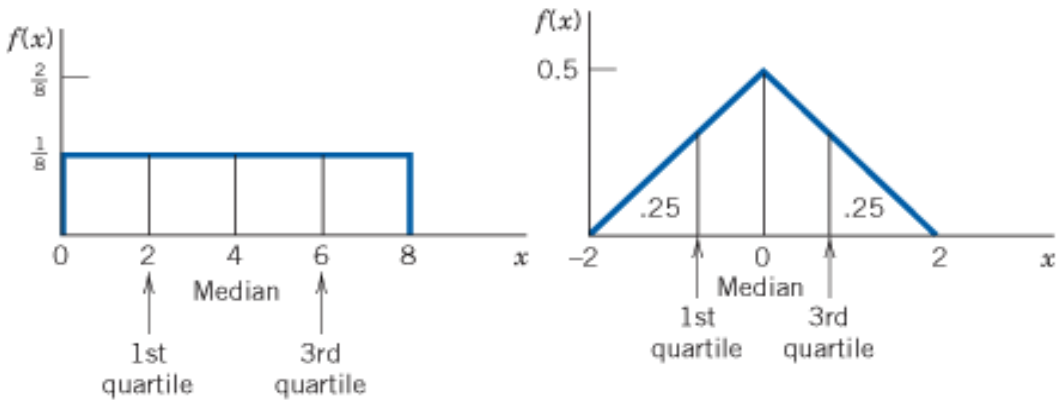
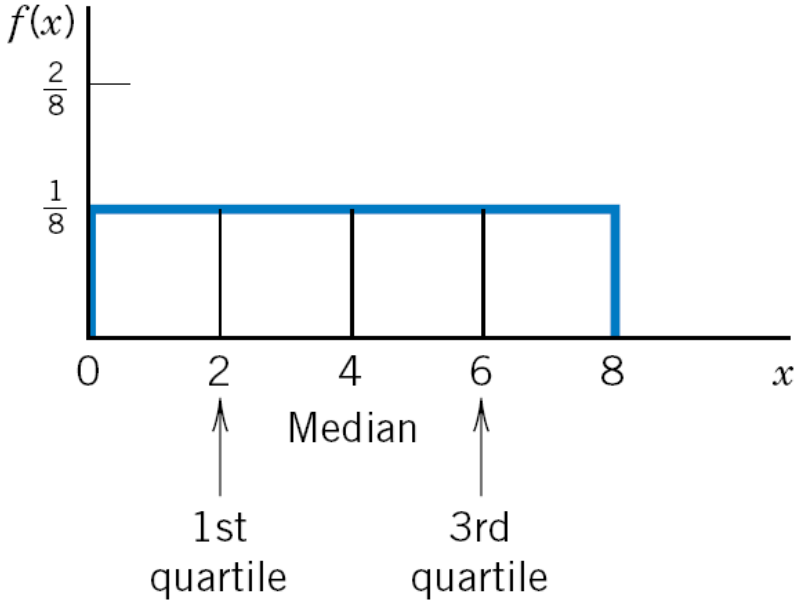


Figure 4 Quartiles of two continuous distributions.



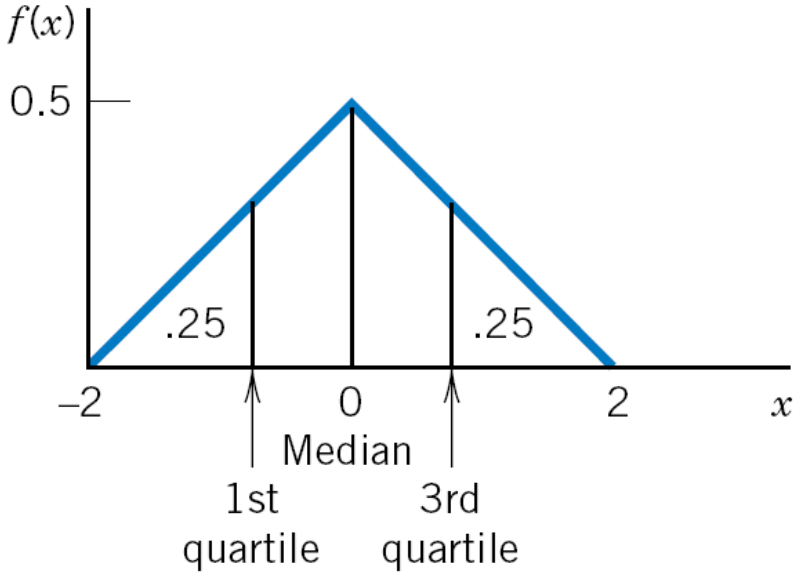


Figure 4, Quartile of two continuous distributions, p. 238.

The population **100*p*-th percentile** is an x value that supports area p to its left and $1 - p$ to its right.

Lower (first) quartile	=	25th percentile
Second quartile (or median)	=	50th percentile
Upper (third) quartile	=	75th percentile

Box, 110*p*-th percentile, p. 238.

Standardized Variable Z

Statisticians often find it convenient to convert random variables to a dimensionless scale. Suppose X, a real estate salesperson's commission for a month, has mean \$6000 and standard deviation \$500. Subtracting the mean produces the deviation $X - 6000$ measured in dollars. Then, dividing by the standard deviation, expressed in dollars, yields the dimensionless variable $Z = (X - 6000)/500$. Moreover, the standardized variable Z can be shown to have mean 0 and standard deviation 1. (See Appendix 1 for details.)

The observed values of standardized variables provide a convenient way to compare SAT and ACT scores or compare heights of male and female partners.

The standardized variable

$$Z = \frac{X - \mu}{\sigma} = \frac{\text{Variable} - \text{Mean}}{\text{Standard deviation}}$$

has mean 0 and sd 1.

Z-Score Formula Equal to Variance Covariance Calculation for 1 Variable

It is convenient to denote the normal density function with mean μ and variance σ^2 by $N(\mu, \sigma^2)$. Therefore, $N(10, 4)$ refers to the function in (4-1) with $\mu = 10$ and $\sigma = 2$. This notation will be extended to the multivariate case later.

The term

$$\left(\frac{x - \mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu) \tag{4-2}$$

in the exponent of the univariate normal density function measures the square of the distance from x to μ in standard deviation units. This can be generalized for a $p \times 1$ vector $\mathbf{x}$ of observations on several variables as

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{4-3}$$

Area and Volume of Normal Curve, Hence Probability Density Function

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{4-3}$$

The $p \times 1$ vector $\boldsymbol{\mu}$ represents the expected value of the random vector $\mathbf{X}$, and the $p \times p$ matrix $\boldsymbol{\Sigma}$ is the variance–covariance matrix of $\mathbf{X}$. [See (2–30) and (2–31).] We shall assume that the symmetric matrix $\boldsymbol{\Sigma}$ is positive definite, so the expression in (4-3) is the square of the generalized distance from $\mathbf{x}$ to $\boldsymbol{\mu}$.

The multivariate normal density is obtained by replacing the univariate distance in (4-2) by the multivariate generalized distance of (4-3) in the density function of (4-1). When this replacement is made, the univariate normalizing constant $(2\pi)^{-1/2}(\sigma^2)^{-1/2}$ must be changed to a more general constant that makes the *volume* under the surface of the multivariate density function unity for any p. This is necessary because, in the multivariate case, probabilities are represented by volumes under the surface over regions defined by intervals of the x_i values. It can be shown (see [1]) that this constant is $(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}$, and consequently, a p-dimensional normal density for the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ has the form

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2} \tag{4-4}$$

where $-\infty < x_i < \infty$, $i = 1, 2, \ldots, p$. We shall denote this p-dimensional normal density by $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which is analogous to the normal density in the univariate case.

Derivation Bivariate Normal Density

Example 4.1 (Bivariate normal density) Let us evaluate the $p = 2$-variate normal density in terms of the individual parameters $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$, $\sigma_{11} = \text{Var}(X_1)$, $\sigma_{22} = \text{Var}(X_2)$, and $\rho_{12} = \sigma_{12}/(\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}) = \text{Corr}(X_1, X_2)$.

Using Result 2A.8, we find that the inverse of the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

is

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

Introducing the correlation coefficient ρ_{12} by writing $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$, we obtain $\sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$, and the squared distance becomes

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

$$= [x_1 - \mu_1, x_2 - \mu_2] \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}$$

Derivation Bivariate Normal Density

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

$$= [x_1-\mu_1, x_2-\mu_2]\frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}\begin{bmatrix}\sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}$$

$$= \frac{\sigma_{22}(x_1-\mu_1)^2+\sigma_{11}(x_2-\mu_2)^2-2\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}$$

$$= \frac{1}{1-\rho_{12}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right] \qquad (4\text{-}5)$$

Derivation Bivariate Normal Density

The last expression is written in terms of the standardized values $(x_1 - \mu_1)/\sqrt{\sigma_{11}}$ and $(x_2 - \mu_2)/\sqrt{\sigma_{22}}$.

Next, since $|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$, we can substitute for $\boldsymbol{\Sigma}^{-1}$ and $|\boldsymbol{\Sigma}|$ in (4-4) to get the expression for the bivariate $(p = 2)$ normal density involving the individual parameters $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$, and ρ_{12}:

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}} \times \exp\left\{-\frac{1}{2(1 - \rho_{12}^2)}\left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\} \tag{4-6}$$

Probability Density Function

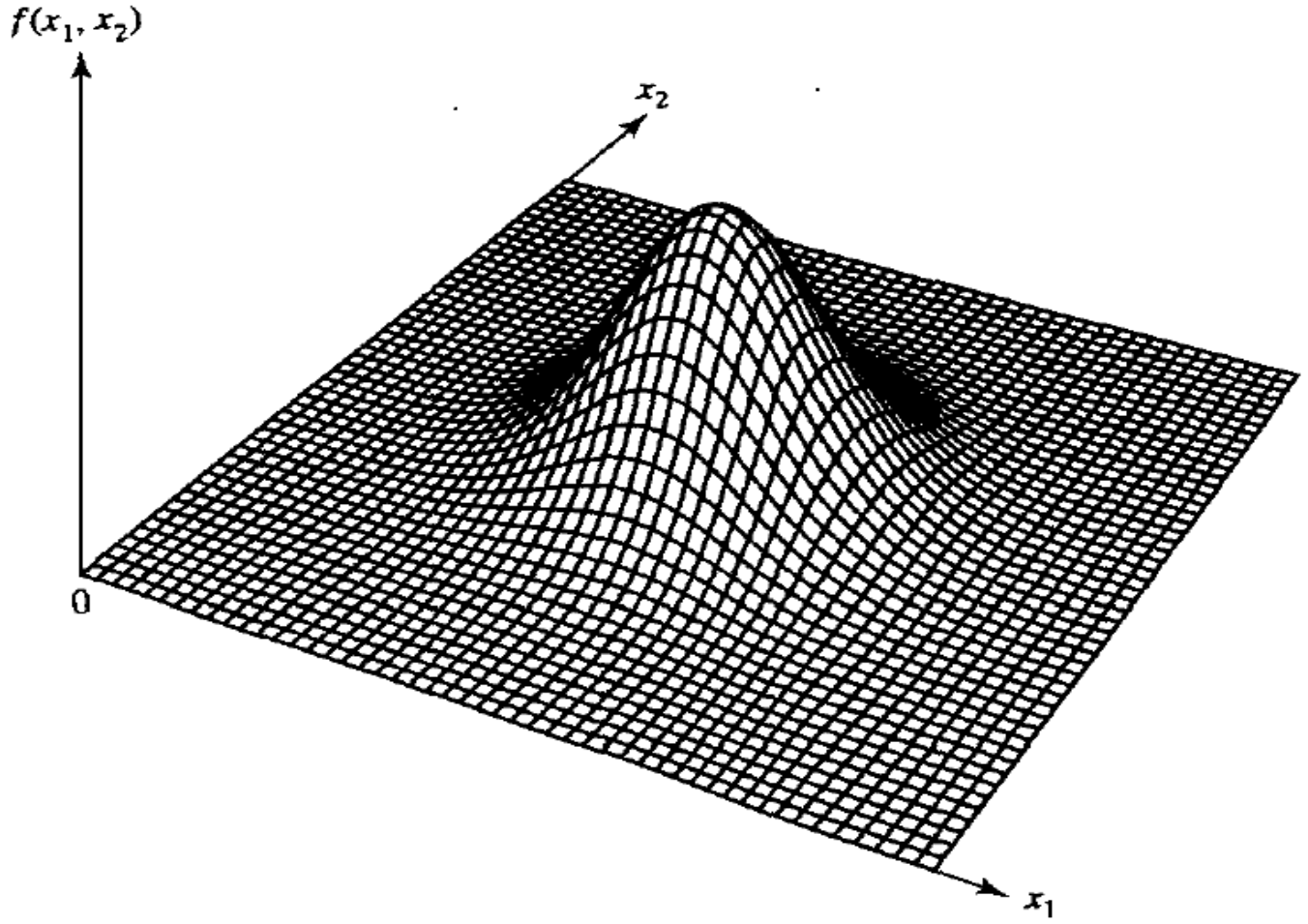
$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

The expression in (4-6) is somewhat unwieldy, and the compact general form in (4-4) is more informative in many ways. On the other hand, the expression in (4-6) is useful for discussing certain properties of the normal distribution. For example, if the random variables X_1 and X_2 are uncorrelated, so that $\rho_{12} = 0$, the joint density can be written as the product of two univariate normal densities each of the form of (4-1).

That is, $f(x_1, x_2) = f(x_1)f(x_2)$ and X_1 and X_2 are independent. [See (2-28).] This result is true in general. (See Result 4.5.)

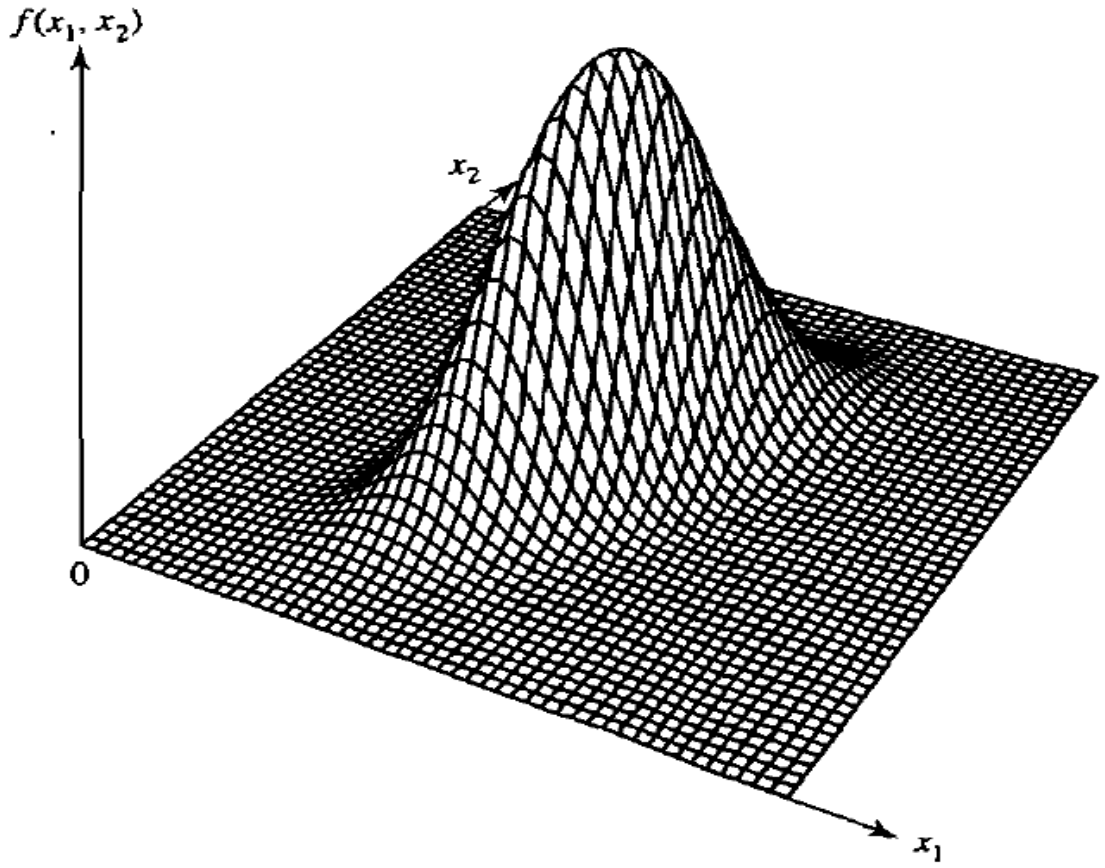
Two bivariate distributions with $\sigma_{11} = \sigma_{22}$ are shown in Figure 4.2. In Figure 4.2(a), X_1 and X_2 are independent ($\rho_{12} = 0$). In Figure 4.2(b), $\rho_{12} = .75$. Notice how the presence of correlation causes the probability to concentrate along a line. ■

Bivariate Distribution – 3D Scatter Plot: Correlation = 0



(a)

Bivariate Distribution – 3D Scatter Plot: Correlation = .75



(b)

Figure 4.2 Two bivariate normal distributions. (a) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = 0$. (b) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = .75$.

Constant Probability Density Contour = Surface of An Ellipsoid Centered at μ

From the expression in (4-4) for the density of a p-dimensional normal variable, it should be clear that the paths of $\mathbf{x}$ values yielding a constant height for the density are ellipsoids. That is, the multivariate normal density is constant on surfaces where the square of the distance $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is constant. These paths are called *contours*:

$$\begin{aligned} \textit{Constant probability density contour} &= \{\text{all } \mathbf{x} \text{ such that } (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2\} \\ &= \text{surface of an ellipsoid centered at } \boldsymbol{\mu} \end{aligned}$$

The axes of each ellipsoid of constant density are in the direction of the eigenvectors of $\boldsymbol{\Sigma}^{-1}$, and their lengths are proportional to the reciprocals of the square roots of the eigenvalues of $\boldsymbol{\Sigma}^{-1}$. Fortunately, we can avoid the calculation of $\boldsymbol{\Sigma}^{-1}$ when determining the axes, since these ellipsoids are also determined by the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$. We state the correspondence formally for later reference.

1/λ, e Is an Eigenvalue-Eigenvector Pair for Σ^{-1}

Result 4.1. If Σ is positive definite, so that Σ^{-1} exists, then

$$\Sigma\mathbf{e} = \lambda\mathbf{e} \quad \text{implies} \quad \Sigma^{-1}\mathbf{e} = \left(\frac{1}{\lambda}\right)\mathbf{e}$$

so $(\lambda, \mathbf{e})$ is an eigenvalue–eigenvector pair for Σ corresponding to the pair $(1/\lambda, \mathbf{e})$ for Σ^{-1}. Also, Σ^{-1} is positive definite.

Proof. For Σ positive definite and $\mathbf{e} \neq \mathbf{0}$ an eigenvector, we have $0 < \mathbf{e}'\Sigma\mathbf{e} = \mathbf{e}'(\Sigma\mathbf{e}) = \mathbf{e}'(\lambda\mathbf{e}) = \lambda\mathbf{e}'\mathbf{e} = \lambda$. Moreover, $\mathbf{e} = \Sigma^{-1}(\Sigma\mathbf{e}) = \Sigma^{-1}(\lambda\mathbf{e})$, or $\mathbf{e} = \lambda\Sigma^{-1}\mathbf{e}$, and division by $\lambda > 0$ gives $\Sigma^{-1}\mathbf{e} = (1/\lambda)\mathbf{e}$. Thus, $(1/\lambda, \mathbf{e})$ is an eigenvalue–eigenvector pair for Σ^{-1}. Also, for any $p \times 1$ $\mathbf{x}$, by (2-21)

$$\begin{aligned}\mathbf{x}'\Sigma^{-1}\mathbf{x} &= \mathbf{x}'\left(\sum_{i=1}^{p}\left(\frac{1}{\lambda_i}\right)\mathbf{e}_i\mathbf{e}_i'\right)\mathbf{x} \\ &= \sum_{i=1}^{p}\left(\frac{1}{\lambda_i}\right)(\mathbf{x}'\mathbf{e}_i)^2 \geq 0\end{aligned}$$

since each term $\lambda_i^{-1}(\mathbf{x}'\mathbf{e}_i)^2$ is nonnegative. In addition, $\mathbf{x}'\mathbf{e}_i = 0$ for all i only if $\mathbf{x} = \mathbf{0}$. So $\mathbf{x} \neq \mathbf{0}$ implies that $\sum_{i=1}^{p}(1/\lambda_i)(\mathbf{x}'\mathbf{e}_i)^2 > 0$, and it follows that Σ^{-1} is positive definite. ■

Contours of Constant Density for p-dimensional Distribution Are Ellipsoids

The following summarizes these concepts:

Contours of constant density for the p-dimensional normal distribution are ellipsoids defined by $\mathbf{x}$ such the that

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2 \qquad (4\text{-}7)$$

These ellipsoids are centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\,\mathbf{e}_i$, where $\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i\mathbf{e}_i$ for $i = 1, 2, \ldots, p$.

A contour of constant density for a bivariate normal distribution with $\sigma_{11} = \sigma_{22}$ is obtained in the following example.

Obtaining the Axes of Constant Probability Density Contours for Bivariate Normal Distribution

Example 4.2 (Contours of the bivariate normal density) We shall obtain the axes of constant probability density contours for a bivariate normal distribution when $\sigma_{11} = \sigma_{22}$. From (4-7), these axes are given by the eigenvalues and eigenvectors of Σ. Here $|\Sigma - \lambda \mathbf{I}| = 0$ becomes

$$0 = \begin{vmatrix} \sigma_{11} - \lambda & \sigma_{12} \\ \sigma_{12} & \sigma_{11} - \lambda \end{vmatrix} = (\sigma_{11} - \lambda)^2 - \sigma_{12}^2$$

$$= (\lambda - \sigma_{11} - \sigma_{12})(\lambda - \sigma_{11} + \sigma_{12})$$

Consequently, the eigenvalues are $\lambda_1 = \sigma_{11} + \sigma_{12}$ and $\lambda_2 = \sigma_{11} - \sigma_{12}$. The eigenvector $\mathbf{e}_1$ is determined from

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = (\sigma_{11} + \sigma_{12}) \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

or

$$\sigma_{11}e_1 + \sigma_{12}e_2 = (\sigma_{11} + \sigma_{12})e_1$$

$$\sigma_{12}e_1 + \sigma_{11}e_2 = (\sigma_{11} + \sigma_{12})e_2$$

Eigenvectors Have Length Unity, But the Major Axis Will be Associated with The Largest Eigenvalue

or

$$\sigma_{11}e_1 + \sigma_{12}e_2 = (\sigma_{11} + \sigma_{12})e_1$$
$$\sigma_{12}e_1 + \sigma_{11}e_2 = (\sigma_{11} + \sigma_{12})e_2$$

These equations imply that $e_1 = e_2$, and after normalization, the first eigenvalue–eigenvector pair is

$$\lambda_1 = \sigma_{11} + \sigma_{12}, \qquad \mathbf{e}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

Similarly, $\lambda_2 = \sigma_{11} - \sigma_{12}$ yields the eigenvector $\mathbf{e}_2' = [1/\sqrt{2}, -1/\sqrt{2}]$.

When the covariance σ_{12} (or correlation ρ_{12}) is positive, $\lambda_1 = \sigma_{11} + \sigma_{12}$ is the *largest* eigenvalue, and its associated eigenvector $\mathbf{e}_1' = [1/\sqrt{2}, 1/\sqrt{2}]$ lies along the 45° line through the point $\boldsymbol{\mu}' = [\mu_1, \mu_2]$. This is true for any positive value of the covariance (correlation). Since the axes of the constant-density ellipses are given by $\pm c\sqrt{\lambda_1}\,\mathbf{e}_1$ and $\pm c\sqrt{\lambda_2}\,\mathbf{e}_2$ [see (4–7)], and the eigenvectors each have length unity, the major axis will be associated with the largest eigenvalue. For positively correlated normal random variables, then, the *major* axis of the constant-density ellipses will be along the 45° line through $\boldsymbol{\mu}$. (See Figure 4.3.)

Bivariate Normal Distribution – Constant Density Contour for x_1 and x_2

When the covariance σ_{12} (or correlation ρ_{12}) is positive, $\lambda_1 = \sigma_{11} + \sigma_{12}$ is the *largest* eigenvalue, and its associated eigenvector $\mathbf{e}_1' = [1/\sqrt{2}, 1/\sqrt{2}]$ lies along the 45° line through the point $\boldsymbol{\mu}' = [\mu_1, \mu_2]$. This is true for any positive value of the covariance (correlation). Since the axes of the constant-density ellipses are given by $\pm c\sqrt{\lambda_1}\,\mathbf{e}_1$ and $\pm c\sqrt{\lambda_2}\,\mathbf{e}_2$ [see (4–7)], and the eigenvectors each have length unity, the major axis will be associated with the largest eigenvalue. For positively correlated normal random variables, then, the *major* axis of the constant-density ellipses will be along the 45° line through $\boldsymbol{\mu}$. (See Figure 4.3.)

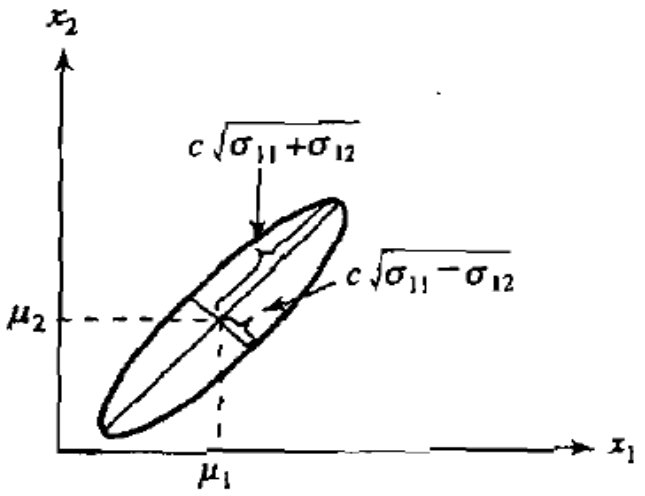


Figure 4.3 A constant-density contour for a bivariate normal distribution with $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} > 0$ (or $\rho_{12} > 0$).

When the Covariance (Correlation is Negative or Positive, λ Will be the Largest Eigenvalue

When the covariance (correlation) is negative, $\lambda_2 = \sigma_{11} - \sigma_{12}$ will be the largest eigenvalue, and the major axes of the constant-density ellipses will lie along a line at right angles to the 45° line through μ. (These results are true only for $\sigma_{11} = \sigma_{22}$.)

To summarize, the axes of the ellipses of constant density for a bivariate normal distribution with $\sigma_{11} = \sigma_{22}$ are determined by

$$\pm c\sqrt{\sigma_{11} + \sigma_{12}}\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \text{and} \quad \pm c\sqrt{\sigma_{11} - \sigma_{12}}\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$$

■

Solid Ellipsoid of x Values Satisfies

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2 \qquad (4\text{-}7)$$

We show in Result 4.7 that the choice $c^2 = \chi_p^2(\alpha)$, where $\chi_p^2(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with p degrees of freedom, leads to contours that contain $(1 - \alpha) \times 100\%$ of the probability. Specifically, the following is true for a p-dimensional normal distribution:

The solid ellipsoid of $\mathbf{x}$ values satisfying

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha) \qquad (4\text{-}8)$$

has probability $1 - \alpha$.

The constant-density contours containing 50% and 90% of the probability under the bivariate normal surfaces in Figure 4.2 are pictured in Figure 4.4.

Constant-Density Contours Containing 50% and 90% of the Probability Under Bivariate Normal Surfaces Contour Plot Figure 4.2

The constant-density contours containing 50% and 90% of the probability under the bivariate normal surfaces in Figure 4.2 are pictured in Figure 4.4.

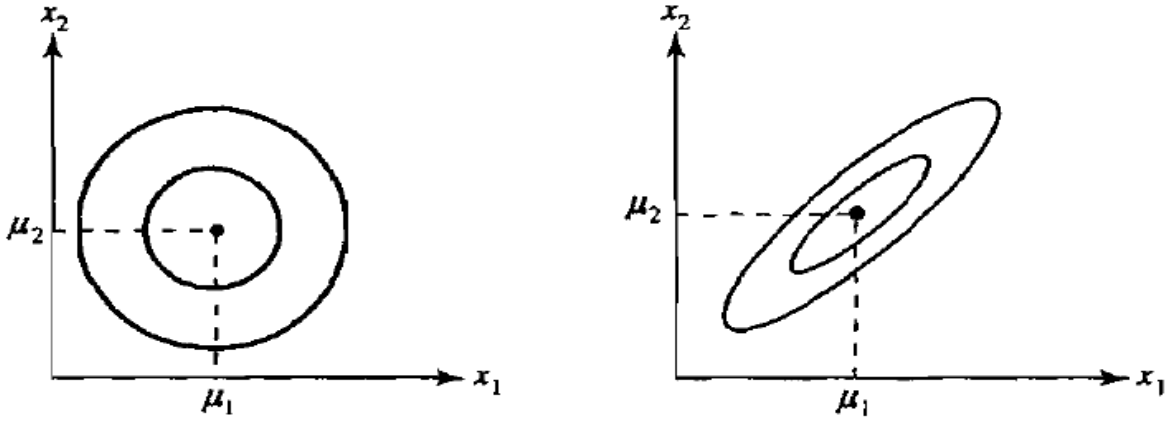


Figure 4.4 The 50% and 90% contours for the bivariate normal distributions in Figure 4.2.

The p-variate normal density in (4-4) has a maximum value when the squared distance in (4-3) is zero—that is, when $\mathbf{x} = \boldsymbol{\mu}$. Thus, $\boldsymbol{\mu}$ is the point of maximum density, or *mode*, as well as the expected value of $\mathbf{X}$, or *mean*. The fact that $\boldsymbol{\mu}$ is the mean of the multivariate normal distribution follows from the symmetry exhibited by the constant-density contours: These contours are centered, or balanced, at $\boldsymbol{\mu}$.

Properties of the Multivariate Normal Distribution (Including Covariance)

Additional Properties of the Multivariate Normal Distribution

Certain properties of the normal distribution will be needed repeatedly in our explanations of statistical models and methods. These properties make it possible to manipulate normal distributions easily and, as we suggested in Section 4.1, are partly responsible for the popularity of the normal distribution. The key properties, which we shall soon discuss in some mathematical detail, can be stated rather simply.

The following are true for a random vector $\mathbf{X}$ having a multivariate normal distribution:

1. Linear combinations of the components of $\mathbf{X}$ are normally distributed.
2. All subsets of the components of $\mathbf{X}$ have a (multivariate) normal distribution.
3. Zero covariance implies that the corresponding components are independently distributed.
4. The conditional distributions of the components are (multivariate) normal.

Any Linear Combination of Variables a'X = $a_1X_1+a_2X_2+....+a_pX_p$ Distributed as N(a'u, a'Σa)

These statements are reproduced mathematically in the results that follow. Many of these results are illustrated with examples. The proofs that are included should help improve your understanding of matrix manipulations and also lead you to an appreciation for the manner in which the results successively build on themselves.

Result 4.2 can be taken as a working definition of the normal distribution. With this in hand, the subsequent properties are almost immediate. Our partial proof of Result 4.2 indicates how the linear combination definition of a normal density relates to the multivariate density in (4-4).

Result 4.2. If $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination of variables $\mathbf{a}'\mathbf{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p$ is distributed as $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$. Also, if $\mathbf{a}'\mathbf{X}$ is distributed as $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$ for every $\mathbf{a}$, then $\mathbf{X}$ must be $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Proof. The expected value and variance of $\mathbf{a}'\mathbf{X}$ follow from (2-43). Proving that $\mathbf{a}'\mathbf{X}$ is normally distributed if $\mathbf{X}$ is multivariate normal is more difficult. You can find a proof in [1]. The second part of result 4.2 is also demonstrated in [1]. ■

Distribution of Linear Combination of Components of Normal Random Vector

Example 4.3 (The distribution of a linear combination of the components of a normal random vector) Consider the linear combination $\mathbf{a}'\mathbf{X}$ of a multivariate normal random vector determined by the choice $\mathbf{a}' = [1, 0, \ldots, 0]$. Since

$$\mathbf{a}'\mathbf{X} = [1, 0, \ldots, 0] \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = X_1$$

X_1 Distributed as $N(\mu_1, \sigma_{11})$ Marginal Distribution of Any Component X_i of X is $N(u_i, \sigma_{ii})$

and

$$\mathbf{a}'\boldsymbol{\mu} = [1, 0, \ldots, 0] \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \mu_1$$

we have

$$\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = [1, 0, \ldots, 0] \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \sigma_{11}$$

and it follows from Result 4.2 that X_1 is distributed as $N(\mu_1, \sigma_{11})$. More generally, the marginal distribution of any component X_i of $\mathbf{X}$ is $N(\mu_i, \sigma_{ii})$. ■

Linear Combination b'(AX) is Linear Combination X of the form a'X with a= A'b

The next result considers several linear combinations of a multivariate normal vector $\mathbf{X}$.

Result 4.3. If $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the q linear combinations

$$\underset{(q\times p)}{\mathbf{A}}\ \underset{(p\times 1)}{\mathbf{X}} = \begin{bmatrix} a_{11}X_1 + \cdots + a_{1p}X_p \\ a_{21}X_1 + \cdots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + \cdots + a_{qp}X_p \end{bmatrix}$$

are distributed as $N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. Also, $\underset{(p\times 1)}{\mathbf{X}} + \underset{(p\times 1)}{\mathbf{d}}$, where $\mathbf{d}$ is a vector of constants, is distributed as $N_p(\boldsymbol{\mu} + \mathbf{d}, \boldsymbol{\Sigma})$.

Proof. The expected value $E(\mathbf{AX})$ and the covariance matrix of $\mathbf{AX}$ follow from (2-45). Any linear combination $\mathbf{b}'(\mathbf{AX})$ is a linear combination of $\mathbf{X}$, of the form $\mathbf{a}'\mathbf{X}$ with $\mathbf{a} = \mathbf{A}'\mathbf{b}$. Thus, the conclusion concerning $\mathbf{AX}$ follows directly from Result 4.2.

A'X is Distributed as N(a'μ, a'Σa); Note: Adding A Constant Leaves the Variance Unchanged

The second part of the result can be obtained by considering $\mathbf{a}'(\mathbf{X} + \mathbf{d}) = \mathbf{a}'\mathbf{X} + (\mathbf{a}'\mathbf{d})$, where $\mathbf{a}'\mathbf{X}$ is distributed as $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$. It is known from the univariate case that adding a constant $\mathbf{a}'\mathbf{d}$ to the random variable $\mathbf{a}'\mathbf{X}$ leaves the variance unchanged and translates the mean to $\mathbf{a}'\boldsymbol{\mu} + \mathbf{a}'\mathbf{d} = \mathbf{a}'(\boldsymbol{\mu} + \mathbf{d})$. Since $\mathbf{a}$ was arbitrary, $\mathbf{X} + \mathbf{d}$ is distributed as $N_p(\boldsymbol{\mu} + \mathbf{d}, \boldsymbol{\Sigma})$. ■

Distribution 2 Linear Combinations of the Components of Normal Random Vector

Example 4.4 (The distribution of two linear combinations of the components of a normal random vector) For $\mathbf{X}$ distributed as $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, find the distribution of

$$\begin{bmatrix} X_1 - X_2 \\ X_2 - X_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \mathbf{AX}$$

Distribution of AX is Multivariate Normal with Mean Aμ and AΣA'

By Result 4.3, the distribution of $\mathbf{AX}$ is multivariate normal with mean

$$\mathbf{A}\boldsymbol{\mu} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \end{bmatrix}$$

and covariance matrix

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{22} & \sigma_{13} - \sigma_{23} \\ \sigma_{12} - \sigma_{13} & \sigma_{22} - \sigma_{23} & \sigma_{23} - \sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} \\ \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} & \sigma_{22} - 2\sigma_{23} + \sigma_{33} \end{bmatrix}$$

Mean Vector Aμ and Covariance Matrix AΣA' Verified by Direct Calculation Means and Covariances Two Random Variables x and y

Alternatively, the mean vector $\mathbf{A}\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ may be verified by direct calculation of the means and covariances of the two random variables $Y_1 = X_1 - X_2$ and $Y_2 = X_2 - X_3$. ■

All Subsets X Are Normally Distributed. Partition X, Mean Vector u and Covariance Matrix Σ, Then X1 Distributed As $N_q(\mu_1, \Sigma_{11})$

We have mentioned that all subsets of a multivariate normal random vector $\mathbf{X}$ are themselves normally distributed. We state this property formally as Result 4.4.

Result 4.4. All subsets of $\mathbf{X}$ are normally distributed. If we respectively partition $\mathbf{X}$, its mean vector $\boldsymbol{\mu}$, and its covariance matrix $\boldsymbol{\Sigma}$ as

$$\underset{(p\times 1)}{\mathbf{X}} = \begin{bmatrix} \underset{(q\times 1)}{\mathbf{X}_1} \\ \hdashline \underset{((p-q)\times 1)}{\mathbf{X}_2} \end{bmatrix} \qquad \underset{(p\times 1)}{\boldsymbol{\mu}} = \begin{bmatrix} \underset{(q\times 1)}{\boldsymbol{\mu}_1} \\ \hdashline \underset{((p-q)\times 1)}{\boldsymbol{\mu}_2} \end{bmatrix}$$

and

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \left[\begin{array}{c:c} \underset{(q\times q)}{\boldsymbol{\Sigma}_{11}} & \underset{(q\times(p-q))}{\boldsymbol{\Sigma}_{12}} \\ \hdashline \underset{((p-q)\times q)}{\boldsymbol{\Sigma}_{21}} & \underset{((p-q)\times(p-q))}{\boldsymbol{\Sigma}_{22}} \end{array}\right]$$

then $\mathbf{X}_1$ is distributed as $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

Proof Distribution of Subset Normal Random Vector

$\mathbf{X}_1$ is distributed as $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

Proof. Set $\underset{(q\times p)}{\mathbf{A}} = \left[\underset{(q\times q)}{\mathbf{I}} \vdots \underset{(q\times(p-q))}{\mathbf{0}}\right]$ in Result 4.3, and the conclusion follows.

To apply Result 4.4 to an *arbitrary* subset of the components of $\mathbf{X}$, we simply relabel the subset of interest as $\mathbf{X}_1$ and select the corresponding component means and covariances as $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_{11}$, respectively. ■

Example: Distribution of a Subset of Normal Random Vector

Example 4.5 (The distribution of a subset of a normal random vector)

If $\mathbf{X}$ is distributed as $N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, find the distribution of $\begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$. We set

$$\mathbf{X}_1 = \begin{bmatrix} X_2 \\ X_4 \end{bmatrix}, \qquad \boldsymbol{\mu}_1 = \begin{bmatrix} \mu_2 \\ \mu_4 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_{11} = \begin{bmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{24} & \sigma_{44} \end{bmatrix}$$

and note that with this assignment, $\mathbf{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ can respectively be rearranged and partitioned as

$$\mathbf{X} = \begin{bmatrix} X_2 \\ X_4 \\ \hdashline X_1 \\ X_3 \\ X_5 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_2 \\ \mu_4 \\ \hdashline \mu_1 \\ \mu_3 \\ \mu_5 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \left[\begin{array}{cc:ccc} \sigma_{22} & \sigma_{24} & \sigma_{12} & \sigma_{23} & \sigma_{25} \\ \sigma_{24} & \sigma_{44} & \sigma_{14} & \sigma_{34} & \sigma_{45} \\ \hdashline \sigma_{12} & \sigma_{14} & \sigma_{11} & \sigma_{13} & \sigma_{15} \\ \sigma_{23} & \sigma_{34} & \sigma_{13} & \sigma_{33} & \sigma_{35} \\ \sigma_{25} & \sigma_{45} & \sigma_{15} & \sigma_{35} & \sigma_{55} \end{array}\right]$$

Normal Distribution For Any Subset Can Be Expressed by Simply Selecting Means and Covariances from the Original μ and Σ

or

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ (2\times 1) \\ \cdots\cdots \\ \mathbf{X}_2 \\ (3\times 1) \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ (2\times 1) \\ \cdots\cdots \\ \boldsymbol{\mu}_2 \\ (3\times 1) \end{bmatrix}, \quad \boldsymbol{\Sigma} = \left[\begin{array}{c:c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ (2\times 2) & (2\times 3) \\ \hdashline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \\ (3\times 2) & (3\times 3) \end{array}\right]$$

Thus, from Result 4.4, for

$$\mathbf{X}_1 = \begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$$

we have the distribution

$$N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) = N_2\left(\begin{bmatrix} \mu_2 \\ \mu_4 \end{bmatrix}, \begin{bmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{24} & \sigma_{44} \end{bmatrix}\right)$$

It is clear from this example that the normal distribution for any subset can be expressed by simply selecting the appropriate means and covariances from the original $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The formal process of relabeling and partitioning is unnecessary. ■

Zero Correlation Between Normal Random Variables is Equivalent to Statistical Independence

We are now in a position to state that zero correlation between normal random variables or sets of normal random variables is equivalent to statistical independence.

Result 4.5.

(a) If $\underset{(q_1 \times 1)}{\mathbf{X}_1}$ and $\underset{(q_2 \times 1)}{\mathbf{X}_2}$ are independent, then $\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$, a $q_1 \times q_2$ matrix of zeros.

(b) If $\begin{bmatrix} \mathbf{X}_1 \\ \hline \mathbf{X}_2 \end{bmatrix}$ is $N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{bmatrix}, \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] \right)$, then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

Zero Correlation Between Normal Random Variables is Equivalent to Statistical Independence (cont.)

(c) If $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent and are distributed as $N_{q_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $N_{q_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, respectively, then $\begin{bmatrix} \mathbf{X}_1 \\ \hdashline \mathbf{X}_2 \end{bmatrix}$ has the multivariate normal distribution

$$N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \hdashline \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \hdashline \mathbf{0}' & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$$

Proof. (See Exercise 4.14 for partial proofs based upon factoring the density function when $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.) ■

Example: Equivalence of Zero Covariance and Independence for Normal Variables

Example 4.6 (The equivalence of zero covariance and independence for normal variables) Let $\underset{(3\times 1)}{\mathbf{X}}$ be $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Are X_1 and X_2 independent? What about (X_1, X_2) and X_3?

Since X_1 and X_2 have covariance $\sigma_{12} = 1$, they are not independent. However, partitioning $\mathbf{X}$ and $\boldsymbol{\Sigma}$ as

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \hline X_3 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \left[\begin{array}{cc:c} 4 & 1 & 0 \\ 1 & 3 & 0 \\ \hdashline 0 & 0 & 2 \end{array}\right] = \left[\begin{array}{c:c} \underset{(2\times 2)}{\boldsymbol{\Sigma}_{11}} & \underset{(2\times 1)}{\boldsymbol{\Sigma}_{12}} \\ \hdashline \underset{(1\times 2)}{\boldsymbol{\Sigma}_{21}} & \underset{(1\times 1)}{\boldsymbol{\Sigma}_{22}} \end{array}\right]$$

we see that $\mathbf{X}_1 = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ and X_3 have covariance matrix $\boldsymbol{\Sigma}_{12} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Therefore, (X_1, X_2) and X_3 are independent by Result 4.5. This implies X_3 is independent of X_1 and also of X_2. ■

Conditional Distribution of X_1, Given $X_2 = x_2$ is normal

We pointed out in our discussion of the bivariate normal distribution that $\rho_{12} = 0$ (zero correlation) implied independence because the joint density function [see (4-6)] could then be written as the product of the marginal (normal) densities of X_1 and X_2. This fact, which we encouraged you to verify directly, is simply a special case of Result 4.5 with $q_1 = q_2 = 1$.

Result 4.6. Let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \cdots \\ \mathbf{X}_2 \end{bmatrix}$ be distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \cdots \\ \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$, and $|\boldsymbol{\Sigma}_{22}| > 0$. Then the conditional distribution of $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, is normal and has

$$\text{Mean} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

Covariance Does Not Depend on Value of x_2 of the Conditioning Variable

$$\text{Covariance} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Note that the covariance does not depend on the value $\mathbf{x}_2$ of the conditioning variable.

Proof. We shall give an indirect proof. (See Exercise 4.13, which uses the densities directly.) Take

$$\underset{(p\times p)}{\mathbf{A}} = \left[\begin{array}{c|c} \underset{(q\times q)}{\mathbf{I}} & \underset{q\times(p-q)}{-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}} \\ \hline \underset{(p-q)\times q}{\mathbf{0}} & \underset{(p-q)\times(p-q)}{\mathbf{I}} \end{array}\right]$$

so

$$\mathbf{A}(\mathbf{X}-\boldsymbol{\mu}) = \mathbf{A}\left[\begin{array}{c} \mathbf{X}_1-\boldsymbol{\mu}_1 \\ \hline \mathbf{X}_2-\boldsymbol{\mu}_2 \end{array}\right] = \left[\begin{array}{c} \mathbf{X}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2-\boldsymbol{\mu}_2) \\ \hline \mathbf{X}_2-\boldsymbol{\mu}_2 \end{array}\right]$$

is jointly normal with covariance matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ given by

Jointly Normal Covariance Matrix AΣA’

is jointly normal with covariance matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ given by

$$\left[\begin{array}{c|c} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0}' \\ \hline (-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1})' & \mathbf{I} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{0}' \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{array}\right].$$

Since $\mathbf{X}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$ and $\mathbf{X}_2 - \boldsymbol{\mu}_2$ have zero covariance, they are independent. Moreover, the quantity $\mathbf{X}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$ has distribution $N_q(\mathbf{0}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$. Given that $\mathbf{X}_2 = \mathbf{x}_2$, $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ is a constant. Because $\mathbf{X}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$ and $\mathbf{X}_2 - \boldsymbol{\mu}_2$ are independent, the conditional distribution of $\mathbf{X}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ is the same as the unconditional distribution of $\mathbf{X}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$. Since $\mathbf{X}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$ is $N_q(\mathbf{0}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$, so is the random vector $\mathbf{X}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ when $\mathbf{X}_2$ has the particular value $\mathbf{x}_2$. Equivalently, given that $\mathbf{X}_2 = \mathbf{x}_2$, $\mathbf{X}_1$ is distributed as $N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$. ■

Conditional Density Bivariate Normal Distribution

Example 4.7 (The conditional density of a bivariate normal distribution) The conditional density of X_1, given that $X_2 = x_2$ for any bivariate distribution, is defined by

$$f(x_1|x_2) = \{\text{conditional density of } X_1 \text{ given that } X_2 = x_2\} = \frac{f(x_1, x_2)}{f(x_2)}$$

where $f(x_2)$ is the marginal distribution of X_2. If $f(x_1, x_2)$ is the bivariate normal density, show that $f(x_1|x_2)$ is

$$N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2), \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right)$$

Conditional Density Bivariate Normal Distribution (cont.)

Here $\sigma_{11} - \sigma_{12}^2/\sigma_{22} = \sigma_{11}(1 - \rho_{12}^2)$. The two terms involving $x_1 - \mu_1$ in the exponent of the bivariate normal density [see Equation (4-6)] become, apart from the multiplicative constant $-1/2(1 - \rho_{12}^2)$,

$$\frac{(x_1 - \mu_1)^2}{\sigma_{11}} - 2\rho_{12}\frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$$

$$= \frac{1}{\sigma_{11}}\left[x_1 - \mu_1 - \rho_{12}\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(x_2 - \mu_2)\right]^2 - \frac{\rho_{12}^2}{\sigma_{22}}(x_2 - \mu_2)^2$$

Because $\rho_{12} = \sigma_{12}/\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$, or $\rho_{12}\sqrt{\sigma_{11}}/\sqrt{\sigma_{22}} = \sigma_{12}/\sigma_{22}$, the complete exponent is

$$\frac{-1}{2(1 - \rho_{12}^2)}\left(\frac{(x_1 - \mu_1)^2}{\sigma_{11}} - 2\rho_{12}\frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} + \frac{(x_2 - \mu_2)^2}{\sigma_{22}}\right)$$

$$= \frac{-1}{2\sigma_{11}(1 - \rho_{12}^2)}\left(x_1 - \mu_1 - \rho_{12}\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(x_2 - \mu_2)\right)^2$$

$$- \frac{1}{2(1 - \rho_{12}^2)}\left(\frac{1}{\sigma_{22}} - \frac{\rho_{12}^2}{\sigma_{22}}\right)(x_2 - \mu_2)^2$$

Joint Density of X_1 and X_2 By the Marginal Density Yields Conditional Density

$$- \frac{1}{2(1-\rho_{12}^2)}\left(\frac{1}{\sigma_{22}} - \frac{\rho_{12}^2}{\sigma_{22}}\right)(x_2-\mu_2)^2$$

$$= \frac{-1}{2\sigma_{11}(1-\rho_{12}^2)}\left(x_1 - \mu_1 - \frac{\sigma_{12}}{\sigma_{22}}(x_2-\mu_2)\right)^2 - \frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_{22}}$$

The constant term $2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}$ also factors as

$$\sqrt{2\pi}\cdot\sqrt{\sigma_{22}} \times \sqrt{2\pi}\sqrt{\sigma_{11}(1-\rho_{12}^2)}$$

Dividing the joint density of X_1 and X_2 by the marginal density

$$f(x_2) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_{22}}} e^{-(x_2-\mu_2)^2/2\sigma_{22}}$$

and canceling terms yields the conditional density

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)}$$

Result: Conditional Distribution of X_1 Given that $X_2 = x_2$ is Normally Distributed Agreeing With

Covariance $= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

and canceling terms yields the conditional density

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_{11}(1-\rho_{12}^2)}} e^{-[x_1-\mu_1-(\sigma_{12}/\sigma_{22})(x_2-\mu_2)]^2/2\sigma_{11}(1-\rho_{12}^2)},$$

$$-\infty < x_1 < \infty$$

Thus, with our customary notation, the conditional distribution of X_1 given that $X_2 = x_2$ is $N(\mu_1 + (\sigma_{12}/\sigma_{22})(x_2 - \mu_2),\ \sigma_{11}(1-\rho_{12}^2))$. Now, $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \sigma_{11} - \sigma_{12}^2/\sigma_{22} = \sigma_{11}(1-\rho_{12}^2)$ and $\Sigma_{12}\Sigma_{22}^{-1} = \sigma_{12}/\sigma_{22}$, agreeing with Result 4.6, which we obtained by an indirect method. ■

Summary of Multivariate Normal Applications

For the multivariate normal situation, it is worth emphasizing the following:

1. All conditional distributions are (multivariate) normal.
2. The conditional mean is of the form

$$\begin{matrix} \mu_1 + \beta_{1,q+1}(x_{q+1} - \mu_{q+1}) + \cdots + \beta_{1,p}(x_p - \mu_p) \\ \vdots \\ \mu_q + \beta_{q,q+1}(x_{q+1} - \mu_{q+1}) + \cdots + \beta_{q,p}(x_p - \mu_p) \end{matrix} \tag{4-9}$$

where the β's are defined by

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} = \begin{bmatrix} \beta_{1,q+1} & \beta_{1,q+2} & \cdots & \beta_{1,p} \\ \beta_{2,q+1} & \beta_{2,q+2} & \cdots & \beta_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q,q+1} & \beta_{q,q+2} & \cdots & \beta_{q,p} \end{bmatrix}$$

Chi-Square Distribution Determines the Variability of the Sample Variance For Samples From a Univariate Normal Population

3. The conditional covariance, $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, does not depend upon the value(s) of the conditioning variable(s).

We conclude this section by presenting two final properties of multivariate normal random vectors. One has to do with the probability content of the ellipsoids of constant density. The other discusses the distribution of another form of linear combinations.

The chi-square distribution determines the variability of the sample variance $s^2 = s_{11}$ for samples from a univariate normal population. It also plays a basic role in the multivariate case.

Result 4.7. Let $\mathbf{X}$ be distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $|\boldsymbol{\Sigma}| > 0$. Then

(a) $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ is distributed as χ_p^2, where χ_p^2 denotes the chi-square distribution with p degrees of freedom.

(b) The $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution assigns probability $1 - \alpha$ to the solid ellipsoid $\{\mathbf{x}: (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha)\}$, where $\chi_p^2(\alpha)$ denotes the upper (100α)th percentile of the χ_p^2 distribution.

Proof Chi-Square Distribution Determines the Variability of Sample Variance from Univariate Normal Population

Proof. We know that χ^2_p is defined as the distribution of the sum $Z_1^2 + Z_2^2 + \cdots + Z_p^2$, where $Z_1, Z_2, \ldots, Z_p$ are independent $N(0,1)$ random variables. Next, by the spectral decomposition [see Equations (2-16) and (2-21) with $\mathbf{A} = \mathbf{\Sigma}$, and see Result 4.1], $\mathbf{\Sigma}^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i'$, where $\mathbf{\Sigma}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, so $\mathbf{\Sigma}^{-1}\mathbf{e}_i = (1/\lambda_i)\mathbf{e}_i$. Consequently,

$$(\mathbf{X} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p} (1/\lambda_i)(\mathbf{X} - \boldsymbol{\mu})'\mathbf{e}_i\mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p} (1/\lambda_i)(\mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu}))^2 =$$

$$\sum_{i=1}^{p} [(1/\sqrt{\lambda_i})\, \mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu})]^2 = \sum_{i=1}^{p} Z_i^2, \text{ for instance. Now, we can write } \mathbf{Z} = \mathbf{A}(\mathbf{X} - \boldsymbol{\mu}),$$

X-μ is Distributed as N_p (o, Σ)

where

$$\underset{(p\times 1)}{\mathbf{Z}} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix}, \qquad \underset{(p\times p)}{\mathbf{A}} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}}\mathbf{e}_1' \\ \frac{1}{\sqrt{\lambda_2}}\mathbf{e}_2' \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}}\mathbf{e}_p' \end{bmatrix}$$

and $\mathbf{X} - \boldsymbol{\mu}$ is distributed as $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Therefore, by Result 4.3, $\mathbf{Z} = \mathbf{A}(\mathbf{X} - \boldsymbol{\mu})$ is distributed as $N_p(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$, where

$$\underset{(p\times p)(p\times p)(p\times p)}{\mathbf{A}\ \boldsymbol{\Sigma}\ \mathbf{A}'} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}}\mathbf{e}_1' \\ \frac{1}{\sqrt{\lambda_2}}\mathbf{e}_2' \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}}\mathbf{e}_p' \end{bmatrix} \left[\sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i' \right] \left[\frac{1}{\sqrt{\lambda_1}}\mathbf{e}_1 \;\middle|\; \frac{1}{\sqrt{\lambda_2}}\mathbf{e}_2 \;\middle|\; \cdots \;\middle|\; \frac{1}{\sqrt{\lambda_p}}\mathbf{e}_p \right]$$

Independent Standard Normal Variables With Chi-Square Distribution

$$= \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1' \\ \sqrt{\lambda_2}\,\mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_p}\,\mathbf{e}_p' \end{bmatrix} \left[\frac{1}{\sqrt{\lambda_1}}\mathbf{e}_1 \;\middle|\; \frac{1}{\sqrt{\lambda_2}}\mathbf{e}_2 \;\middle|\; \cdots \;\middle|\; \frac{1}{\sqrt{\lambda_p}}\mathbf{e}_p \right] = \mathbf{I}$$

By Result 4.5, $Z_1, Z_2, \ldots, Z_p$ are *independent* standard normal variables, and we conclude that $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ has a χ^2_p-distribution.

For Part b, we note that $P[(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \le c^2]$ is the probability assigned to the ellipsoid $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \le c^2$ by the density $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. But from Part a, $P[(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \le \chi^2_p(\alpha)] = 1 - \alpha$, and Part b holds. ■

Pearson's Chi-Square Goodness of Fit Test with Categorical Data

2. Pearson's χ^2 Test for Goodness of Fit

We first consider the type of problem illustrated in Example 1, where the data consist of frequency counts observed from a random sample and the null hypothesis specifies the unknown cell probabilities. Our primary goal is to test if the model given by the null hypothesis fits the data, and this is appropriately called **testing for goodness of fit**.

For general discussion, suppose a random sample of size n is classified into k categories or cells labeled $1, 2, \ldots, k$ and let $n_1, n_2, \ldots, n_k$ denote the respective cell frequencies. If we denote the cell probabilities by $p_1, p_2, \ldots, p_k$, a null hypothesis that completely specifies the cell probabilities is of the form

$$H_0: \quad p_1 = p_{10}, \ldots, p_k = p_{k0}$$

where $p_{10}, \ldots, p_{k0}$ are given numerical values that satisfy $p_{10} + \cdots + p_{k0} = 1$.

Chi-Square Goodness of Fit Test: Comparison of the Observed to the Expected

From Chapter 5 recall that if the probability of an event is p, then the expected number of occurrences of the event in n trials is np. Therefore, once the cell probabilities are specified, the expected cell frequencies can be readily computed by **multiplying** these probabilities by the sample size n. A goodness of fit test attempts to determine if a conspicuous discrepancy exists between the observed cell frequencies and those expected under H_0. (See Table 4.)

TABLE 4 The Basis of a Goodness of Fit Test

Cells	1	2	⋯	k	Total
Observed frequency O	n_1	n_2	⋯	n_k	n
Probability under H_0	p_{10}	p_{20}	⋯	p_{k0}	1
Expected frequency E under H_0	np_{10}	np_{20}	⋯	np_{k0}	n

TABLE 4 The Basis of a Goodness of Fit Test

Cells	1	2	$\cdots$	k	Total
Observed frequency O	n_1	n_2	$\cdots$	n_k	n
Probability under H_0	p_{10}	p_{20}	$\cdots$	p_{k0}	1
Expected frequency E under H_0	np_{10}	np_{20}	$\cdots$	np_{k0}	n

Table 4 (p. 532)

The Basis of a Goodness of Fit Test

Chi-Square Goodness of Fit Formula: Comparison of the Observed to the Expected

TABLE 4 The Basis of a Goodness of Fit Test

Cells	1	2	⋯	k	Total
Observed frequency O	n_1	n_2	⋯	n_k	n
Probability under H_0	p_{10}	p_{20}	⋯	p_{k0}	1
Expected frequency E under H_0	np_{10}	np_{20}	⋯	np_{k0}	n

A useful measure for the overall discrepancy between the observed and expected frequencies is given by the **chi-square** or χ^2 **statistic**.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{\text{cells}} \frac{(O-E)^2}{E}$$

where O and E symbolize an observed frequency and the corresponding expected frequency. The discrepancy in each cell is measured by the squared difference between the observed and the expected frequencies divided by the expected frequency. The χ^2 measure is the sum of these quantities for all cells.

Chi-Square Goodness of Fit Depends on the Relative Difference to the Expected

A useful measure for the overall discrepancy between the observed and expected frequencies is given by the **chi-square** or χ^2 **statistic**.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{\text{cells}} \frac{(O-E)^2}{E}$$

where O and E symbolize an observed frequency and the corresponding expected frequency. The discrepancy in each cell is measured by the squared difference between the observed and the expected frequencies divided by the expected frequency. The χ^2 measure is the sum of these quantities for all cells.

Notice that the term $(O-E)^2/E$ depends on more than just the difference $O-E$ between the observed and expected value. For instance $(95-100)^2/100 = .25$ is much smaller than $(5-10)^2/10 = 2.5$ although the difference is 5 in both cases. It is the relative difference between O and E, not just the actual difference $O-E$, that determines the contribution of $(O-E)^2/E$ to the value of χ^2.

Pearson's Chi-Square Test for Goodness of Fit (Expected Frequency in Each Cell at Least 5)

A description of the χ^2 table (Appendix B, Table 6) and how to obtain an upper α point, χ^2_α, is given in Section 4 of Chapter 9.

Pearson's x^2 Test for Goodness of Fit (Based on Large n)

Null hypothesis

$$H_0: p_1 = p_{10}, \ldots, p_k = p_{k0}$$

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{\text{cells}} \frac{(O-E)^2}{E}$$

Rejection region

$$\chi^2 \geq \chi^2_\alpha$$

where χ^2_α is the upper α point of the χ^2 distribution with

$$\text{d.f.} = k - 1 = (\text{Number of cells}) - 1$$

It should be remembered that Pearson's χ^2 test is an approximate test that is valid only for large samples. As a rule of thumb, n should be large enough so that the expected frequency of each cell is at least 5.

Pearson's χ^2 Test for Goodness of Fit (Based on Large n)

Null hypothesis

$$H_0: \quad p_1 = p_{10}, \ldots, p_k = p_{k0}$$

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{\text{cells}} \frac{(O - E)^2}{E}$$

Rejection region

$$\chi^2 \geq \chi^2_\alpha$$

where χ^2_α is the upper α point of the χ^2 distribution with

$$\text{d.f.} = k - 1 = (\text{Number of cells}) - 1$$

Box on Page 533

Pearson's χ^2 Test for Goodness of Fit (Based on Large n)

Example: Chi-Square Goodness of Fit Test as Applied to Genetic Model

Example 4

χ^2 Goodness of Fit for a Genetic Model

Referring to Example 1, test the goodness of fit of the genetic model to the data in Table 1. Take $\alpha = .05$.

SOLUTION

Following the structure of Table 4, the computations for the χ^2 statistic are exhibited in Table 5 where the last line gives the calculation

$$\sum_{\text{cells}} \frac{(O-E)^2}{E} = \frac{(18-25)^2}{25} + \frac{(55-50)^2}{50} + \frac{(27-25)^2}{25}$$

$$= 1.96 + .50 + .16 = 2.62$$

TABLE 5 The χ^2 Goodness of Fit Test for the Data in Table 1

Cell	*A*	*B*	*C*	Total
Observed frequency O	18	55	27	100
Probability under H_0	.25	.50	.25	1.0
Expected frequency E	25	50	25	100
$\frac{(O-E)^2}{E}$	1.96	.50	.16	$2.62 = \chi^2$ d.f. = 2

TABLE 5 The χ^2 Goodness of Fit Test for the Data in Table 1

Cell	A	B	C	Total
Observed frequency O	18	55	27	100
Probability under H_0	.25	.50	.25	1.0
Expected frequency E	25	50	25	100
$\frac{(O - E)^2}{E}$	1.96	.50	.16	$2.62 = \chi^2$ d.f. = 2

Table 5 (p. 533)

The χ^2 Goodness of Fit Test for the Data in Table 1

Example: Chi-Square Goodness of Fit Test as Applied to Genetic Model

Cell	*A*	*B*	*C*	Total
Observed frequency O	18	55	27	100
Probability under H_0	.25	.50	.25	1.0
Expected frequency E	25	50	25	100
$\frac{(O-E)^2}{E}$	1.96	.50	.16	$2.62 = \chi^2$ d.f. = 2

We use the χ^2 statistic with rejection region $R: \chi^2 \geq 5.99$ since $\chi^2_{.05} = 5.99$ with d.f. = 2 (Appendix B, Table 6). Because the observed $\chi^2 = 2.62$ is smaller than this value, the null hypothesis is not rejected at $\alpha = .05$. We conclude that the data in Table 1 do not contradict the genetic model.

The P-value = .270, shown in Figure 1, confirms the conclusion.

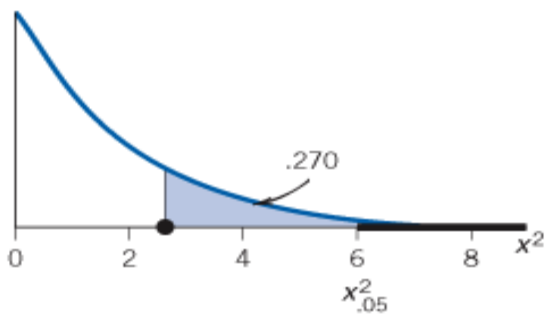


Figure 1 The P-value for Example 4 based on the χ^2 distribution with 2 d.f.

Chi-Square Distribution Defined by Degrees of Freedom

The χ^2 statistic measures the overall discrepancy between the observed frequencies and those expected under a given null hypothesis. Example 4 demonstrates its application when the frequency counts arise from a single random sample and the categories refer to only one characteristic—namely, the genotype of the offspring. Basically, the same principle extends to testing hypotheses with more complex types of categorical data such as the contingency tables illustrated in Examples 2 and 3. In preparation for these developments, we state two fundamental properties of the χ^2 statistic:

Properties of Pearson's χ^2 Statistic

1. **Additivity:**
   If χ^2 statistics are computed from independent samples, then their sum is also a χ^2 statistic whose d.f. equals the sum of the d.f.'s of the components.
2. **Loss of d.f. due to estimation of parameters:**
   If H_0 does not completely specify the cell probabilities, then some parameters have to be estimated in order to obtain the expected cell frequencies. In that case, the d.f. of χ^2 is reduced by the number of parameters estimated.

$$\text{d.f. of } \chi^2 = (\text{No. of cells}) - 1 - (\text{No. of parameters estimated})$$

Properties of Pearson's χ^2 Statistic

1. **Additivity:** If χ^2 statistics are computed from independent samples, then their sum is also a χ^2 statistic whose d.f. equals the sum of the d.f.'s of the components.
2. **Loss of d.f. due to estimation of parameters:** If H_0 does not completely specify the cell probabilities, then some parameters have to be estimated in order to obtain the expected cell frequencies. In that case, the d.f. of χ^2 is reduced by the number of parameters estimated.

$$\text{d.f. of } \chi^2 = (\text{No. of cells}) - 1 - (\text{No. of parameters estimated})$$

Box on Page 534

Properties of Pearson's χ^2 Statistic

Test of Homogeneity: Contingency Table with One Margin Fixed

3. Contingency Table With One Margin Fixed (Test of Homogeneity)

From each population, we draw a random sample of a predetermined sample size and classify each response in categories. These data form a two-way contingency table where one classification refers to the populations and the other to the response under study. Our objective is to test whether the populations are alike, or **homogeneous**, with respect to cell probabilities. To do so, we will determine if the observed proportions in each response category are nearly the same for all populations.

Let us pursue our development of the χ^2 **test of homogenity** with the data of Table 2.

Example 5

Developing a χ^2 Test to Compare Two Diets

Referring to Example 2, test the null hypothesis that there is no difference between the quality of the two diets.

SOLUTION

For ease of reference, the data in Table 2 are reproduced in Table 6. Here the populations correspond to the two diets and the response is recorded in three categories. The row totals 80 and 70 are the fixed sample sizes.

TABLE 6 A 2 × 3 Contingency Table with Fixed Row Totals

	Excellent	Average	Poor	Total
Diet *A*	37	24	19	80
Diet *B*	17	33	20	70
Total	54	57	39	150

Test of Homogeneity: Contingency Table with One Margin Fixed

TABLE 6 A 2 × 3 Contingency Table with Fixed Row Totals

	Excellent	Average	Poor	Total
Diet *A*	37	24	19	80
Diet *B*	17	33	20	70
Total	54	57	39	150

We have already formulated the null hypothesis of "homogeneity" or "no difference between the diets" as [see Table 2(b)]

$$H_0: \quad p_{A1} = p_{B1} \quad p_{A2} = p_{B2} \quad p_{A3} = p_{B3}$$

If we denote these common probabilities under H_0 by p_1, p_2, and p_3, respectively, the expected cell frequencies in each row would be obtained by multiplying these probabilities by the sample size. In particular, the expected frequencies in the first row are $80p_1$, $80p_2$, and $80p_3$, and those in the second row are $70p_1$, $70p_2$, and $70p_3$. However, the p_i's are not specified by H_0. Therefore, we have to estimate these parameters in order to obtain the numerical values of the expected frequencies.

TABLE 6 A 2 × 3 Contingency Table with Fixed Row Totals

	Excellent	Average	Poor	Total
Diet *A*	37	24	19	80
Diet *B*	17	33	20	70
Total	54	57	39	150

Table 6 (p. 537)

A 2 X 3 Contingency Table with Fixed Row Totals

Test of Homogeneity: Contingency Table with One Margin Fixed

The column totals 54, 57, and 39 in Table 6 are the frequency counts of the three response categories in the combined sample of size 150. Under H_0, the estimated probabilities are

$$\hat{p}_1 = \frac{54}{150} \quad \hat{p}_2 = \frac{57}{150} \quad \hat{p}_3 = \frac{39}{150}$$

We use these estimates to calculate the expected frequencies in the first row as

$$80 \times \frac{54}{150} = \frac{80 \times 54}{150} = 28.8 \quad \frac{80 \times 57}{150} = 30.4 \quad \frac{80 \times 39}{150} = 20.8$$

and similarly for the second row. Referring to Table 6, notice the interesting pattern in these calculations:

$$\text{Expected cell frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Chi-Square Test of Homogeneity: Add Each Chi-Square Value for Each Cell

$$\text{Expected cell frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Table 7(a) presents the observed frequencies O along with the expected frequencies E. The latter are given in parentheses. Table 7(b) computes the discrepancy measure $(O-E)^2/E$ for the individual cells. Adding these over all the cells, we obtain the value of the χ^2 statistic.

$$\begin{aligned} x^2 &= 2.335 + 1.347 + .156 + \\ &\quad 2.668 + 1.540 + .178 = 8.224 \end{aligned}$$

TABLE 7(a) The Observed and Expected Frequencies of the Data in Table 6

	Excellent	Average	Poor
Diet *A*	37 (28.8)	24 (30.4)	19 (20.8)
Diet *B*	17 (25.2)	33 (26.6)	37 (18.2)

TABLE 7(b) The Values of $(O - E)^2 / E$

	Excellent	Average	Poor	
Diet *A*	2.335	1.347	.156	
Diet *B*	2.668	1.540	.178	
				$8.224 = \chi^2$

TABLE 7(a) The Observed and Expected Frequencies of the Data in Table 6

	Excellent	Average	Poor
Diet *A*	37 (28.8)	24 (30.4)	19 (20.8)
Diet *B*	17 (25.2)	33 (26.6)	20 (18.2)

Table 7a (p. 538)

The Observed and Expected Frequencies of the Data in Table 6

Compare Computed Chi-Square Against the Critical Value from the Table

TABLE 7(b) The Values of $(O - E)^2 / E$

	Excellent	Average	Poor	
Diet A	2.335	1.347	.156	
Diet B	2.668	1.540	.178	
				$8.224 = \chi^2$

In order to determine the degrees of freedom, we employ the properties of the χ^2 statistic stated in Section 2. Our χ^2 statistic has been computed from two independent samples; each contributes $3 - 1 = 2$ d.f. because there are three categories. The added d.f. $= 2 + 2 = 4$ must now be reduced by the number of parameters we have estimated. Since p_1, p_2, and p_3 satisfy the relation $p_1 + p_2 + p_3 = 1$, there are really two undetermined parameters among them. Therefore, our χ^2 statistic has d.f. $= 4 - 2 = 2$.

With d.f. $= 2$, the tabulated upper 5% point of χ^2 is 5.99 (Appendix B, Table 6). Since the observed $\chi^2 = 8.224$ is larger, the null hypothesis is rejected at $\alpha = .05$. A computer calculation gives the P-value .016. Therefore, a significant difference between the quality of the two diets is indicated by the data. Having obtained a significant χ^2, we should now examine Tables 7(a) and 7(b) and try to locate the source of the significance. We find that large contributions to χ^2 come from the "excellent" category, where the relative frequency is $37/80$, or 46%, for diet A and $17/70$, or 24%, for diet B. These data indicate that diet A is superior.

Interpretation of Statistical Distance

Remark: (Interpretation of statistical distance) Result 4.7 provides an interpretation of a squared statistical distance. When $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

is the squared statistical distance from $\mathbf{X}$ to the population mean vector $\boldsymbol{\mu}$. If one component has a much larger variance than another, it will contribute less to the squared distance. Moreover, two highly correlated random variables will contribute less than two variables that are nearly uncorrelated. Essentially, the use of the inverse of the covariance matrix, (1) standardizes all of the variables and (2) eliminates the effects of correlation. From the proof of Result 4.7,

$$(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = Z_1^2 + Z_2^2 + \cdots + Z_p^2$$

Linear Combination of Vectors (Not a Variable Written as Combination of Other Random Univariate Variables)

In terms of $\Sigma^{-\frac{1}{2}}$ (see (2-22)), $\mathbf{Z} = \Sigma^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$ has a $N_p(\mathbf{0}, \mathbf{I}_p)$ distribution, and

$$(\mathbf{X} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}) = (\mathbf{X} - \boldsymbol{\mu})'\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$$
$$= \mathbf{Z}'\mathbf{Z} = Z_1^2 + Z_2^2 + \cdots + Z_p^2$$

The squared statistical distance is calculated as if, first, the random vector $\mathbf{X}$ were transformed to p independent standard normal random variables and then the usual squared distance, the sum of the squares of the variables, were applied.

Next, consider the linear combination of vector random variables

$$c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \cdots + c_n\mathbf{X}_n = \underset{(p\times n)}{[\mathbf{X}_1 \mid \mathbf{X}_2 \mid \cdots \mid \mathbf{X}_n]} \underset{(n\times 1)}{\mathbf{c}} \tag{4-10}$$

This linear combination differs from the linear combinations considered earlier in that it defines a $p \times 1$ *vector* random variable that is a linear combination of vectors. Previously, we discussed a *single* random variable that could be written as a linear combination of other univariate random variables.

X_1, X_2, etc. Mutually Independent With X_j Distributed as $N_p(\mu_j,\Sigma)$

Result 4.8. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be mutually independent with $\mathbf{X}_j$ distributed as $N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. (Note that each $\mathbf{X}_j$ has the *same* covariance matrix $\boldsymbol{\Sigma}$.) Then

$$\mathbf{V}_1 = c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \cdots + c_n\mathbf{X}_n$$

is distributed as $N_p\left(\sum_{j=1}^{n} c_j\boldsymbol{\mu}_j, \left(\sum_{j=1}^{n} c_j^2\right)\boldsymbol{\Sigma}\right)$. Moreover, $\mathbf{V}_1$ and $\mathbf{V}_2 = b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \cdots + b_n\mathbf{X}_n$ are jointly multivariate normal with covariance matrix

$$\begin{bmatrix} \left(\sum_{j=1}^{n} c_j^2\right)\boldsymbol{\Sigma} & (\mathbf{b}'\mathbf{c})\,\boldsymbol{\Sigma} \\ (\mathbf{b}'\mathbf{c})\,\boldsymbol{\Sigma} & \left(\sum_{j=1}^{n} b_j^2\right)\boldsymbol{\Sigma} \end{bmatrix}$$

Consequently, $\mathbf{V}_1$ and $\mathbf{V}_2$ are independent if $\mathbf{b}'\mathbf{c} = \sum_{j=1}^{n} c_j b_j = 0$.

Proof: The np Component Vector

Proof. By Result 4.5(c), the np component vector

$$[X_{11}, \ldots, X_{1p}, X_{21}, \ldots, X_{2p}, \ldots, X_{np}] = [\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'] = \underset{(1\times np)}{\mathbf{X}'}$$

is multivariate normal. In particular, $\underset{(np\times 1)}{\mathbf{X}}$ is distributed as $N_{np}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{x}})$, where

$$\underset{(np\times 1)}{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_n \end{bmatrix} \quad \text{and} \quad \underset{(np\times np)}{\boldsymbol{\Sigma}_{\mathbf{x}}} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix}$$

Proof: The np Component Vector

The choice

$$\underset{(2p\times np)}{\mathbf{A}} = \begin{bmatrix} c_1\mathbf{I} & c_2\mathbf{I} & \cdots & c_n\mathbf{I} \\ b_1\mathbf{I} & b_2\mathbf{I} & \cdots & b_n\mathbf{I} \end{bmatrix}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix, gives

$$\mathbf{AX} = \begin{bmatrix} \sum_{j=1}^{n} c_j\mathbf{X}_j \\ \sum_{j=1}^{n} b_j\mathbf{X}_j \end{bmatrix} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix}$$

and $\mathbf{AX}$ is normal $N_{2p}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}')$ by Result 4.3. Straightforward block multiplication shows that $\mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}'$ has the first block diagonal term

$$[c_1\boldsymbol{\Sigma}, c_2\boldsymbol{\Sigma}, \ldots, c_n\boldsymbol{\Sigma}][c_1\mathbf{I}, c_2\mathbf{I}, \ldots, c_n\mathbf{I}]' = \left(\sum_{j=1}^{n} c_j^2\right)\boldsymbol{\Sigma}$$

The off-diagonal term is

$$[c_1\boldsymbol{\Sigma}, c_2\boldsymbol{\Sigma}, \ldots, c_n\boldsymbol{\Sigma}][b_1\mathbf{I}, b_2\mathbf{I}, \ldots, b_n\mathbf{I}]' = \left(\sum_{j=1}^{n} c_j b_j\right)\boldsymbol{\Sigma}$$

V_1 and V_2 Are Independent, i.e. Zero Correlation is the Coefficient Vectors b and c Are Perpendicular

The off-diagonal term is

$$[c_1\boldsymbol{\Sigma}, c_2\boldsymbol{\Sigma}, \ldots, c_n\boldsymbol{\Sigma}][b_1\mathbf{I}, b_2\mathbf{I}, \ldots, b_n\mathbf{I}]' = \left(\sum_{j=1}^{n} c_j b_j\right)\boldsymbol{\Sigma}$$

This term is the covariance matrix for $\mathbf{V}_1, \mathbf{V}_2$. Consequently, when $\sum_{j=1}^{n} c_j b_j = \mathbf{b}'\mathbf{c} = 0$, so that $\left(\sum_{j=1}^{n} c_j b_j\right)\boldsymbol{\Sigma} = \underset{(p\times p)}{\mathbf{0}}$, $\mathbf{V}_1$ and $\mathbf{V}_2$ are independent by Result 4.5(b). ■

For sums of the type in (4-10), the property of zero correlation is equivalent to requiring the coefficient vectors $\mathbf{b}$ and $\mathbf{c}$ to be perpendicular.

Linear Combinations Random Vectors

Example 4.8 (Linear combinations of random vectors) Let $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$, and $\mathbf{X}_4$ be independent and identically distributed 3×1 random vectors with

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{bmatrix}$$

We first consider a linear combination $\mathbf{a}'\mathbf{X}_1$ of the three components of $\mathbf{X}_1$. This is a random variable with mean

$$\mathbf{a}'\boldsymbol{\mu} = 3a_1 - a_2 + a_3$$

and variance

$$\mathbf{a}'\boldsymbol{\Sigma}\,\mathbf{a} = 3a_1^2 + a_2^2 + 2a_3^2 - 2a_1a_2 + 2a_1a_3$$

That is, a linear combination $\mathbf{a}'\mathbf{X}_1$ of the components of a random vector is a single random variable consisting of a sum of terms that are each a constant times a variable. This is very different from a linear combination of random vectors, say,

$$c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + c_3\mathbf{X}_3 + c_4\mathbf{X}_4$$

Find the Mean and Covariance Matrix for Each Linear Combination of Vectors and The Covariance Between Them

which is itself a random vector. Here each term in the sum is a constant times a random vector.

Now consider two linear combinations of random vectors

$$\frac{1}{2}\mathbf{X}_1 + \frac{1}{2}\mathbf{X}_2 + \frac{1}{2}\mathbf{X}_3 + \frac{1}{2}\mathbf{X}_4$$

and

$$\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 - 3\mathbf{X}_4$$

Find the mean vector and covariance matrix for each linear combination of vectors and also the covariance between them.

By Result 4.8 with $c_1 = c_2 = c_3 = c_4 = 1/2$, the first linear combination has mean vector

$$(c_1 + c_2 + c_3 + c_4)\boldsymbol{\mu} = 2\boldsymbol{\mu} = \begin{bmatrix} 6 \\ -2 \\ 2 \end{bmatrix}$$

Covariance Matrix for Combinations Random Vectors

and covariance matrix

$$(c_1^2 + c_2^2 + c_3^2 + c_4^2)\boldsymbol{\Sigma} = 1 \times \boldsymbol{\Sigma} = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{bmatrix}$$

For the second linear combination of random vectors, we apply Result 4.8 with $b_1 = b_2 = b_3 = 1$ and $b_4 = -3$ to get mean vector

$$(b_1 + b_2 + b_3 + b_4)\boldsymbol{\mu} = 0\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

and covariance matrix

$$(b_1^2 + b_2^2 + b_3^2 + b_4^2)\boldsymbol{\Sigma} = 12 \times \boldsymbol{\Sigma} = \begin{bmatrix} 36 & -12 & 12 \\ -12 & 12 & 0 \\ 12 & 0 & 24 \end{bmatrix}$$

Zero Covariance, With Two Linear Combinations of Vectors Independent

Finally, the covariance matrix for the two linear combinations of random vectors is

$$(c_1b_1 + c_2b_2 + c_3b_3 + c_4b_4)\boldsymbol{\Sigma} = 0\boldsymbol{\Sigma} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Every component of the first linear combination of random vectors has zero covariance with every component of the second linear combination of random vectors.

If, in addition, each $\mathbf{X}$ has a trivariate normal distribution, then the two linear combinations have a joint six-variate normal distribution, and the two linear combinations of vectors are independent. ■

Sampling from Multivariate Normal Distribution: Multivariate Normal Likelihood

4.3 Sampling from a Multivariate Normal Distribution and Maximum Likelihood Estimation

We discussed sampling and selecting random samples briefly in Chapter 3. In this section, we shall be concerned with samples from a multivariate normal population—in particular, with the sampling distribution of $\bar{\mathbf{X}}$ and $\mathbf{S}$.

The Multivariate Normal Likelihood

Let us assume that the $p \times 1$ vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ represent a random sample from a multivariate normal population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Since $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are mutually independent and each has distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the joint density function of all the observations is the product of the marginal normal densities:

$$\left\{\begin{array}{c}\text{Joint density} \\ \text{of } \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\end{array}\right\} = \prod_{j=1}^{n} \left\{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j-\boldsymbol{\mu})/2}\right\}$$

$$= \frac{1}{(2\pi)^{np/2}} \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} e^{-\sum_{j=1}^{n}(\mathbf{x}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j-\boldsymbol{\mu})/2} \qquad (4\text{-}11)$$

Maximum Likelihood Estimation: Values That Maximize the Joint Density

When the numerical values of the observations become available, they may be substituted for the $\mathbf{x}_j$ in Equation (4-11). The resulting expression, now considered as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the fixed set of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, is called the *likelihood*.

Many good statistical procedures employ values for the population parameters that "best" explain the observed data. One meaning of *best* is to select the parameter values that *maximize* the joint density evaluated at the observations. This technique is called *maximum likelihood estimation*, and the maximizing parameter values are called *maximum likelihood estimates*.

At this point, we shall consider maximum likelihood estimation of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for a multivariate normal population. To do so, we take the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ as fixed and consider the joint density of Equation (4-11) evaluated at these values. The result is the likelihood function. In order to simplify matters, we rewrite the likelihood function in another form. We shall need some additional properties for the trace of a square matrix. (The trace of a matrix is the sum of its diagonal elements, and the properties of the trace are discussed in Definition 2A.28 and Result 2A.12.)

Maximum Likelihood Symmetric Matrix and k x 1 vector

Result 4.9. Let $\mathbf{A}$ be a $k \times k$ symmetric matrix and $\mathbf{x}$ be a $k \times 1$ vector. Then

(a) $\mathbf{x}'\mathbf{A}\mathbf{x} = \operatorname{tr}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \operatorname{tr}(\mathbf{A}\mathbf{x}\mathbf{x}')$

(b) $\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{k} \lambda_i$, where the λ_i are the eigenvalues of $\mathbf{A}$.

Maximum Likelihood Estimation of μ and Σ

Maximum Likelihood Estimation of μ and Σ

The next result will eventually allow us to obtain the maximum likelihood estimators of μ and Σ.

Result 4.10. Given a $p \times p$ symmetric positive definite matrix $\mathbf{B}$ and a scalar $b > 0$, it follows that

$$\frac{1}{|\Sigma|^b} e^{-\operatorname{tr}(\Sigma^{-1}\mathbf{B})/2} \leq \frac{1}{|\mathbf{B}|^b} (2b)^{pb} e^{-bp}$$

for all positive definite $\underset{(p\times p)}{\Sigma}$, with equality holding only for $\Sigma = (1/2b)\mathbf{B}$.

Generalized Variance Determines the Peakedness of the Likelihood Function

$$L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \text{constant} \times (\text{generalized variance})^{-n/2} \tag{4-19}$$

The generalized variance determines the "peakedness" of the likelihood function and, consequently, is a natural measure of variability when the parent population is multivariate normal.

Maximum likelihood estimators possess an *invariance property*. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$, and consider estimating the parameter $h(\boldsymbol{\theta})$, which is a function of $\boldsymbol{\theta}$. Then the *maximum likelihood estimate* of

$$\underset{\text{(a function of } \boldsymbol{\theta})}{h(\boldsymbol{\theta})} \quad \text{is given by} \quad \underset{\text{(same function of } \hat{\boldsymbol{\theta}})}{h(\hat{\boldsymbol{\theta}})} \tag{4-20}$$

Properties Maximum Likelihood

(See [1] and [15].) For example,

1. The maximum likelihood estimator of $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}$, where $\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}}$ and $\hat{\boldsymbol{\Sigma}} = ((n-1)/n)\mathbf{S}$ are the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.
2. The maximum likelihood estimator of $\sqrt{\sigma_{ii}}$ is $\sqrt{\hat{\sigma}_{ii}}$, where

$$\hat{\sigma}_{ii} = \frac{1}{n}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2$$

is the maximum likelihood estimator of $\sigma_{ii} = \mathrm{Var}(X_i)$.

Sufficient Statistics x-bar and s

Sufficient Statistics

From expression (4-15), the joint density depends on the whole set of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ only through the sample mean $\bar{\mathbf{x}}$ and the sum-of-squares-and-cross-products matrix $\sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' = (n-1)\mathbf{S}$. We express this fact by saying that $\bar{\mathbf{x}}$ and $(n-1)\mathbf{S}$ (or $\mathbf{S}$) are *sufficient statistics*:

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a multivariate normal population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Then

$\bar{\mathbf{X}}$ and $\mathbf{S}$ are *sufficient statistics* (4-21)

Information About μ and Σ In the Data Matrix is Contained in x-bar and s Regardless of Sample Size n

The importance of sufficient statistics for normal populations is that all of the information about $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the data matrix $\mathbf{X}$ is contained in $\bar{\mathbf{x}}$ and $\mathbf{S}$, regardless of the sample size n. This generally is not true for nonnormal populations. Since many multivariate techniques begin with sample means and covariances, it is prudent to check on the *adequacy* of the multivariate normal assumption. (See Section 4.6.) If the data cannot be regarded as multivariate normal, techniques that depend solely on $\bar{\mathbf{x}}$ and $\mathbf{S}$ may be ignoring other useful sample information.

Sampling Distribution of X-bar and S

4.4 The Sampling Distribution of $\bar{X}$ and S

The tentative assumption that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ constitute a random sample from a normal population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ completely determines the sampling distributions of $\bar{\mathbf{X}}$ and $\mathbf{S}$. Here we present the results on the sampling distributions of $\bar{\mathbf{X}}$ and $\mathbf{S}$ by drawing a parallel with the familiar univariate conclusions.

In the univariate case ($p = 1$), we know that $\bar{X}$ is normal with mean μ = (population mean) and variance

$$\frac{1}{n}\sigma^2 = \frac{\text{population variance}}{\text{sample size}}$$

The result for the multivariate case ($p \geq 2$) is analogous in that $\bar{\mathbf{X}}$ has a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $(1/n)\boldsymbol{\Sigma}$.

Sample Variance Distribution as σ^2 Times A Chi-Square Variable Having n-1 Degrees of Freedom

For the sample variance, recall that $(n-1)s^2 = \sum_{j=1}^{n} (X_j - \bar{X})^2$ is distributed as σ^2 times a chi-square variable having $n-1$ degrees of freedom (d.f.). In turn, this chi-square is the distribution of a sum of squares of independent standard normal random variables. That is, $(n-1)s^2$ is distributed as $\sigma^2(Z_1^2 + \cdots + Z_{n-1}^2) = (\sigma Z_1)^2 + \cdots + (\sigma Z_{n-1})^2$. The individual terms σZ_i are independently distributed as $N(0, \sigma^2)$. It is this latter form that is suitably generalized to the basic sampling distribution for the sample covariance matrix.

Sampling Distribution of the Sample Covariance Matrix is Called the Wishart Distribution

The sampling distribution of the sample covariance matrix is called the *Wishart distribution*, after its discoverer; it is defined as the sum of independent products of multivariate normal random vectors. Specifically,

$$W_m(\cdot \mid \Sigma) = \text{Wishart distribution with } m \text{ d.f.} \tag{4-22}$$

$$= \text{distribution of } \sum_{j=1}^{m} \mathbf{Z}_j \mathbf{Z}_j'$$

where the $\mathbf{Z}_j$ are each independently distributed as $N_p(\mathbf{0}, \Sigma)$.

We summarize the sampling distribution results as follows:

Wishart Distribution

We summarize the sampling distribution results as follows:

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample of size n from a p-variate *normal* distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then

1. $\bar{\mathbf{X}}$ is distributed as $N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$.
2. $(n-1)\mathbf{S}$ is distributed as a Wishart random matrix with $n-1$ d.f. (4-23)
3. $\bar{\mathbf{X}}$ and $\mathbf{S}$ are independent.

Because $\boldsymbol{\Sigma}$ is unknown, the distribution of $\bar{\mathbf{X}}$ cannot be used directly to make inferences about $\boldsymbol{\mu}$. However, $\mathbf{S}$ provides independent information about $\boldsymbol{\Sigma}$, and the distribution of $\mathbf{S}$ does not depend on $\boldsymbol{\mu}$. This allows us to construct a statistic for making inferences about $\boldsymbol{\mu}$, as we shall see in Chapter 5.

For the present, we record some further results from multivariable distribution theory. The following properties of the Wishart distribution are derived directly from its definition as a sum of the independent products, $\mathbf{Z}_j\mathbf{Z}_j'$. Proofs can be found in [1].

Properties of the Wishart Distribution

Properties of the Wishart Distribution

1. If $\mathbf{A}_1$ is distributed as $W_{m_1}(\mathbf{A}_1 \mid \boldsymbol{\Sigma})$ independently of $\mathbf{A}_2$, which is distributed as $W_{m_2}(\mathbf{A}_2 \mid \boldsymbol{\Sigma})$, then $\mathbf{A}_1 + \mathbf{A}_2$ is distributed as $W_{m_1+m_2}(\mathbf{A}_1 + \mathbf{A}_2 \mid \boldsymbol{\Sigma})$. That is, the degrees of freedom add. (4-24)
2. If $\mathbf{A}$ is distributed as $W_m(\mathbf{A} \mid \boldsymbol{\Sigma})$, then $\mathbf{CAC}'$ is distributed as $W_m(\mathbf{CAC}' \mid \mathbf{C\Sigma C}')$.

Although we do not have any particular need for the probability density function of the Wishart distribution, it may be of some interest to see its rather complicated form. The density does not exist unless the sample size n is greater than the number of variables p. When it does exist, its value at the positive definite matrix $\mathbf{A}$ is

$$w_{n-1}(\mathbf{A} \mid \boldsymbol{\Sigma}) = \frac{|\mathbf{A}|^{(n-p-2)/2} e^{-\mathrm{tr}[\mathbf{A}\boldsymbol{\Sigma}^{-1}]/2}}{2^{p(n-1)/2} \pi^{p(p-1)/4} |\boldsymbol{\Sigma}|^{(n-1)/2} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n-i)\right)}, \quad \mathbf{A} \text{ positive definite} \tag{4-25}$$

where $\Gamma(\cdot)$ is the gamma function. (See [1] and [11].)

Large-Sample Behavior of X-bar and S

4.5 Large-Sample Behavior of $\bar{X}$ and S

Suppose the quantity X is determined by a large number of independent causes $V_1, V_2, \ldots, V_n$, where the random variables V_i representing the causes have approximately the same variability. If X is the sum

$$X = V_1 + V_2 + \cdots + V_n$$

then the central limit theorem applies, and we conclude that X has a distribution that is nearly normal. This is true for virtually any parent distribution of the V_i's, provided that n is large enough.

The univariate central limit theorem also tells us that the sampling distribution of the sample mean, $\bar{X}$ for a large sample size is nearly normal, whatever the form of the underlying population distribution. A similar result holds for many other important univariate statistics.

It turns out that certain multivariate statistics, like $\bar{\mathbf{X}}$ and $\mathbf{S}$, have large-sample properties analogous to their univariate counterparts. As the sample size is increased without bound, certain regularities govern the sampling variation in $\bar{\mathbf{X}}$ and $\mathbf{S}$, irrespective of the form of the parent population. Therefore, the conclusions presented in this section do not require multivariate normal populations. The only requirements are that the parent population, whatever its form, have a mean $\boldsymbol{\mu}$ and a finite covariance $\boldsymbol{\Sigma}$.

Law of Large Numbers

Result 4.12 (Law of large numbers). Let $Y_1, Y_2, \ldots, Y_n$ be independent observations from a population with mean $E(Y_i) = \mu$. Then

$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

converges in probability to μ as n increases without bound. That is, for any prescribed accuracy $\varepsilon > 0$, $P[-\varepsilon < \bar{Y} - \mu < \varepsilon]$ approaches unity as $n \to \infty$.

Proof: Law of Large Numbers

Proof. See [9]. ∎

As a direct consequence of the law of large numbers, which says that each $\bar{X}_i$ converges in probability to μ_i, $i = 1, 2, \ldots, p$,

$$\bar{\mathbf{X}} \text{ converges in probability to } \boldsymbol{\mu} \tag{4-26}$$

Also, each sample covariance s_{ik} converges in probability to σ_{ik}, $i, k = 1, 2, \ldots, p$, and

$$\mathbf{S} \,(\text{or } \hat{\boldsymbol{\Sigma}} = \mathbf{S}_n) \text{ converges in probability to } \boldsymbol{\Sigma} \tag{4-27}$$

Statement (4-27) follows from writing

$$\begin{aligned}(n-1)s_{ik} &= \sum_{j=1}^{n} (X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k) \\ &= \sum_{j=1}^{n} (X_{ji} - \mu_i + \mu_i - \bar{X}_i)(X_{jk} - \mu_k + \mu_k - \bar{X}_k) \\ &= \sum_{j=1}^{n} (X_{ji} - \mu_i)(X_{jk} - \mu_k) + n(\bar{X}_i - \mu_i)(\bar{X}_k - \mu_k)\end{aligned}$$

Central Limit Theorem and Normal Distribution

Letting $Y_j = (X_{ji} - \mu_i)(X_{jk} - \mu_k)$, with $E(Y_j) = \sigma_{ik}$, we see that the first term in s_{ik} converges to σ_{ik} and the second term converges to zero, by applying the law of large numbers.

The practical interpretation of statements (4-26) and (4-27) is that, with high probability, $\bar{\mathbf{X}}$ will be close to $\boldsymbol{\mu}$ and $\mathbf{S}$ will be close to $\boldsymbol{\Sigma}$ whenever the sample size is large. The statement concerning $\bar{\mathbf{X}}$ is made even more precise by a multivariate version of the central limit theorem.

Result 4.13 (The central limit theorem). Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent observations from any population with mean $\boldsymbol{\mu}$ and finite covariance $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}) \text{ has an approximate } N_p(\mathbf{0}, \boldsymbol{\Sigma}) \text{ distribution}$$

for large sample sizes. Here n should also be large relative to p.

Chi-Square Distribution Approximates Normal Distribution

Proof. See [1]. ■

The approximation provided by the central limit theorem applies to discrete, as well as continuous, multivariate populations. Mathematically, the limit is exact, and the approach to normality is often fairly rapid. Moreover, from the results in Section 4.4, we know that $\bar{\mathbf{X}}$ is exactly normally distributed when the underlying population is normal. Thus, we would expect the central limit theorem approximation to be quite good for moderate n when the parent population is nearly normal.

As we have seen, when n is large, $\mathbf{S}$ is close to $\boldsymbol{\Sigma}$ with high probability. Consequently, replacing $\boldsymbol{\Sigma}$ by $\mathbf{S}$ in the approximating normal distribution for $\bar{\mathbf{X}}$ will have a negligible effect on subsequent probability calculations.

Result 4.7 can be used to show that $n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ has a χ^2_p distribution when $\bar{\mathbf{X}}$ is distributed as $N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$ or, equivalently, when $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ has an $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ distribution. The χ^2_p distribution is *approximately* the sampling distribution of $n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ when $\bar{\mathbf{X}}$ is approximately normally distributed. Replacing $\boldsymbol{\Sigma}^{-1}$ by $\mathbf{S}^{-1}$ does not seriously affect this approximation for n large and much greater than p.

Normal Distribution When n-p Large

We summarize the major conclusions of this section as follows:

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent observations from a population with mean $\boldsymbol{\mu}$ and finite (nonsingular) covariance $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}\,(\overline{\mathbf{X}} - \boldsymbol{\mu}) \text{ is approximately } N_p(\mathbf{0}, \boldsymbol{\Sigma})$$

and (4-28)

$$n(\overline{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}) \text{ is approximately } \chi^2_p$$

for $n - p$ large.

In the next three sections, we consider ways of verifying the assumption of normality and methods for transforming nonnormal observations into observations that are approximately normal.

Assessing The Assumption of Normality

4.6 Assessing the Assumption of Normality

As we have pointed out, most of the statistical techniques discussed in subsequent chapters assume that each vector observation $\mathbf{X}_j$ comes from a multivariate normal distribution. On the other hand, in situations where the sample size is large and the techniques depend solely on the behavior of $\bar{\mathbf{X}}$, or distances involving $\bar{\mathbf{X}}$ of the form $n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$, the assumption of normality for the individual observations is less crucial. But to some degree, the *quality* of inferences made by these methods depends on how closely the true parent population resembles the multivariate normal form. It is imperative, then, that procedures exist for detecting cases where the data exhibit moderate to extreme departures from what is expected under multivariate normality.

Do the Observations of A Variable Violate the Assumptions of Normality?

We want to answer this question: Do the observations $\mathbf{X}_j$ appear to violate the assumption that they came from a normal population? Based on the properties of normal distributions, we know that all linear combinations of normal variables are normal and the contours of the multivariate normal density are ellipsoids. Therefore, we address these questions:

1. Do the marginal distributions of the elements of $\mathbf{X}$ appear to be normal? What about a few linear combinations of the components X_i?
2. Do the scatter plots of pairs of observations on different characteristics give the elliptical appearance expected from normal populations?
3. Are there any "wild" observations that should be checked for accuracy?

How to Evaluate: Univariate and Bivariate Graphs (Marginal Distributions and Scatter Plots)

It will become clear that our investigations of normality will concentrate on the behavior of the observations in one or two dimensions (for example, marginal distributions and scatter plots). As might be expected, it has proved difficult to construct a "good" overall test of joint normality in more than two dimensions because of the large number of things that can go wrong. To some extent, we must pay a price for concentrating on univariate and bivariate examinations of normality: We can never be sure that we have not missed some feature that is revealed only in higher dimensions. (It is possible, for example, to construct a nonnormal bivariate distribution with normal marginals. [See Exercise 4.8.]) Yet many types of nonnormality are often reflected in the marginal distributions and scatter plots. Moreover, for most practical work, one-dimensional and two-dimensional investigations are ordinarily sufficient. Fortunately, pathological data sets that are normal in lower dimensional representations, but nonnormal in higher dimensions, are not frequently encountered in practice.

Evaluating Normality Univariate Marginal Distributions

Evaluating the Normality of the Univariate Marginal Distributions

Dot diagrams for smaller n and histograms for $n > 25$ or so help reveal situations where one tail of a univariate distribution is much longer than the other. If the histogram for a variable X_i appears reasonably symmetric, we can check further by counting the number of observations in certain intervals. A univariate normal distribution assigns probability .683 to the interval $(\mu_i - \sqrt{\sigma_{ii}}, \mu_i + \sqrt{\sigma_{ii}})$ and probability .954 to the interval $(\mu_i - 2\sqrt{\sigma_{ii}}, \mu_i + 2\sqrt{\sigma_{ii}})$. Consequently, with a large sample size n, we expect the observed proportion $\hat{p}_{i1}$ of the observations lying in the

Evaluating Normality Using the 68-95-99% Probability Rule

interval $(\bar{x}_i - \sqrt{s_{ii}}, \bar{x}_i + \sqrt{s_{ii}})$ to be about .683. Similarly, the observed proportion $\hat{p}_{i2}$ of the observations in $(\bar{x}_i - 2\sqrt{s_{ii}}, \bar{x}_i + 2\sqrt{s_{ii}})$ should be about .954. Using the normal approximation to the sampling distribution of $\hat{p}_i$ (see [9]), we observe that either

$$|\hat{p}_{i1} - .683| > 3\sqrt{\frac{(.683)(.317)}{n}} = \frac{1.396}{\sqrt{n}}$$

or

$$|\hat{p}_{i2} - .954| > 3\sqrt{\frac{(.954)(.046)}{n}} = \frac{.628}{\sqrt{n}} \qquad (4\text{-}29)$$

would indicate departures from an assumed normal distribution for the ith characteristic. When the observed proportions are too small, parent distributions with thicker tails than the normal are suggested.

Univariate Assessment Using Q-Q Plot

Plots are always useful devices in any data analysis. Special plots called *Q–Q plots* can be used to assess the assumption of normality. These plots can be made for the marginal distributions of the sample observations on each variable. They are, in effect, plots of the sample quantile versus the quantile one would expect to observe if the observations actually were normally distributed. When the points lie very nearly along a straight line, the normality assumption remains tenable. Normality is suspect if the points deviate from a straight line. Moreover, the pattern of the deviations can provide clues about the nature of the nonnormality. Once the reasons for the non-normality are identified, corrective action is often possible. (See Section 4.8.)

To simplify notation, let $x_1, x_2, \ldots, x_n$ represent n observations on any single characteristic X_i. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ represent these observations after they are ordered according to magnitude. For example, $x_{(2)}$ is the second smallest observation and $x_{(n)}$ is the largest observation. The $x_{(j)}$'s are the sample quantiles. When the $x_{(j)}$ are distinct, exactly j observations are less than or equal to $x_{(j)}$. (This is theoretically always true when the observations are of the continuous type, which we usually assume.) The proportion j/n of the sample at or to the left of $x_{(j)}$ is often approximated by $\left(j - \frac{1}{2}\right)/n$ for analytical convenience.[1]

Q-Q Plot or Quantiles Defined by Cumulative Probability

For a standard normal distribution, the quantiles $q_{(j)}$ are defined by the relation

$$P[Z \leq q_{(j)}] = \int_{-\infty}^{q(j)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = p_{(j)} = \frac{j - \frac{1}{2}}{n} \qquad (4\text{-}30)$$

(See Table 1 in the appendix). Here $p_{(j)}$ is the probability of getting a value less than or equal to $q_{(j)}$ in a single drawing from a standard normal population.

The idea is to look at the pairs of quantiles $(q_{(j)}, x_{(j)})$ with the same associated cumulative probability $\left(j - \frac{1}{2}\right)/n$. If the data arise from a normal population, the pairs $(q_{(j)}, x_{(j)})$ will be approximately linearly related, since $\sigma q_{(j)} + \mu$ is nearly the expected sample quantile.[2]

Constructing a Q-Q Plot

Example 4.9 (Constructing a Q–Q plot) A sample of $n = 10$ observations gives the values in the following table:

Ordered observations $x_{(j)}$	Probability levels $\left(j - \frac{1}{2}\right)/n$	Standard normal quantiles $q_{(j)}$
−1.00	.05	−1.645
−.10	.15	−1.036
.16	.25	−.674
.41	.35	−.385
.62	.45	−.125
.80	.55	.125
1.26	.65	.385
1.54	.75	.674
1.71	.85	1.036
2.30	.95	1.645

Here, for example, $P[Z \leq .385] = \int_{-\infty}^{.385} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = .65$. [See (4-30).]

Q-Q Plot Illustrated

Let us now construct the Q–Q plot and comment on its appearance. The Q–Q plot for the foregoing data, which is a plot of the ordered data $x_{(j)}$ against the normal quantiles $q_{(j)}$, is shown in Figure 4.5. The pairs of points $(q_{(j)}, x_{(j)})$ lie very nearly along a straight line, and we would not reject the notion that these data are normally distributed—particularly with a sample size as small as $n = 10$.

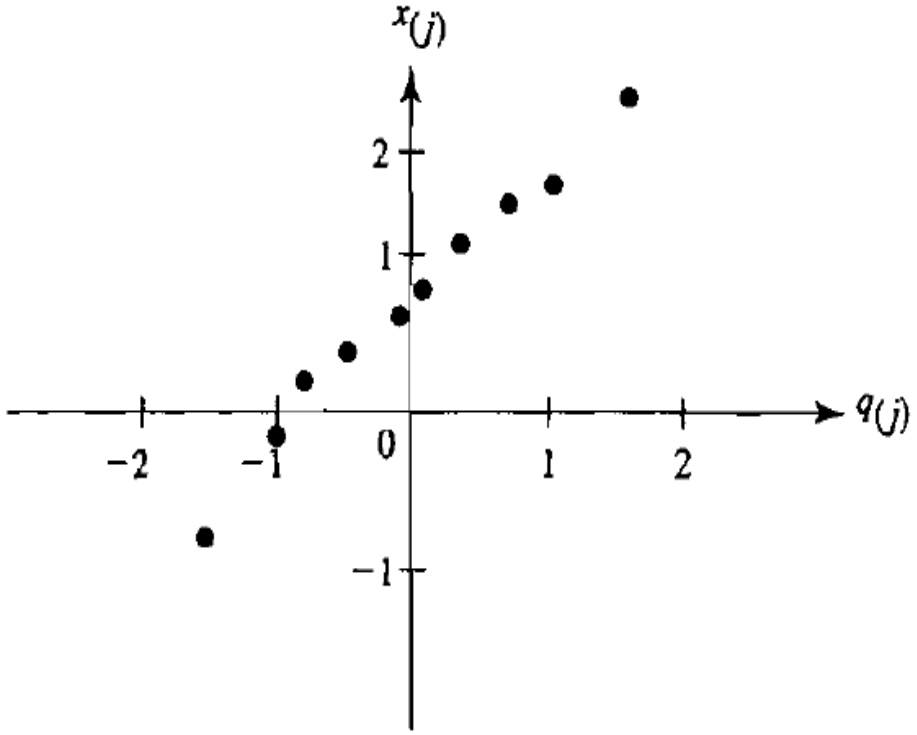


Figure 4.5 A Q–Q plot for the data in Example 4.9. ■

Steps Required for Q-Q Plots

The calculations required for Q–Q plots are easily programmed for electronic computers. Many statistical programs available commercially are capable of producing such plots.

The steps leading to a Q–Q plot are as follows:

1. Order the original observations to get $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and their corresponding probability values $\left(1 - \frac{1}{2}\right)/n, \left(2 - \frac{1}{2}\right)/n, \ldots, \left(n - \frac{1}{2}\right)/n$;
2. Calculate the standard normal quantiles $q_{(1)}, q_{(2)}, \ldots, q_{(n)}$; and
3. Plot the pairs of observations $(q_{(1)}, x_{(1)}), (q_{(2)}, x_{(2)}), \ldots, (q_{(n)}, x_{(n)})$, and examine the "straightness" of the outcome.

Q-Q Plot Requirement: Sample Size of at Least n = 20

Q–Q plots are not particularly informative unless the sample size is moderate to large—for instance, $n \geq 20$. There can be quite a bit of variability in the straightness of the *Q–Q* plot for small samples, even when the observations are known to come from a normal population.

Q-Q Plot for Radiation Data

Example 4.10 (A Q–Q plot for radiation data) The quality-control department of a manufacturer of microwave ovens is required by the federal government to monitor the amount of radiation emitted when the doors of the ovens are closed. Observations of the radiation emitted through closed doors of $n = 42$ randomly selected ovens were made. The data are listed in Table 4.1.

Table 4.1 Radiation Data (Door Closed)

Oven no.	Radiation	Oven no.	Radiation	Oven no.	Radiation
1	.15	16	.10	31	.10
2	.09	17	.02	32	.20
3	.18	18	.10	33	.11
4	.10	19	.01	34	.30
5	.05	20	.40	35	.02
6	.12	21	.10	36	.20
7	.08	22	.05	37	.20
8	.05	23	.03	38	.30
9	.08	24	.05	39	.30
10	.10	25	.15	40	.40

Sample Data Table – Radiation Data

Table 4.1 Radiation Data (Door Closed)

Oven no.	Radiation	Oven no.	Radiation	Oven no.	Radiation
1	.15	16	.10	31	.10
2	.09	17	.02	32	.20
3	.18	18	.10	33	.11
4	.10	19	.01	34	.30
5	.05	20	.40	35	.02
6	.12	21	.10	36	.20
7	.08	22	.05	37	.20
8	.05	23	.03	38	.30
9	.08	24	.05	39	.30
10	.10	25	.15	40	.40
11	.07	26	.10	41	.30
12	.02	27	.15	42	.05
13	.01	28	.09		
14	.10	29	.08		
15	.10	30	.18		

Source: Data courtesy of J. D. Cryer.

Pairs Are Assembled and Q-Q Plot Generated

In order to determine the probability of exceeding a prespecified tolerance level, a probability distribution for the radiation emitted was needed. Can we regard the observations here as being normally distributed?

A computer was used to assemble the pairs $(q_{(j)}, x_{(j)})$ and construct the Q–Q plot, pictured in Figure 4.6 on page 181. It appears from the plot that the data as a whole are not normally distributed. The points indicated by the circled locations in the figure are outliers—values that are too large relative to the rest of the observations.

For the radiation data, several observations are equal. When this occurs, those observations with like values are associated with the same normal quantile. This quantile is calculated using the average of the quantiles the tied observations would have if they all differed slightly. ■

Q-Q Plot Radiation Data

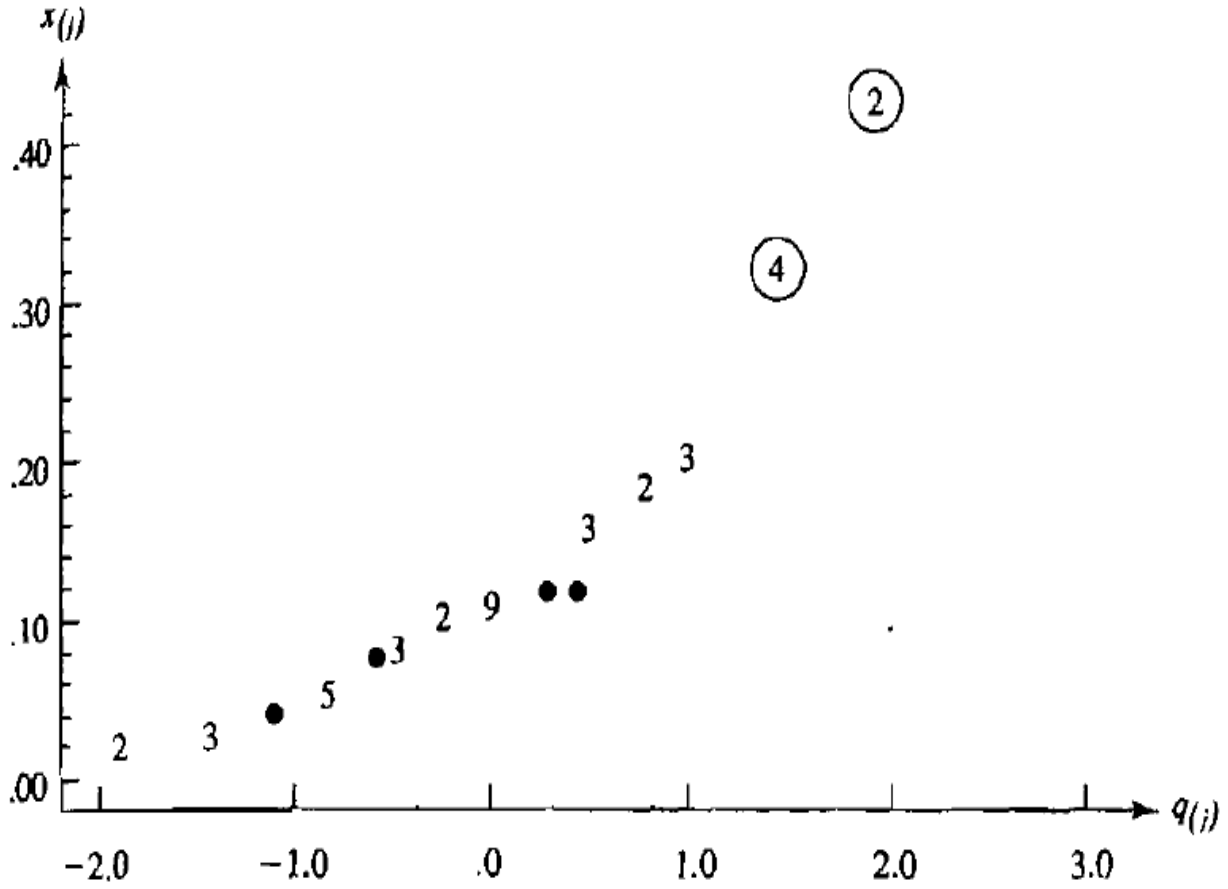


Figure 4.6 A Q–Q plot of the radiation data (door closed) from Example 4.10. (The integers in the plot indicate the number of points occupying the same location.)

Q-Q Plot Can Be Measured By Calculating the Correlation Coefficient of the Points in the Plot

The straightness of the Q–Q plot can be measured by calculating the correlation coefficient of the points in the plot. The correlation coefficient for the Q–Q plot is defined by

$$r_Q = \frac{\sum_{j=1}^{n} (x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n} (x_{(j)} - \bar{x})^2} \sqrt{\sum_{j=1}^{n} (q_{(j)} - \bar{q})^2}} \tag{4-31}$$

and a powerful test of normality can be based on it. (See [5], [10], and [12].) Formally, we reject the hypothesis of normality at level of significance α if r_Q falls *below* the appropriate value in Table 4.2.

Critical Points for the Q-Q Plot Correlation Coefficient Test for Normality

Table 4.2 Critical Points for the Q–Q Plot Correlation Coefficient Test for Normality

Sample size n	Significance levels α		
	.01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
20	.9269	.9508	.9604
25	.9410	.9591	.9665
30	.9479	.9652	.9715
35	.9538	.9682	.9740
40	.9599	.9726	.9771
45	.9632	.9749	.9792
50	.9671	.9768	.9809
55	.9695	.9787	.9822
60	.9720	.9801	.9836
75	.9771	.9838	.9866
100	.9822	.9873	.9895
150	.9879	.9913	.9928
200	.9905	.9931	.9942
300	.9935	.9953	.9960

Correlation Coefficient for Normality

Example 4.11 (A correlation coefficient test for normality) Let us calculate the correlation coefficient r_Q from the Q–Q plot of Example 4.9 (see Figure 4.5) and test for normality.

Using the information from Example 4.9, we have $\bar{x} = .770$ and

$$\sum_{j=1}^{10} (x_{(j)} - \bar{x})q_{(j)} = 8.584, \sum_{j=1}^{10} (x_{(j)} - \bar{x})^2 = 8.472, \text{ and } \sum_{j=1}^{10} q_{(j)}^2 = 8.795$$

Since always, $\bar{q} = 0$,

$$r_Q = \frac{8.584}{\sqrt{8.472}\sqrt{8.795}} = .994$$

A test of normality at the 10% level of significance is provided by referring $r_Q = .994$ to the entry in Table 4.2 corresponding to $n = 10$ and $\alpha = .10$. This entry is .9351. Since $r_Q > .9351$, we do not reject the hypothesis of normality. ■

Shapiro and Wilk Statistic and Correspondence to Points in the Normal Scores Plot

Instead of r_Q, some software packages evaluate the original statistic proposed by Shapiro and Wilk [12]. Its correlation form corresponds to replacing $q_{(j)}$ by a function of the expected value of standard normal-order statistics and their covariances. We prefer r_Q because it corresponds directly to the points in the normal-scores plot. For large sample sizes, the two statistics are nearly the same (see [13]), so either can be used to judge lack of fit.

Linear combinations of more than one characteristic can be investigated. Many statisticians suggest plotting

$$\hat{\mathbf{e}}_1'\mathbf{x}_j \quad \text{where} \quad \mathbf{S}\,\hat{\mathbf{e}}_1 = \hat{\lambda}_1\hat{\mathbf{e}}_1$$

in which $\hat{\lambda}_1$ is the largest eigenvalue of $\mathbf{S}$. Here $\mathbf{x}_j' = [x_{j1}, x_{j2}, \ldots, x_{jp}]$ is the jth observation on the p variables $X_1, X_2, \ldots, X_p$. The linear combination $\hat{\mathbf{e}}_p'\mathbf{x}_j$ corresponding to the smallest eigenvalue is also frequently singled out for inspection. (See Chapter 8 and [6] for further details.)

Evaluating Bivariate Normality

Evaluating Bivariate Normality

We would like to check on the assumption of normality for all distributions of $2, 3, \ldots, p$ dimensions. However, as we have pointed out, for practical work it is usually sufficient to investigate the univariate and bivariate distributions. We considered univariate marginal distributions earlier. It is now of interest to examine the bivariate case.

In Chapter 1, we described scatter plots for pairs of characteristics. If the observations were generated from a multivariate normal distribution, each bivariate distribution would be normal, and the contours of constant density would be ellipses. The scatter plot should conform to this structure by exhibiting an overall pattern that is nearly elliptical.

Moreover, by Result 4.7, the set of bivariate outcomes $\mathbf{x}$ such that

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leq \chi_2^2(.5)$$

Approximately 50% of Sample Observations Lie in the Ellipse

has probability .5. Thus, we should expect *roughly* the same percentage, 50%, of sample observations to lie in the ellipse given by

$$\{\text{all } \mathbf{x} \text{ such that } (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq \chi^2_2(.5)\}$$

where we have replaced $\boldsymbol{\mu}$ by its estimate $\bar{\mathbf{x}}$ and $\boldsymbol{\Sigma}^{-1}$ by its estimate $\mathbf{S}^{-1}$. If not, the normality assumption is suspect.

Checking Bivariate Normality

Example 4.12 (Checking bivariate normality) Although not a random sample, data consisting of the pairs of observations (x_1 = sales, x_2 = profits) for the 10 largest companies in the world are listed in Exercise 1.4. These data give

$$\bar{\mathbf{x}} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

so

$$\mathbf{S}^{-1} = \frac{1}{103{,}623.12} \begin{bmatrix} 26.19 & -303.62 \\ -303.62 & 7476.45 \end{bmatrix}$$

$$= \begin{bmatrix} .000253 & -.002930 \\ -.002930 & .072148 \end{bmatrix}$$

From Table 3 in the appendix, $\chi^2_2(.5) = 1.39$. Thus, any observation $\mathbf{x}' = [x_1, x_2]$ satisfying

$$\begin{bmatrix} x_1 - 155.60 \\ x_2 - 14.70 \end{bmatrix}' \begin{bmatrix} .000253 & -.002930 \\ -.002930 & .072148 \end{bmatrix} \begin{bmatrix} x_1 - 155.60 \\ x_2 - 14.70 \end{bmatrix} \leq 1.39$$

is on or inside the estimated 50% contour. Otherwise the observation is outside this contour. The first pair of observations in Exercise 1.4 is $[x_1, x_2]' = [108.28, 17.05]$.

Checking Bivariate Normality

From Table 3 in the appendix, $\chi^2_2(.5) = 1.39$. Thus, any observation $\mathbf{x}' = [x_1, x_2]$ satisfying

$$\begin{bmatrix} x_1 - 155.60 \\ x_2 - 14.70 \end{bmatrix}' \begin{bmatrix} .000253 & -.002930 \\ -.002930 & .072148 \end{bmatrix} \begin{bmatrix} x_1 - 155.60 \\ x_2 - 14.70 \end{bmatrix} \leq 1.39$$

is on or inside the estimated 50% contour. Otherwise the observation is outside this contour. The first pair of observations in Exercise 1.4 is $[x_1, x_2]' = [108.28, 17.05]$. In this case

$$\begin{bmatrix} 108.28 - 155.60 \\ 17.05 - 14.70 \end{bmatrix}' \begin{bmatrix} .000253 & -.002930 \\ -.002930 & .072148 \end{bmatrix} \begin{bmatrix} 108.28 - 155.60 \\ 17.05 - 14.70 \end{bmatrix}$$

$$= 1.61 > 1.39$$

and this point falls outside the 50% contour. The remaining nine points have generalized distances from $\bar{\mathbf{x}}$ of .30, .62, 1.79, 1.30, 4.38, 1.64, 3.53, 1.71, and 1.16, respectively. Since four of these distances are less than 1.39, a proportion, .40, of the data falls within the 50% contour. If the observations were normally distributed, we would expect about half, or 5, of them to be within this contour. This difference in proportions might ordinarily provide evidence for rejecting the notion of bivariate normality; however, our sample size of 10 is too small to reach this conclusion. (See also Example 4.13.) ■

Chi-Square Plot or Gamma Plot Based on Squared Generalized Distances

Computing the fraction of the points within a contour and subjectively comparing it with the theoretical probability is a useful, but rather rough, procedure.

A somewhat more formal method for judging the joint normality of a data set is based on the squared generalized distances

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}), \qquad j = 1, 2, \ldots, n \tag{4-32}$$

where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are the sample observations. The procedure we are about to describe is not limited to the bivariate case; it can be used for all $p \geq 2$.

When the parent population is multivariate normal and both n and $n - p$ are greater than 25 or 30, each of the squared distances $d_1^2, d_2^2, \ldots, d_n^2$ should behave like a chi-square random variable. [See Result 4.7 and Equations (4-26) and (4-27).] Although these distances are *not* independent or exactly chi-square distributed, it is helpful to plot them as if they were. The resulting plot is called a *chi-square plot* or *gamma plot*, because the chi-square distribution is a special case of the more general gamma distribution. (See [6].)

Constructing Chi-Square Plot

To construct the chi-square plot,

1. Order the squared distances in (4-32) from smallest to largest as $d^2_{(1)} \le d^2_{(2)} \le \cdots \le d^2_{(n)}$.
2. Graph the pairs $\left(q_{c,p}\left(\left(j - \frac{1}{2}\right)/n\right), d^2_{(j)}\right)$, where $q_{c,p}\left(\left(j - \frac{1}{2}\right)/n\right)$ is the $100\left(j - \frac{1}{2}\right)/n$ quantile of the chi-square distribution with p degrees of freedom.

Quantiles are specified in terms of proportions, whereas percentiles are specified in terms of percentages.

The quantiles $q_{c,p}\left(\left(j - \frac{1}{2}\right)/n\right)$ are related to the upper percentiles of a chi-squared distribution. In particular, $q_{c,p}\left(\left(j - \frac{1}{2}\right)/n\right) = \chi^2_p\left(\left(n - j + \frac{1}{2}\right)/n\right)$.

The plot should resemble a straight line through the origin having slope 1. A systematic curved pattern suggests lack of normality. One or two points far above the line indicate large distances, or outlying observations, that merit further attention.

Example Constructing Chi –Square Plot

Example 4.13 (Constructing a chi-square plot) Let us construct a chi-square plot of the generalized distances given in Example 4.12. The ordered distances and the corresponding chi-square percentiles for $p = 2$ and $n = 10$ are listed in the following table:

j	$d^2_{(j)}$	$q_{c,2}\left(\frac{j-\frac{1}{2}}{10}\right)$
1	.30	.10
2	.62	.33
3	1.16	.58
4	1.30	.86
5	1.61	1.20
6	1.64	1.60
7	1.71	2.10
8	1.79	2.77
9	3.53	3.79
10	4.38	5.99

Chi-Square Plot of Ordered Distances

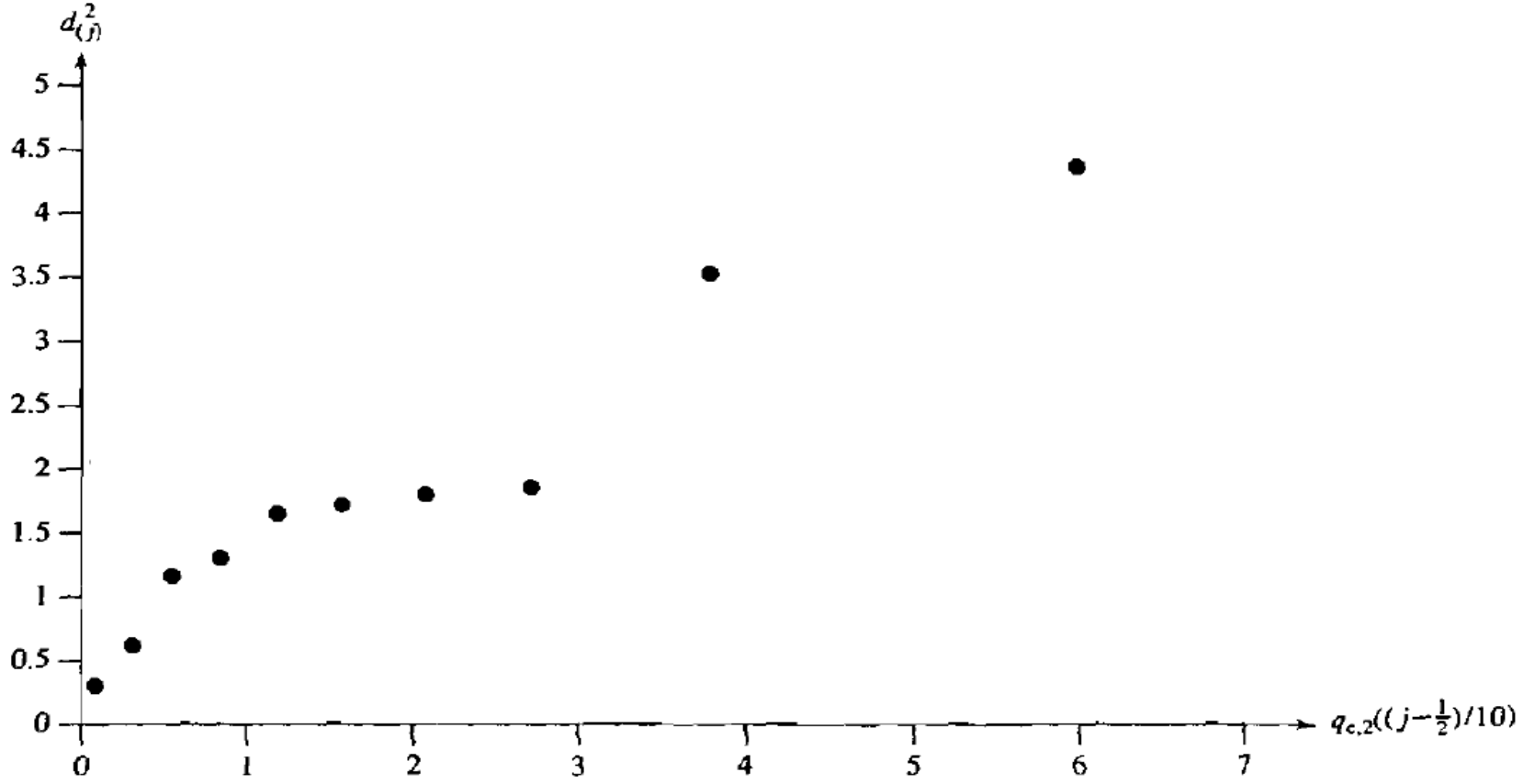


Figure 4.7 A chi-square plot of the ordered distances in Example 4.13.

Straight Line – Transformation May Not Be Necessary

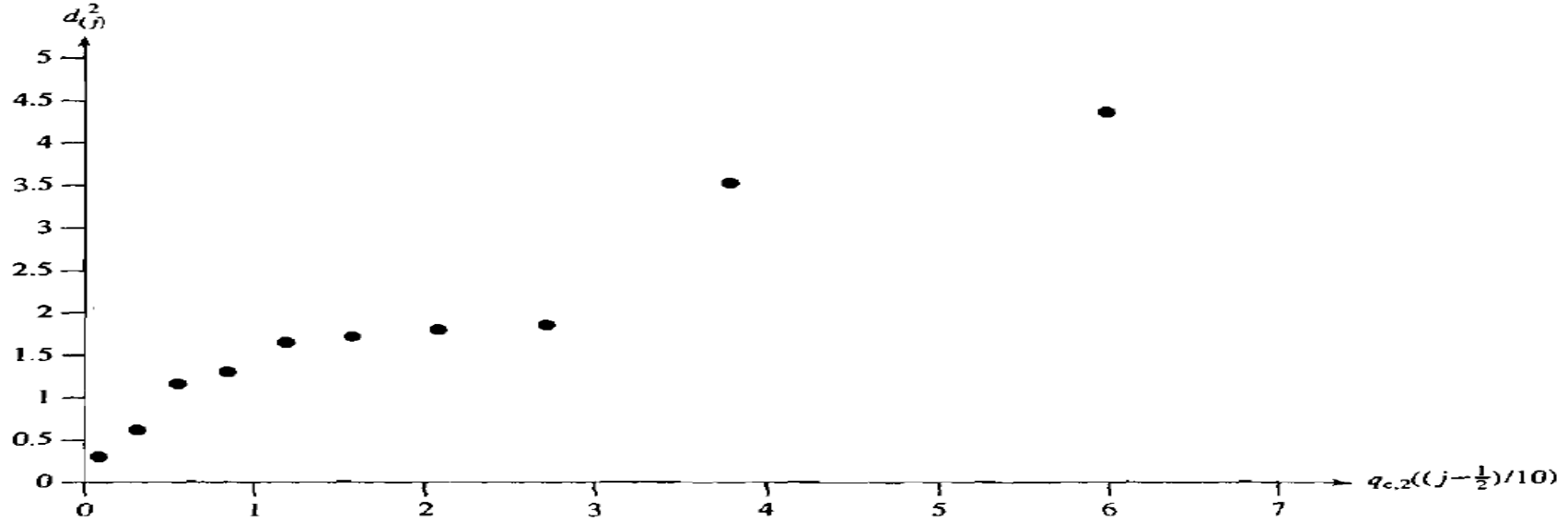


Figure 4.7 A chi-square plot of the ordered distances in Example 4.13.

A graph of the pairs $\left(q_{c,2}\left(\left(j-\frac{1}{2}\right)/10\right), d^2_{(j)}\right)$ is shown in Figure 4.7. The points in Figure 4.7 are reasonably straight. Given the small sample size it is difficult to reject bivariate normality on the evidence in this graph. If further analysis of the data were required, it might be reasonable to transform them to observations more nearly bivariate normal. Appropriate transformations are discussed in Section 4.8. ■

Checking Multivariate Normality – Chi-Squared or d^2 Plot

In addition to inspecting univariate plots and scatter plots, we should check multivariate normality by constructing a chi-squared or d^2 plot. Figure 4.8 contains d^2

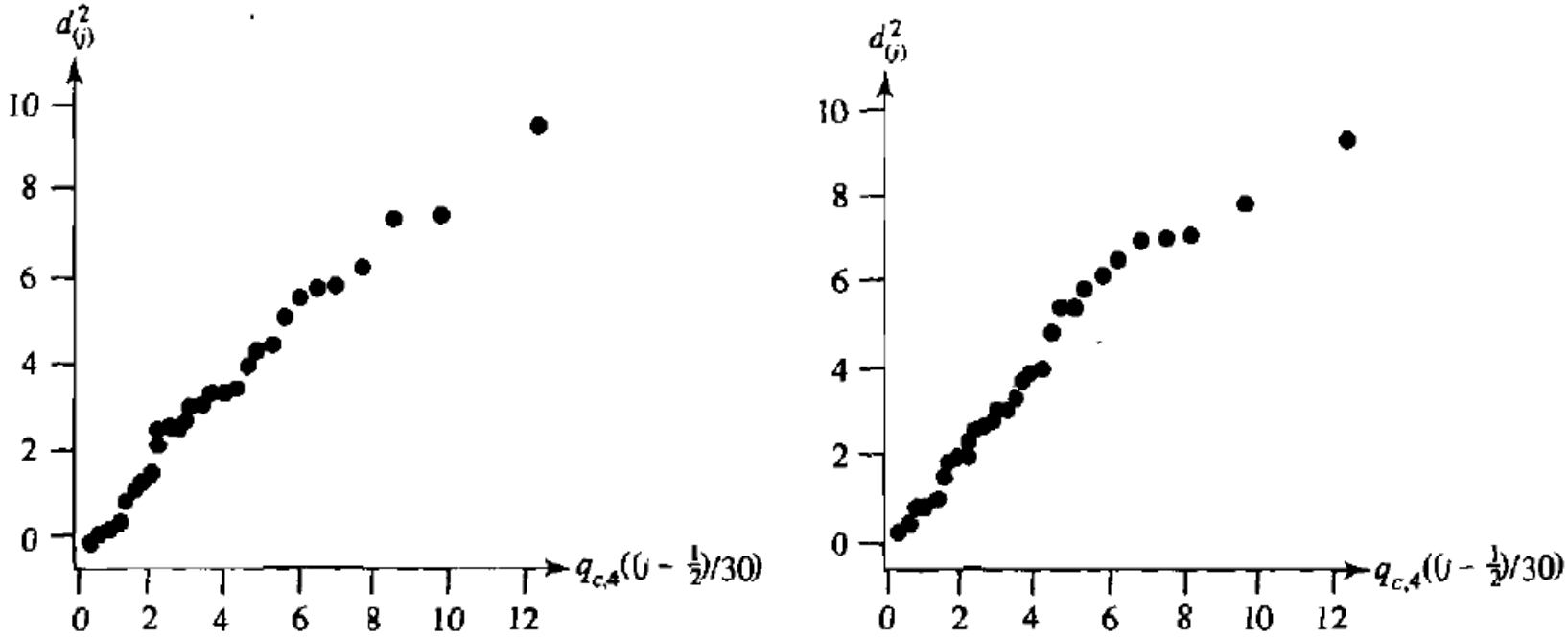


Figure 4.8 Chi-square plots for two simulated four-variate normal data sets with $n = 30$.

Possible Outliers - Detected

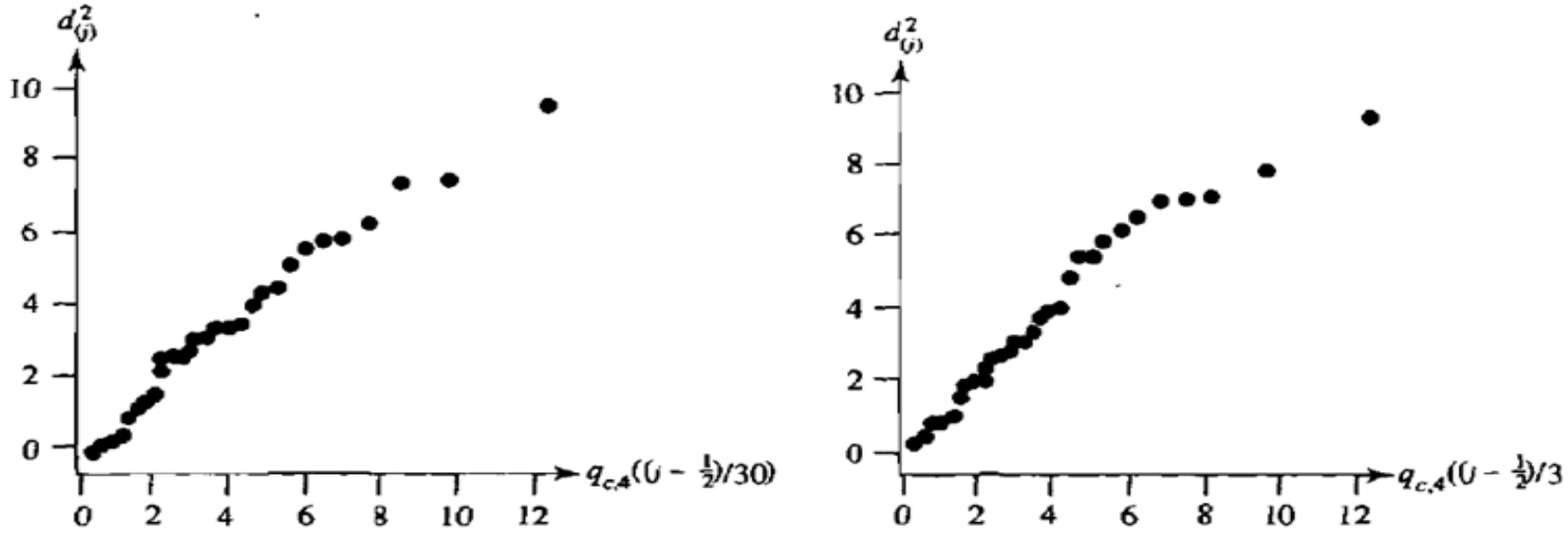


Figure 4.8 Chi-square plots for two simulated four-variate normal data sets with $n = 30$

plots based on two computer-generated samples of 30 four-variate normal random vectors. As expected, the plots have a straight-line pattern, but the top two or three ordered squared distances are quite variable.

The next example contains a real data set comparable to the simulated data set that produced the plots in Figure 4.8.

Evaluating Multivariate Normality for Four-Variable Data Set

Example 4.14 (Evaluating multivariate normality for a four-variable data set) The data in Table 4.3 were obtained by taking four different measures of stiffness, x_1, x_2, x_3, and x_4, of each of $n = 30$ boards. The first measurement involves sending a shock wave down the board, the second measurement is determined while vibrating the board, and the last two measurements are obtained from static tests. The squared distances $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$ are also presented in the table.

Table 4.3 Four Measurements of Stiffness

Observation no.	x_1	x_2	x_3	x_4	d^2	Observation no.	x_1	x_2	x_3	x_4	d^2
1	1889	1651	1561	1778	.60	16	1954	2149	1180	1281	16.85
2	2403	2048	2087	2197	5.48	17	1325	1170	1002	1176	3.50
3	2119	1700	1815	2222	7.62	18	1419	1371	1252	1308	3.99
4	1645	1627	1110	1533	5.21	19	1828	1634	1602	1755	1.36
5	1976	1916	1614	1883	1.40	20	1725	1594	1313	1646	1.46
6	1712	1712	1439	1546	2.22	21	2276	2189	1547	2111	9.90
7	1943	1685	1271	1671	4.99	22	1899	1614	1422	1477	5.06
8	2104	1820	1717	1874	1.49	23	1633	1513	1290	1516	.80
9	2983	2794	2412	2581	12.26	24	2061	1867	1646	2037	2.54
10	1745	1600	1384	1508	.77	25	1856	1493	1356	1533	4.58
11	1710	1591	1518	1667	1.93	26	1727	1412	1238	1469	3.40

Note d^2 Value for Observation No. 9

Table 4.3 Four Measurements of Stiffness

Observation no.	x_1	x_2	x_3	x_4	d^2
1	1889	1651	1561	1778	.60
2	2403	2048	2087	2197	5.48
3	2119	1700	1815	2222	7.62
4	1645	1627	1110	1533	5.21
5	1976	1916	1614	1883	1.40
6	1712	1712	1439	1546	2.22
7	1943	1685	1271	1671	4.99
8	2104	1820	1717	1874	1.49
9	2983	2794	2412	2581	12.26
10	1745	1600	1384	1508	.77
11	1710	1591	1518	1667	1.93
12	2046	1907	1627	1898	.46
13	1840	1841	1595	1741	2.70
14	1867	1685	1493	1678	.13
15	1859	1649	1389	1714	1.08
16	1954	2149	1180	1281	16.85
17	1325	1170	1002	1176	3.50
18	1419	1371	1252	1308	3.99
19	1828	1634	1602	1755	1.36
20	1725	1594	1313	1646	1.46
21	2276	2189	1547	2111	9.90
22	1899	1614	1422	1477	5.06
23	1633	1513	1290	1516	.80
24	2061	1867	1646	2037	2.54
25	1856	1493	1356	1533	4.58
26	1727	1412	1238	1469	3.40
27	2168	1896	1701	1834	2.38
28	1655	1675	1414	1597	3.00
29	2326	2301	2065	2234	6.28
30	1490	1382	1214	1284	2.58

Source: Data courtesy of William Galligan.

The marginal distributions appear quite normal (see Exercise 4.33), with the possible exception of specimen (board) 9.

Removing Two Observations (Specimens) With Largest Squared Distances

The marginal distributions appear quite normal (see Exercise 4.33), with the possible exception of specimen (board) 9.

To further evaluate multivariate normality, we constructed the chi-square plot shown in Figure 4.9. The two specimens with the largest squared distances are clearly removed from the straight-line pattern. Together, with the next largest point or two, they make the plot appear curved at the upper end. We will return to a discussion of this plot in Example 4.15. ■

Properties of Chi-Square Plot (or P-P Plot or Probability Plot in SPSS, R, etc.)

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}), \qquad j = 1, 2, \ldots, n \tag{4-32}$$

We have discussed some rather simple techniques for checking the multivariate normality assumption. Specifically, we advocate calculating the d_j^2, $j = 1, 2, \ldots, n$ [see Equation (4-32)] and comparing the results with χ^2 quantiles. For example, p-variate normality is indicated if

1. Roughly half of the d_j^2 are less than or equal to $q_{c,p}(.50)$.

2. A plot of the ordered squared distances $d_{(1)}^2 \leq d_{(2)}^2 \leq \cdots \leq d_{(n)}^2$ versus $q_{c,p}\left(\frac{1-\frac{1}{2}}{n}\right), q_{c,p}\left(\frac{2-\frac{1}{2}}{n}\right), \ldots, q_{c,p}\left(\frac{n-\frac{1}{2}}{n}\right)$, respectively, is nearly a straight line having slope 1 and that passes through the origin.

(See [6] for a more complete exposition of methods for assessing normality.)

We close this section by noting that all measures of goodness of fit suffer the same serious drawback. When the sample size is small, only the most aberrant behavior will be identified as lack of fit. On the other hand, very large samples invariably produce statistically significant lack of fit. Yet the departure from the specified distribution may be very small and technically unimportant to the inferential conclusions.

Chi-Square Plot (P-P Plot) for Example 4.14 i.e. Measures of Stiffness

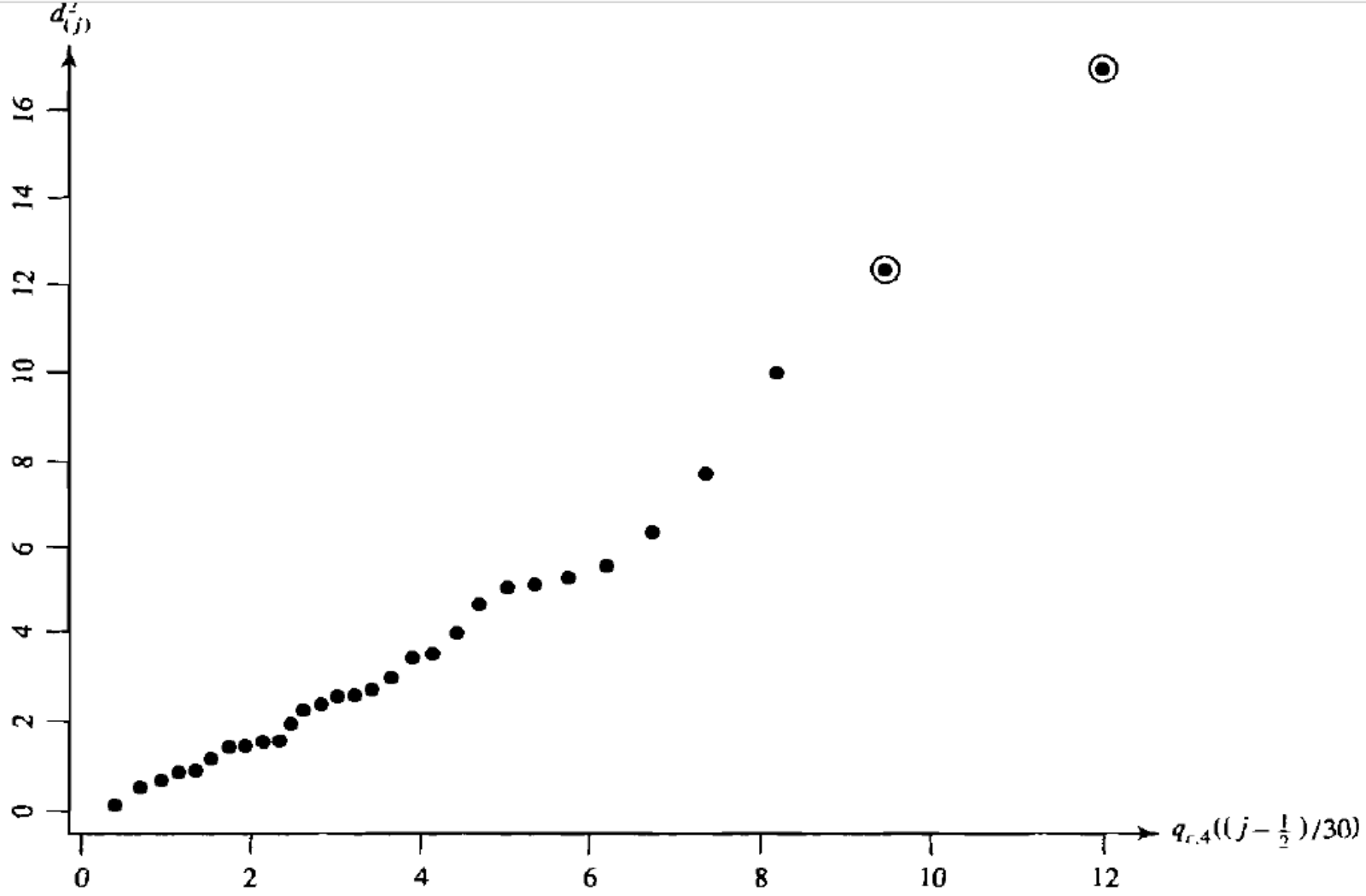


Figure 4.9 A chi-square plot for the data in Example 4.14.

Detecting Outliers and Cleaning Data

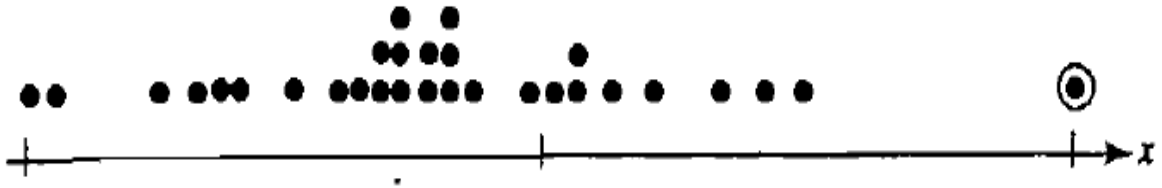
4.7 Detecting Outliers and Cleaning Data

Most data sets contain one or a few unusual observations that do not seem to belong to the pattern of variability produced by the other observations. With data on a single characteristic, unusual observations are those that are either very large or very small relative to the others. The situation can be more complicated with multivariate data. Before we address the issue of identifying these *outliers*, we must emphasize that not all outliers are wrong numbers. They may, justifiably, be part of the group and may lead to a better understanding of the phenomena being studied.

Detecting Outliers: Start with Univariate Graphical Analysis

Outliers are best detected visually whenever this is possible. When the number of observations n is large, dot plots are not feasible. When the number of characteristics p is large, the large number of scatter plots $p(p-1)/2$ may prevent viewing them all. Even so, we suggest first visually inspecting the data whenever possible.

What should we look for? For a single random variable, the problem is one dimensional, and we look for observations that are far from the others. For instance, the dot diagram



reveals a single large observation which is circled.

Bivariate Graph and Two Possible Outliers (Not Detected in Univariate Dot Plot)

In the bivariate case, the situation is more complicated. Figure 4.10 shows a situation with two unusual observations.

The data point circled in the upper right corner of the figure is detached from the pattern, and its second coordinate is large relative to the rest of the x_2 measurements, as shown by the vertical dot diagram. The second outlier, also circled, is far from the elliptical pattern of the rest of the points, but, separately, each of its components has a typical value. This outlier cannot be detected by inspecting the marginal dot diagrams.

In higher dimensions, there can be outliers that cannot be detected from the univariate plots or even the bivariate scatter plots. Here a large value of $(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$ will suggest an unusual observation, even though it cannot be seen visually.

Outliers Detected in Uni- and Bivariate Graphs

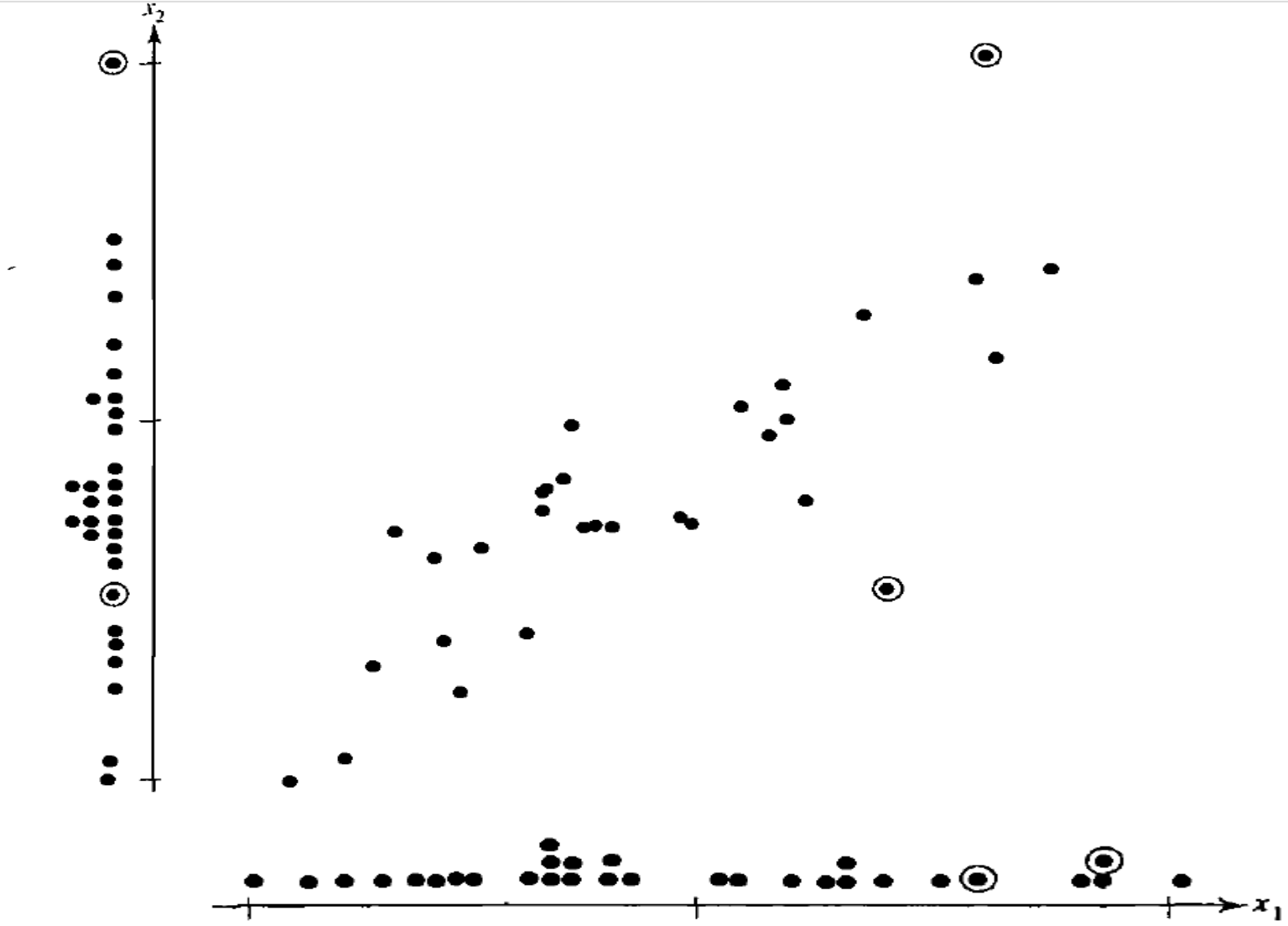


Figure 4.10 Two outliers; one univariate and one bivariate.

Steps for Detecting Outliers

Steps for Detecting Outliers

1. Make a dot plot for each variable.
2. Make a scatter plot for each pair of variables.
3. Calculate the standardized values $z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$ for $j = 1, 2, \ldots, n$ and each column $k = 1, 2, \ldots, p$. Examine these standardized values for large or small values.
4. Calculate the generalized squared distances $(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$. Examine these distances for unusually large values. In a chi-square plot, these would be the points farthest from the origin.

In step 3, "large" must be interpreted relative to the sample size and number of variables. There are $n \times p$ standardized values. When $n = 100$ and $p = 5$, there are 500 values. You expect 1 or 2 of these to exceed 3 or be less than -3, even if the data came from a multivariate distribution that is exactly normal. As a guideline, 3.5 might be considered large for moderate sample sizes.

In step 4, "large" is measured by an appropriate percentile of the chi-square distribution with p degrees of freedom. If the sample size is $n = 100$, we would expect 5 observations to have values of d_j^2 that exceed the upper fifth percentile of the chi-square distribution. A more extreme percentile must serve to determine observations that do not fit the pattern of the remaining data.

Data – Lumber; Values and Their Standardized Values

The data we presented in Table 4.3 concerning lumber have already been cleaned up somewhat. Similar data sets from the same study also contained data on x_5 = tensile strength. Nine observation vectors, out of the total of 112, are given as rows in the following table, along with their standardized values.

x_1	x_2	x_3	x_4	x_5	z_1	z_2	z_3	z_4	z_5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
1631	1528	1452	1559	1602	.06	−.15	.05	.28	−.12
1770	1677	1707	1738	1785	.64	.43	1.07	.94	.60
1376	1190	723	1285	2791	−1.01	−1.47	−2.87	−.73	4.57
1705	1577	1332	1703	1664	.37	.04	−.43	.81	.13
1643	1535	1510	1494	1582	.11	−.12	.28	.04	−.20
1567	1510	1301	1405	1553	−.21	−.22	−.56	−.28	−.31
1528	1591	1714	1685	1698	−.38	.10	1.10	.75	.26
1803	1826	1748	2746	1764	.78	1.01	1.23	4.65	.52
1587	1554	1352	1554	1551	−.13	−.05	−.35	.26	−.32
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Calculation of Standardized Values Based on Mean and Variance n =112

The standardized values are based on the sample mean and variance, calculated from all 112 observations. There are two extreme standardized values. Both are too large with standardized values over 4.5. During their investigation, the researchers recorded measurements by hand in a logbook and then performed calculations that produced the values given in the table. When they checked their records regarding the values pinpointed by this analysis, errors were discovered. The value $x_5 = 2791$ was corrected to 1241, and $x_4 = 2746$ was corrected to 1670. Incorrect readings on an individual variable are quickly detected by locating a large leading digit for the standardized value.

The next example returns to the data on lumber discussed in Example 4.14.

Detecting Outliers in the Data on Lumber

Example 4.15 (Detecting outliers in the data on lumber) Table 4.4 contains the data in Table 4.3, along with the standardized observations. These data consist of four different measures of stiffness x_1, x_2, x_3, and x_4, on each of $n = 30$ boards. Recall that the first measurement involves sending a shock wave down the board, the second measurement is determined while vibrating the board, and the last two measurements are obtained from static tests. The standardized measurements are

Table 4.4 Four Measurements of Stiffness with Standardized Values

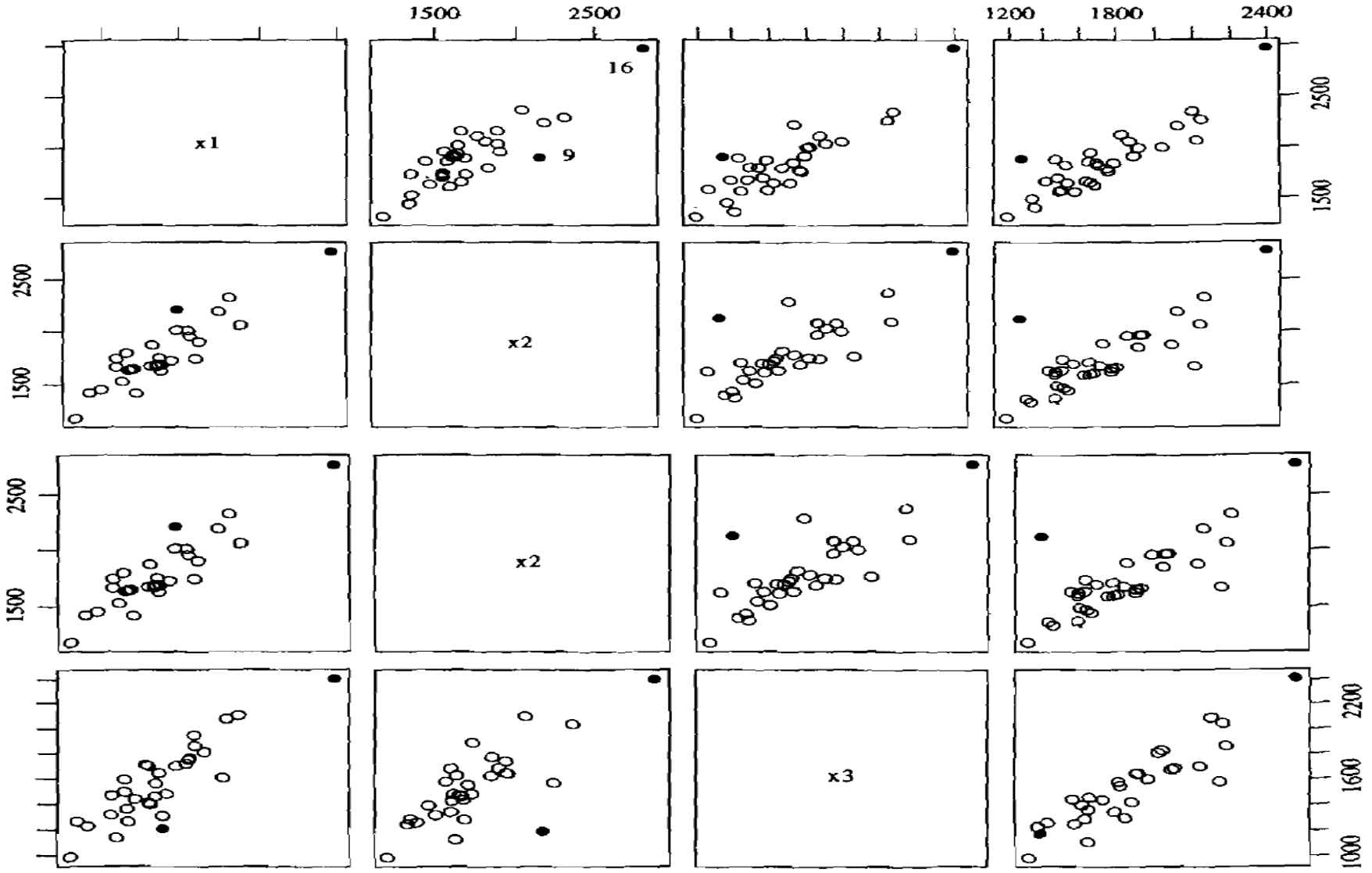
x_1	x_2	x_3	x_4	Observation no.	z_1	z_2	z_3	z_4	d^2
1889	1651	1561	1778	1	−.1	−.3	.2	.2	.60
2403	2048	2087	2197	2	1.5	.9	1.9	1.5	5.48
2119	1700	1815	2222	3	.7	−.2	1.0	1.5	7.62
1645	1627	1110	1533	4	−.8	−.4	−1.3	−.6	5.21
1976	1916	1614	1883	5	.2	.5	.3	.5	1.40
1712	1712	1439	1546	6	−.6	−.1	−.2	−.6	2.22
1943	1685	1271	1671	7	.1	−.2	−.8	−.2	4.99
2104	1820	1717	1874	8	.6	.2	.7	.5	1.49
2983	2794	2412	2581	9	3.3	3.3	3.0	2.7	12.26
1745	1600	1384	1508	10	−.5	−.5	−.4	−.7	.77
1710	1591	1518	1667	11	−.6	−.5	.0	−.2	1.93
2046	1907	1627	1898	12	.4	.5	.4	.5	.46

Variables (Measures) of Stiffness

Table 4.4 Four Measurements of Stiffness with Standardized Values

x_1	x_2	x_3	x_4	Observation no.	z_1	z_2	z_3	z_4	d^2
1889	1651	1561	1778	1	−.1	−.3	.2	.2	.60
2403	2048	2087	2197	2	1.5	.9	1.9	1.5	5.48
2119	1700	1815	2222	3	.7	−.2	1.0	1.5	7.62
1645	1627	1110	1533	4	−.8	−.4	−1.3	−.6	5.21
1976	1916	1614	1883	5	.2	.5	.3	.5	1.40
1712	1712	1439	1546	6	−.6	−.1	−.2	−.6	2.22
1943	1685	1271	1671	7	.1	−.2	−.8	−.2	4.99
2104	1820	1717	1874	8	.6	.2	.7	.5	1.49
2983	2794	2412	2581	9	3.3	3.3	3.0	2.7	12.26
1745	1600	1384	1508	10	−.5	−.5	−.4	−.7	.77
1710	1591	1518	1667	11	−.6	−.5	.0	−.2	1.93
2046	1907	1627	1898	12	.4	.5	.4	.5	.46
1840	1841	1595	1741	13	−.2	.3	.3	.0	2.70
1867	1685	1493	1678	14	−.1	−.2	−.1	−.1	.13
1859	1649	1389	1714	15	−.1	−.3	−.4	−.0	1.08
1954	2149	1180	1281	16	.1	1.3	−1.1	−1.4	16.85
1325	1170	1002	1176	17	−1.8	−1.8	−1.7	−1.7	3.50
1419	1371	1252	1308	18	−1.5	−1.2	−.8	−1.3	3.99
1828	1634	1602	1755	19	−.2	−.4	.3	.1	1.36
1725	1594	1313	1646	20	−.6	−.5	−.6	−.2	1.46
2276	2189	1547	2111	21	1.1	1.4	.1	1.2	9.90
1899	1614	1422	1477	22	−.0	−.4	−.3	−.8	5.06
1633	1513	1290	1516	23	−.8	−.7	−.7	−.6	.80
2061	1867	1646	2037	24	.5	.4	.5	1.0	2.54
1856	1493	1356	1533	25	−.2	−.8	−.5	−.6	4.58
1727	1412	1238	1469	26	−.6	−1.1	−.9	−.8	3.40
2168	1896	1701	1834	27	.8	.5	.6	.3	2.38
1655	1675	1414	1597	28	−.8	−.2	−.3	−.4	3.00
2326	2301	2065	2234	29	1.3	1.7	1.8	1.6	6.28
1490	1382	1214	1284	30	−1.3	−1.2	−1.0	−1.4	2.58

Scatter Plots: Lumber Stiffness



Scatter Plots – Lumber Stiffness With Case No. (or Specimen No.) 9 and 16 Potential Outliers

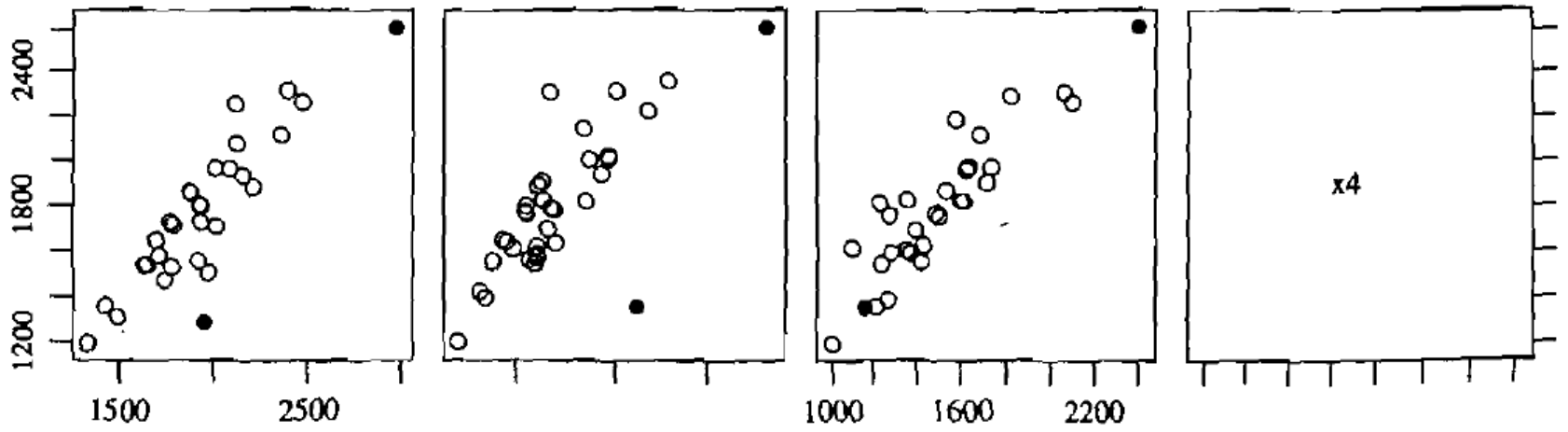


Figure 4.11 Scatter plots for the lumber stiffness data with specimens 9 and 16 plotted as solid dots.

$$z_{jk} = \frac{x_{jk} - \bar{x}_k}{\sqrt{s_{kk}}}, \qquad k = 1, 2, 3, 4; \qquad j = 1, 2, \ldots, 30$$

and the squares of the distances are $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$.

Multivariate Outlier Detected by d^2

are obtained from static tests. The standardized measurements are

$$z_{jk} = \frac{x_{jk} - \bar{x}_k}{\sqrt{s_{kk}}}, \qquad k = 1, 2, 3, 4; \qquad j = 1, 2, \ldots, 30$$

and the squares of the distances are $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$.

The last column in Table 4.4 reveals that specimen 16 is a multivariate outlier, since $\chi_4^2(.005) = 14.86$; yet all of the individual measurements are well within their respective univariate scatters. Specimen 9 also has a large d^2 value.

The two specimens (9 and 16) with large squared distances stand out as clearly different from the rest of the pattern in Figure 4.9. Once these two points are removed, the remaining pattern conforms to the expected straight-line relation. Scatter plots for the lumber stiffness measurements are given in Figure 4.11 above.

Identifying Outliers: Assess for Substantive Content Before Deleting

The solid dots in these figures correspond to specimens 9 and 16. Although the dot for specimen 16 stands out in all the plots, the dot for specimen 9 is "hidden" in the scatter plot of x_3 versus x_4 and nearly hidden in that of x_1 versus x_3. However, specimen 9 is clearly identified as a multivariate outlier when all four variables are considered.

Scientists specializing in the properties of wood conjectured that specimen 9 was unusually clear and therefore very stiff and strong. It would also appear that specimen 16 is a bit unusual, since both of its dynamic measurements are above average and the two static measurements are low. Unfortunately, it was not possible to investigate this specimen further because the material was no longer available. ■

If outliers are identified, they should be examined for content, as was done in the case of the data on lumber stiffness in Example 4.15. Depending upon the nature of the outliers and the objectives of the investigation, outliers may be deleted or appropriately "weighted" in a subsequent analysis.

Even though many statistical techniques assume normal populations, those based on the sample mean vectors usually will not be disturbed by a few moderate outliers. Hawkins [7] gives an extensive treatment of the subject of outliers.

Transformations to Almost Normal Distribution: Theoretical Or Data Driven

4.8 Transformations to Near Normality

If normality is not a viable assumption, what is the next step? One alternative is to ignore the findings of a normality check and proceed as if the data were normally distributed. This practice is not recommended, since, in many instances, it could lead to incorrect conclusions. A second alternative is to make nonnormal data more "normal looking" by considering *transformations* of the data. Normal-theory analyses can then be carried out with the suitably transformed data.

Transformations are nothing more than a reexpression of the data in different units. For example, when a histogram of positive observations exhibits a long right-hand tail, transforming the observations by taking their logarithms or square roots will often markedly improve the symmetry about the mean and the approximation to a normal distribution. It frequently happens that the new units provide more natural expressions of the characteristics being studied.

Appropriate transformations are suggested by (1) theoretical considerations or (2) the data themselves (or both). It has been shown theoretically that data that are counts can often be made more normal by taking their *square roots*. Similarly, the *logit transformation* applied to proportions and *Fisher's z-transformation* applied to correlation coefficients yield quantities that are approximately normally distributed.

Transformations (Scaling) to Near Normality

Helpful Transformations To Near Normality

Original Scale	*Transformed Scale*	
1. Counts, y	$\sqrt{y}$	
2. Proportions, $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$	(4-33)
3. Correlations, r	Fisher's $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$	

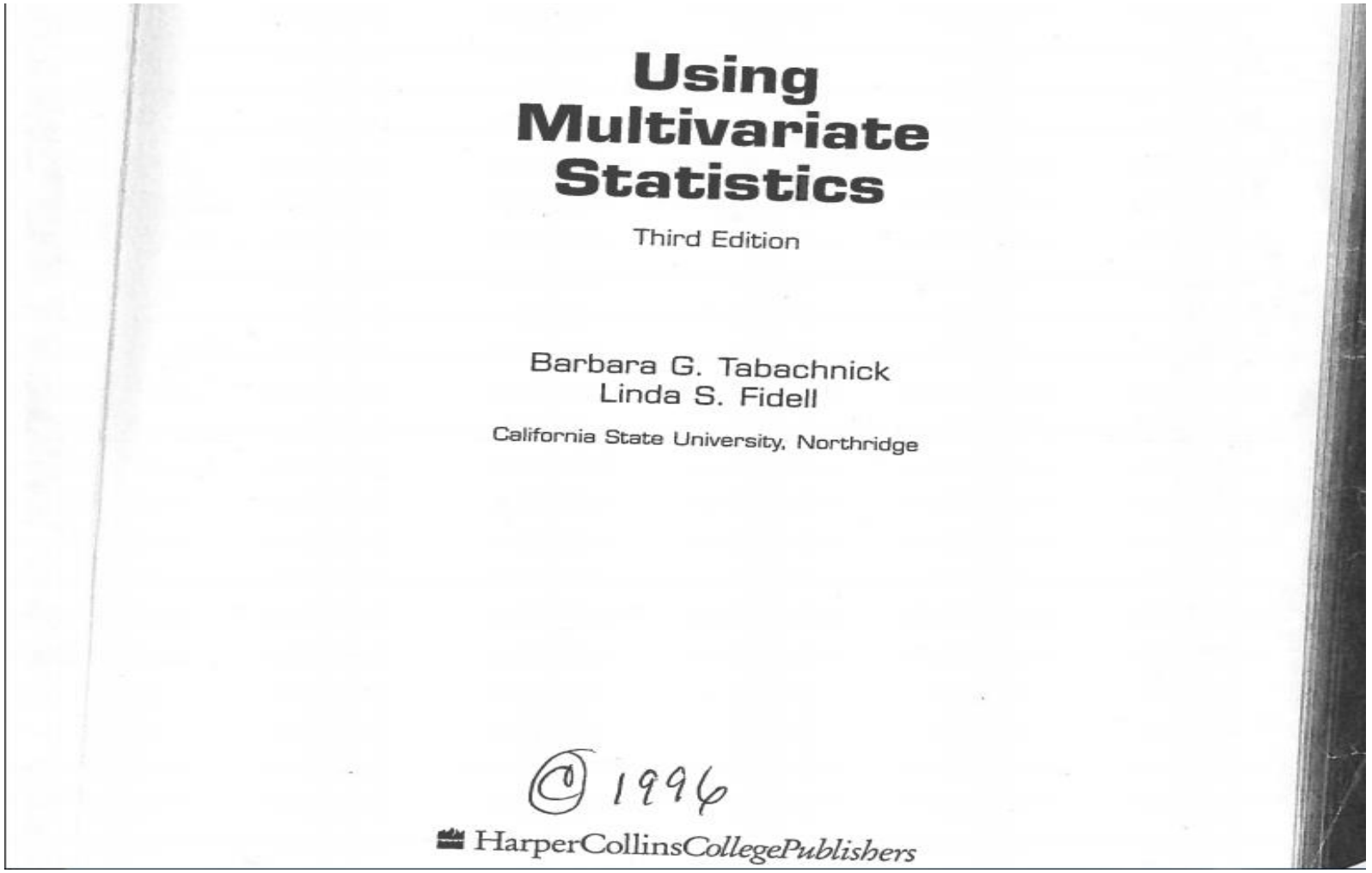
Variance Stabilizing Transformations

TABLE 3.1. Variance stabilizing transformations

$h(\mu)$	z	Description
μ^4	y^{-1}	Reciprocal
μ^2	$log(y)$	Logarithm
μ	$\sqrt{\mu}$	Square root (Poisson)
$\mu(1-\mu)$	$sin^{-1}(\sqrt{\mu})$	Inverse Sine (Binomial)
$(1-\mu^2)^2$	$log\frac{(1+\mu)}{(1-\mu)}$	Correlation

The last three transformations in Table 3.1 are based on theoretical expressions for the variance for the Poisson and Binomial distributions and the distribution of the sample correlation coefficient. In our application to regression models, we do not usually have the luxury of knowing the functional form of the relation. A plot of residuals against $\hat{y}$ may suggest that variability depends on the mean but the exact relation is not specified.

Original Distributions and Application of Scaling to Produce Normality From:



Using Multivariate Statistics

Third Edition

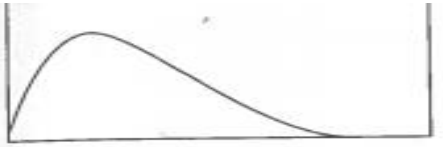
Barbara G. Tabachnick
Linda S. Fidell

California State University, Northridge

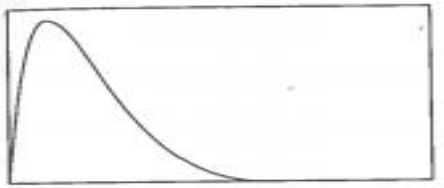
© 1996

HarperCollinsCollegePublishers

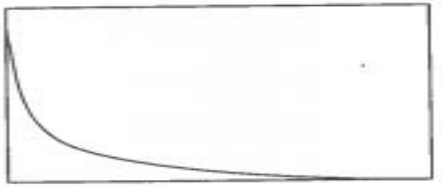
From Original Distribution Shape Can Determine Which Transformation To Apply



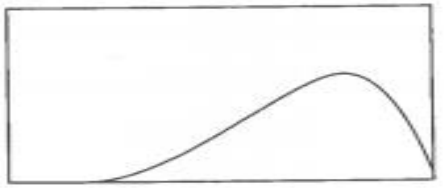
Square root $\sqrt{x}$



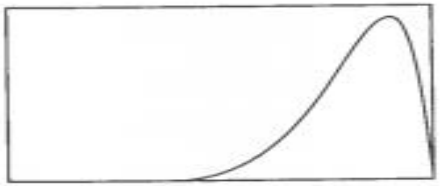
Logarithm $\log_{10}(x)$



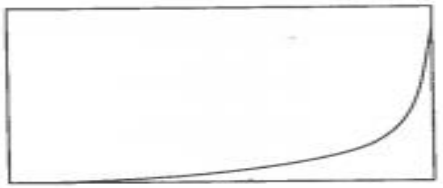
Inverse $1/x$



Reflect and square root $\sqrt{(K-x)}$



Reflect and logarithm $\log_{10}(K-x)$



Reflect and inverse $(1/K-x)$

FIGURE 4.6 ORIGINAL DISTRIBUTIONS AND COMMON TRANSFORMATIONS TO PRODUCE NORMALITY.

"Family" of Power Transformations

In many instances, the choice of a transformation to improve the approximation to normality is not obvious. For such cases, it is convenient to let the data suggest a transformation. A useful family of transformations for this purpose is the family of *power transformations*.

Power transformations are defined only for positive variables. However, this is not as restrictive as it seems, because a single constant can be added to each observation in the data set if some of the values are negative.

Let x represent an arbitrary observation. The power family of transformations is indexed by a parameter λ. A given value for λ implies a particular transformation. For example, consider x^λ with $\lambda = -1$. Since $x^{-1} = 1/x$, this choice of λ corresponds to the reciprocal transformation. We can trace the family of transformations as λ ranges from negative to positive powers of x. For $\lambda = 0$, we define $x^0 = \ln x$. A sequence of possible transformations is

$$\underbrace{\ldots, x^{-1} = \frac{1}{x}, x^0 = \ln x, x^{1/4} = \sqrt[4]{x}, x^{1/2} = \sqrt{x},}_{\text{shrinks large values of } x} \qquad \underbrace{x^2, x^3, \ldots}_{\text{increases large values of } x}$$

Selecting Transformation: Examine Univariate and Bivariate Plots

To select a power transformation, an investigator looks at the marginal dot diagram or histogram and decides whether large values have to be "pulled in" or "pushed out" to improve the symmetry about the mean. Trial-and-error calculations with a few of the foregoing transformations should produce an improvement. The final choice should always be examined by a Q–Q plot or other checks to see whether the tentative normal assumption is satisfactory.

The transformations we have been discussing are data based in the sense that it is only the appearance of the data themselves that influences the choice of an appropriate transformation. There are no external considerations involved, although the transformation actually used is often determined by some mix of information supplied by the data and extra-data factors, such as simplicity or ease of interpretation.

A convenient analytical method is available for choosing a power transformation. We begin by focusing our attention on the univariate case.

Box and Cox [3] consider the slightly modified family of power transformations

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \tag{4-34}$$

which is continuous in λ for $x > 0$. (See [8].) Given the observations $x_1, x_2, \ldots, x_n$, the Box–Cox solution for the choice of an appropriate power λ is the solution that *maximizes* the expression

Appropriate Power of λ Is the Solution That Maximizes the Expression

which is continuous in λ for $x > 0$. (See [8].) Given the observations $x_1, x_2, \ldots, x_n$, the Box–Cox solution for the choice of an appropriate power λ is the solution that *maximizes* the expression

$$\ell(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} (x_j^{(\lambda)} - \overline{x^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{j=1}^{n} \ln x_j \qquad (4\text{-}35)$$

We note that $x_j^{(\lambda)}$ is defined in (4-34) and

$$\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{x_j^{\lambda} - 1}{\lambda} \right) \qquad (4\text{-}36)$$

Creating A New Variable – Transformation of the Existing One

is the arithmetic average of the transformed observations. The first term in (4-35) is, apart from a constant, the logarithm of a normal likelihood function, after maximizing it with respect to the population mean and variance parameters.

The calculation of $\ell(\lambda)$ for many values of λ is an easy task for a computer. It is helpful to have a graph of $\ell(\lambda)$ versus λ, as well as a tabular display of the pairs $(\lambda, \ell(\lambda))$, in order to study the behavior near the maximizing value $\hat{\lambda}$. For instance, if either $\lambda = 0$ (logarithm) or $\lambda = \frac{1}{2}$ (square root) is near $\hat{\lambda}$, one of these may be preferred because of its simplicity.

Rather than program the calculation of (4-35), some statisticians recommend the equivalent procedure of fixing λ, creating the new variable

$$y_j^{(\lambda)} = \frac{x_j^{\lambda} - 1}{\lambda \left[\left(\prod_{i=1}^{n} x_i \right)^{1/n} \right]^{\lambda - 1}} \qquad j = 1, \ldots, n \tag{4-37}$$

and then calculating the sample variance. The minimum of the variance occurs at the same λ that maximizes (4-35).

Box-Cox Transformation: Transformation Continuous Function of λ

Box-Cox Transformation. The assumption that $h(\mu)$ is a power of μ in Table 3.1 led us to transformations that are powers of the original variable, and suggested the general class of power transformations. The search for a data-based approach to determine a transformation led Box and Cox (1964) to consider a method based on maximum likelihood estimation of the transformation parameter. The fundamental assumption of their method is that $y > 0$ and that there exists a power of y that has a linear model with prescribed mean function and satisfies the assumptions of normality and constant variance.

To include the logarithmic transformation as a special case of the power transformation, Box and Cox considered the following form, where the transformation on y is expressed as a function of the power parameter λ:

$$\begin{aligned} y(\lambda) &= \frac{y^{\lambda}-1}{\lambda} \qquad \lambda \neq 0 \\ y(0) &= log(y) \quad \lambda = 0 \end{aligned} \tag{3.15}$$

Note that the definition for $\lambda = 0$ is the limit of the first expression in (3.15) as $\lambda \to 0$. It follows that inclusion of this special case makes the transformation a continuous function of λ.

Estimating Lambda or λ (Power to Which The Variable Values to Be Raised)

Practical Assessment, Research & Evaluation, Vol 15, No 12
Osborne, Applying Box-Cox

While not implemented in all statistical packages[2], there are ways to estimate lambda, the Box-Cox transformation coefficient using any statistical package or by hand to estimate the effects of a selected range of λ automatically. This is discussed in detail in the appendix. Given that λ can take on an almost infinite number of values, we can theoretically calibrate a transformation to be maximally effective in moving a variable toward normality, regardless of whether it is negatively or positively skewed.[3] Additionally, as mentioned above, this family of transformations incorporates many traditional transformations:

λ = 1.00: no transformation needed; produces results identical to original data
λ = 0.50: square root transformation
λ = 0.33: cube root transformation
λ = 0.25: fourth root transformation
λ = 0.00: natural log transformation
λ = -0.50: reciprocal square root transformation
λ = -1.00: reciprocal (inverse) transformation
and so forth.

Box-Cox Transforms of Horse-Kicks With Various λ

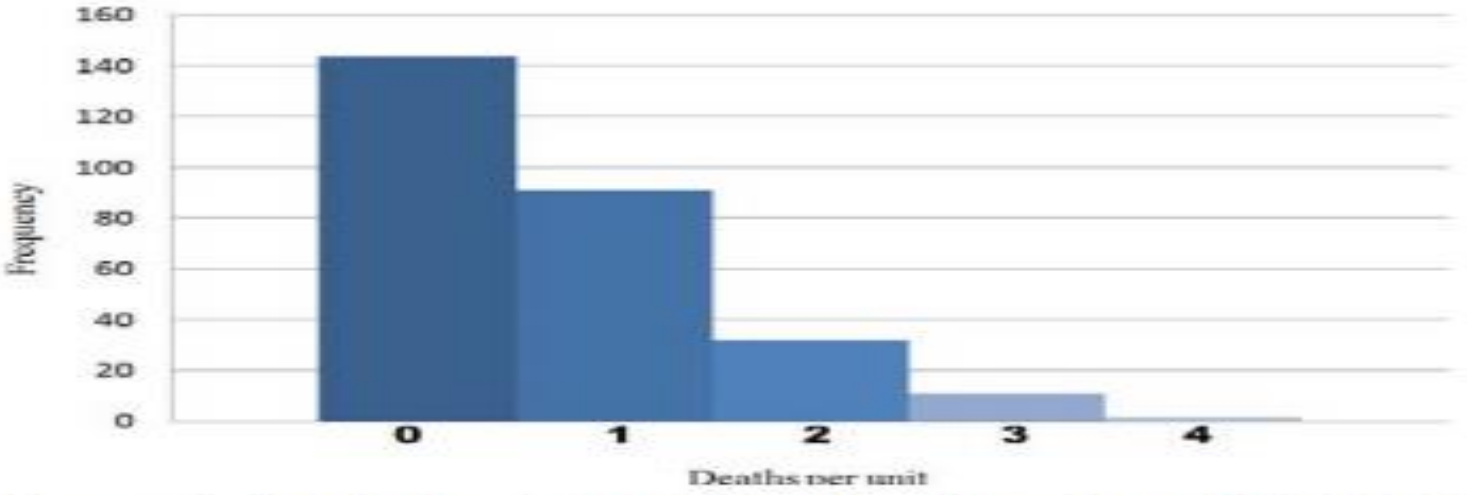


Figure 1. *Deaths from horse kicks, Prussian Army 1875-1894*

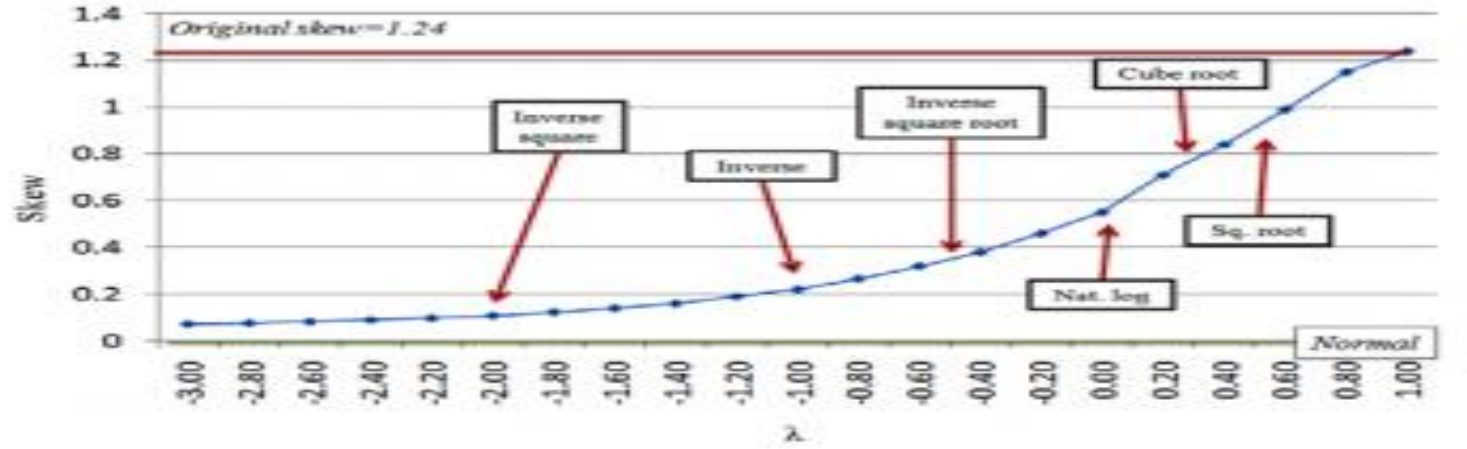


Figure 2.*Box-Cox transforms of horse-kicks with various λ*

Transforming Does Not Always Improve Approximation to Normality

Comment. It is now understood that the transformation obtained by maximizing $\ell(\lambda)$ usually improves the approximation to normality. However, there is no guarantee that even the best choice of λ will produce a transformed set of values that adequately conform to a normal distribution. The outcomes produced by a transformation selected according to (4-35) should always be carefully examined for possible violations of the tentative assumption of normality. This warning applies with equal force to transformations selected by any other technique.

Determining Power Transformation From the Frequency Distribution Or Univariate Graph

Example 4.16 (Determining a power transformation for univariate data) We gave readings of the microwave radiation emitted through the closed doors of $n = 42$ ovens in Example 4.10. The Q–Q plot of these data in Figure 4.6 indicates that the observations deviate from what would be expected if they were normally distributed. Since all the observations are positive, let us perform a power transformation of the data which, we hope, will produce results that are more nearly normal. Restricting our attention to the family of transformations in (4-34), we must find that value of λ maximizing the function $\ell(\lambda)$ in (4-35).

The pairs $(\lambda, \ell(\lambda))$ are listed in the following table for several values of λ:

λ	$\ell(\lambda)$	λ	$\ell(\lambda)$
−1.00	70.52		
−.90	75.65	.40	106.20
−.80	80.46	.50	105.50
−.70	84.94	.60	104.43
−.60	89.06	.70	103.03
−.50	92.79	.80	101.33
−.40	96.10	.90	99.34
−.30	98.97	1.00	97.10
−.20	101.39	1.10	94.64
−.10	103.35	1.20	91.96
.00	104.83	1.30	89.10
.10	105.84	1.40	86.07
.20	106.39	1.50	82.88
.30	106.51		

Value of Maximizing λ

Restricting our attention to the family of transformations in (4-34), we must find that value of λ maximizing the function $\ell(\lambda)$ in (4-35).

The pairs $(\lambda, \ell(\lambda))$ are listed in the following table for several values of λ:

λ	$\ell(\lambda)$	λ	$\ell(\lambda)$
−1.00	70.52		
−.90	75.65	.40	106.20
−.80	80.46	.50	105.50
−.70	84.94	.60	104.43
−.60	89.06	.70	103.03
−.50	92.79	.80	101.33
−.40	96.10	.90	99.34
−.30	98.97	1.00	97.10
−.20	101.39	1.10	94.64
−.10	103.35	1.20	91.96
.00	104.83	1.30	89.10
.10	105.84	1.40	86.07
.20	106.39	1.50	82.88
.30	106.51		

Determining Lambda

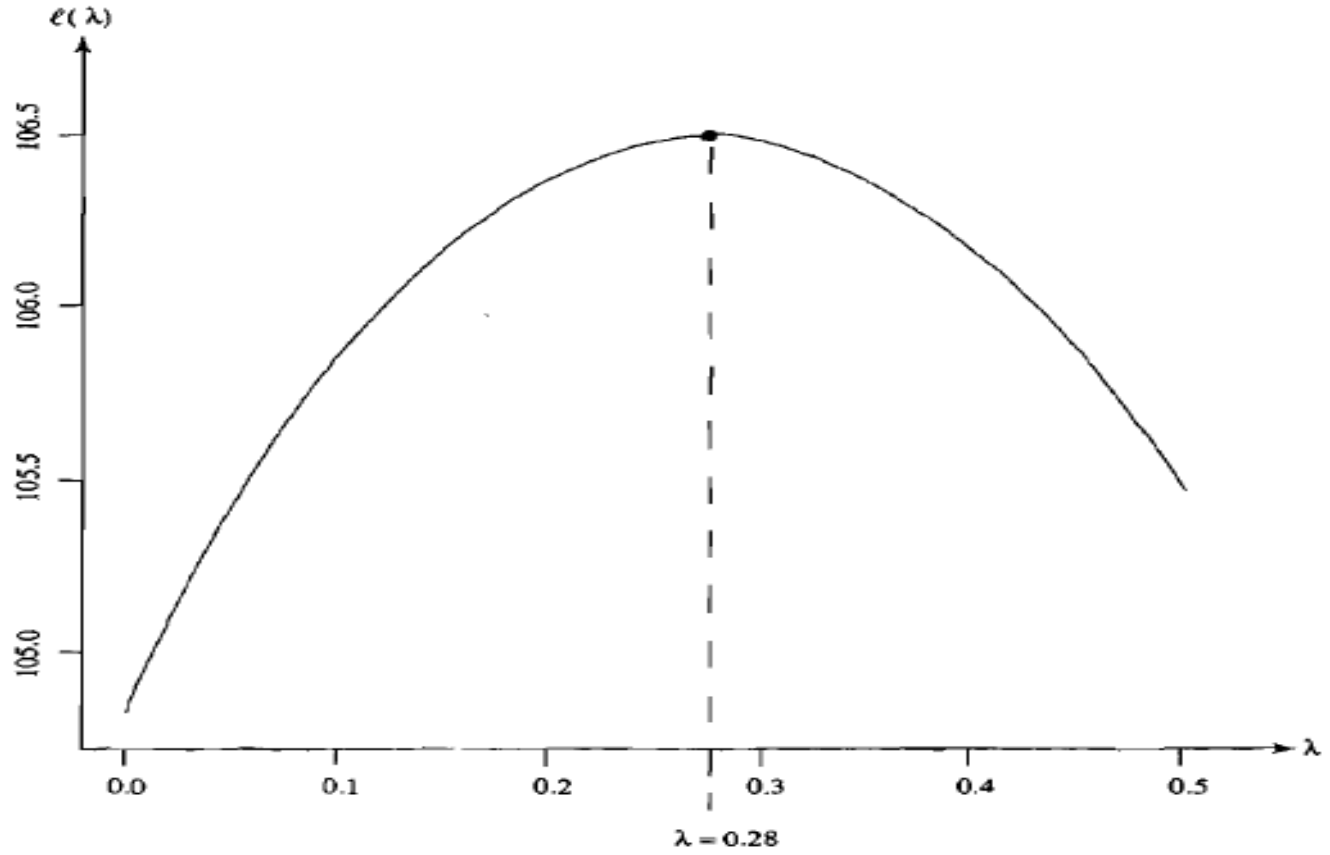


Figure 4.12 Plot of $\ell(\lambda)$ versus λ for radiation data (door closed).

The curve of $\ell(\lambda)$ versus λ that allows the more exact determination $\hat{\lambda} = .28$ is shown in Figure 4.12 .

Selecting λ = .25 or .30

The curve of $\ell(\lambda)$ versus λ that allows the more exact determination $\hat{\lambda} = .28$ is shown in Figure 4.12.

It is evident from both the table and the plot that a value of $\hat{\lambda}$ around .30 maximizes $\ell(\lambda)$. For convenience, we choose $\hat{\lambda} = .25$. The data x_j were reexpressed as

$$x_j^{(1/4)} = \frac{x_j^{1/4} - 1}{\frac{1}{4}} \qquad j = 1, 2, \ldots, 42$$

and a Q–Q plot was constructed from the transformed quantities. This plot is shown in Figure 4.13 on page 196. The quantile pairs fall very close to a straight line, and we would conclude from this evidence that the $x_j^{(1/4)}$ are approximately normal. ■

Q-Q Plot of Transformed Radiation Data

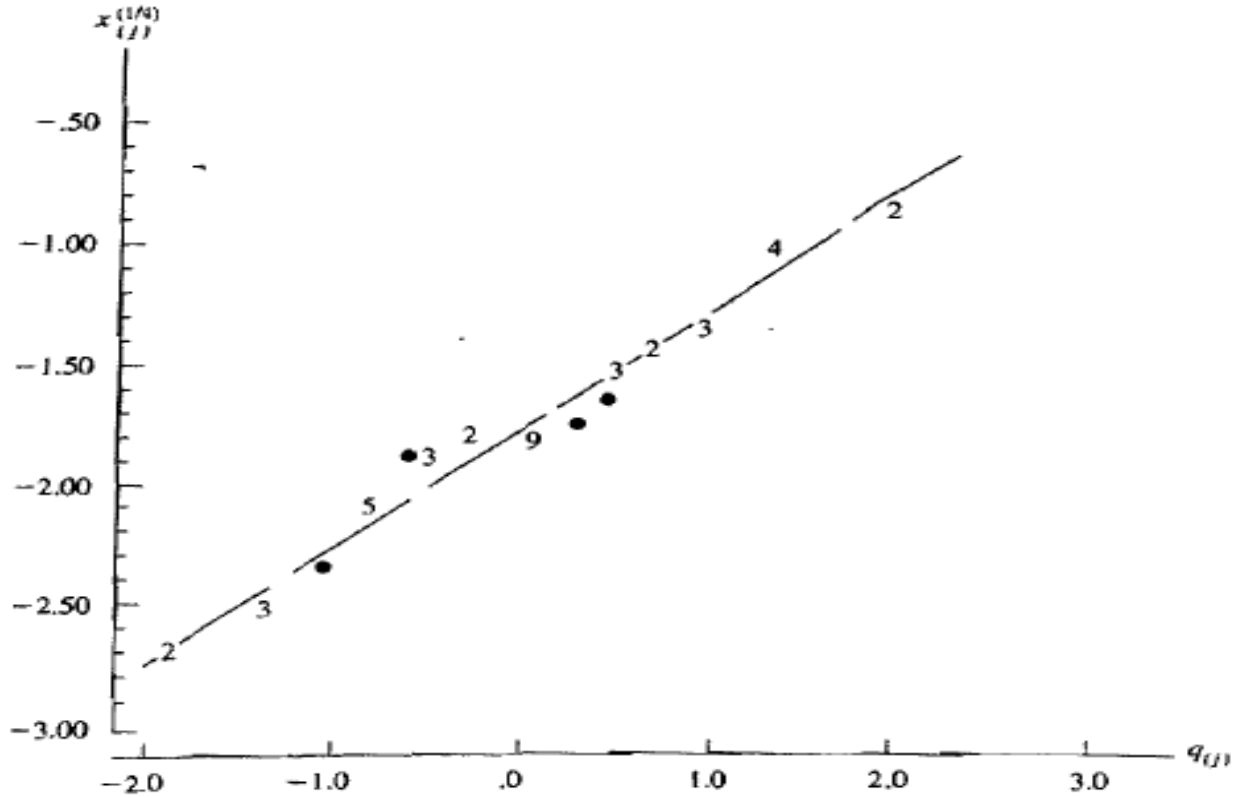


Figure 4.13 A Q–Q plot of the transformed radiation data (door closed). (The integers in the plot indicate the number of points occupying the same location.)