Chapter 5

Continuous Probability Distribution

Introduction

- The normal, or Gaussian or "bell-shaped," distribution is the cornerstone of most methods of estimation and hypothesis testing.

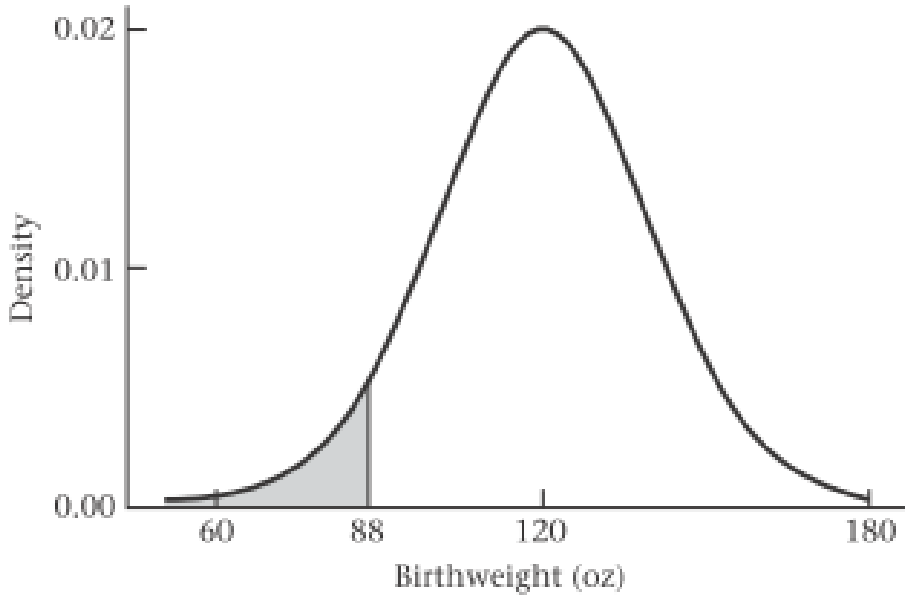
- Many random variables, such as distribution of birth weights or blood pressures in the general population, tend to follow approximately a normal distribution.

- Those variables that are not themselves normal oftentimes are closely approximated by a normal distribution when summed.

- Using normal distribution is desirable since it is easy to use and tables for it are more widely available than are tables for other distributions.

The cumulative-distribution function for the random variable X evaluated at the point a is defined as the probability that X will take on values ≤ a. It is represented by the area under the pdf to the left of a.

Figure 5.3 The pdf for birthweight



Generally, a distinction is not made between $Pr(X < x)$ and $Pr(X \leq x)$ when X is a continuous random variable because they represent the same quantity since the probability of individual values is 0; that is, $Pr(X = x) = 0$.

Expected value and variance for continuous random variables have the same meaning as for discrete random variables.

The **expected value** of a continuous random variable X, denoted by $E(X)$, or μ, is the average value taken on by the random variable.

The variance of a continuous random variable X, denoted by Var(X) or σ^2 is the average squared distance of each value of the random variable from its expected value, which is given by $E(X - \mu)^2$ and can be re-expressed in short form as $E(X^2) - \mu^2$. The standard deviation, or σ, is the squared root of the variance, that is, $\sigma = \sqrt{Var(X)}$.

5.3 Normal Distribution

Figure 5.4 Karl Friedrich Gauss (1777–1855)

Normal distribution is the most widely used continuous distribution.
It is vital to statistical work.

It is also frequently called Gaussian distribution, after the well-known mathematician Karl Friedrich Gauss.

Normal distribution is generally more convenient to work with than any other distribution.

Body weights or DBPs approximately follow a normal distribution. Other distributions that are not themselves normal can be made approximately normal by transforming data onto a different scale, such as a logarithmic scale.

Usually, any random variable that can be expressed as a sum of many other random variables can be well approximated by a normal distribution.

Example., many physiologic measures are determined in part by a combination of several genetic and environmental risk factors can often be well approximated by a normal distribution.

EXAMPLE 5.9

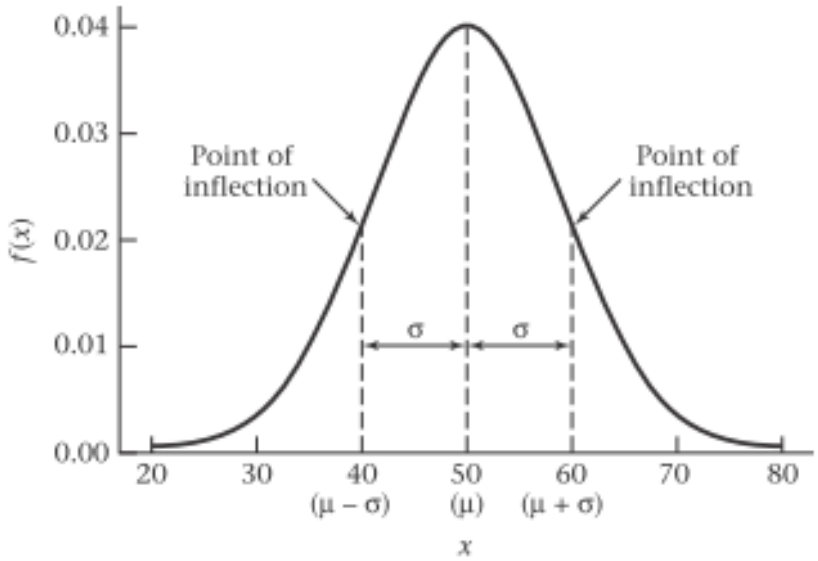
Infectious Disease The number of lymphocytes in a differential of 100 white blood cells (see Example 4.15 for the definition of a differential) tends to be normally distributed because this random variable is a sum of 100 random variables, each representing whether or not an individual cell is a lymphocyte.

Most estimation procedures and hypothesis tests assume the random variable being considered has an underlying normal distribution.

An important area of application of normal distribution is as an approximating distribution to other distributions. It is defined by its pdf, which is given as,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad -\infty < x < \infty$$

Figure 5.5 The pdf for a normal distribution with mean μ (50) and variance σ^2 (100)



The density function follows a bell-shaped curve, with the mode at μ and the most frequently occurring values around μ. The curve is symmetric around μ, with points of inflection on either side of μ at μ - σ and μ + σ.

A point of inflection is a point at which the slope of the curve changes direction. Distance from μ to the points of inflection are an indicator of magnitude of σ.

Using calculus methods, it can be shown that μ and σ^2 are the expected value and variance, respectively, of the distribution.

A normal distribution with mean μ and variance σ^2 will generally be referred to as an $N(\mu,\sigma^2)$ distribution.

The height of the normal distribution is always $1/(\sqrt{2\pi}\sigma)$.

Figure 5.6 Comparison of two normal distributions with the same variance and different means

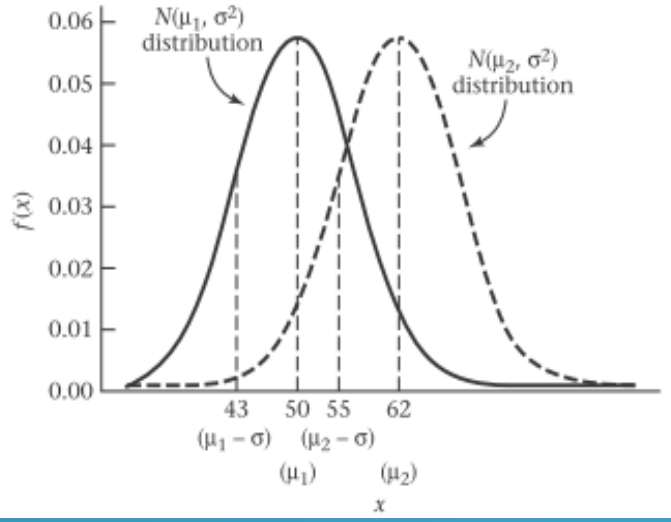
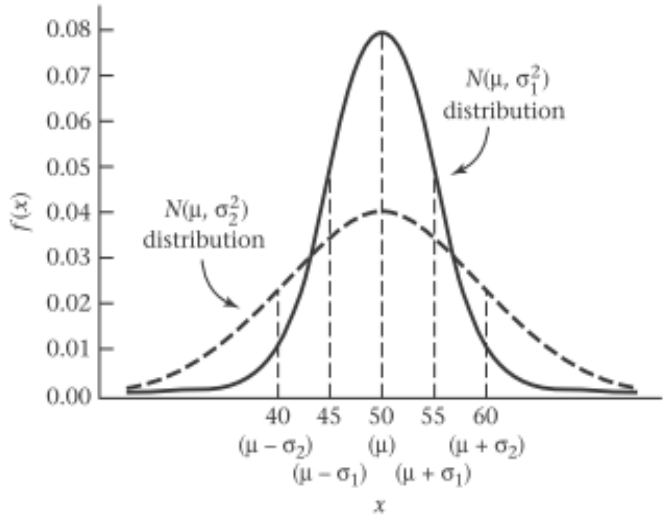


Figure 5.7 Comparison of two normal distributions with the same means and different variances

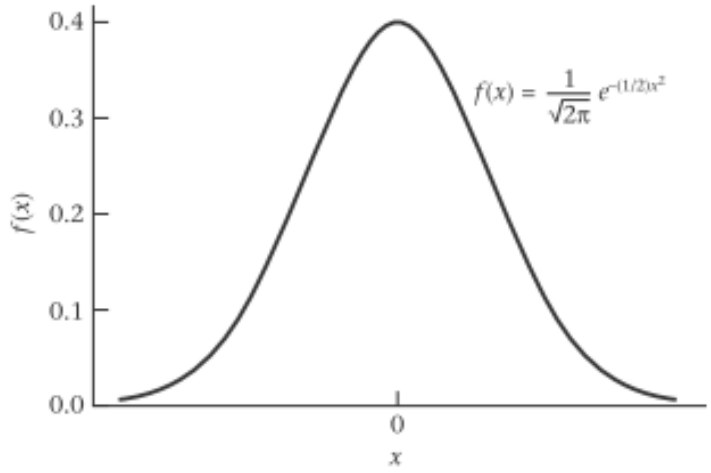


A normal distribution with mean 0 and variance 1 is called a standard, or unit, normal distribution.

This distribution is also called $\mathcal{N}(0,1)$ distribution.

5.4 properties of the standard Normal distribution

Figure 5.8 The pdf for a standard normal distribution



$$f(x) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)x^2}, \quad -\infty < x < +\infty$$

This distribution is symmetric about 0, because f(x) = f(-x).

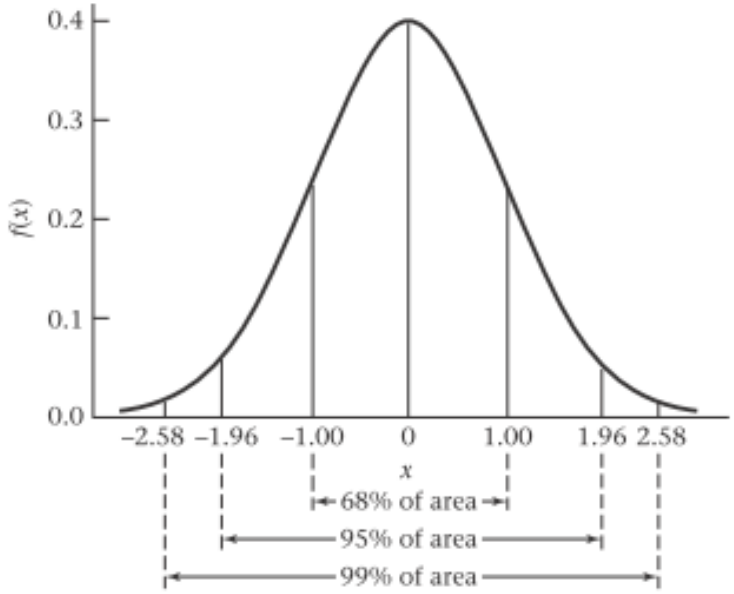
About 68% of the area under the standard normal density lies between +1 and -1, about 95% of the area lies between +2 and -2, and about 99% lies between +2.5 and -2.5.

Pr(-1 < X < 1) = 0.6827

Pr(-1.96 < X < 1.96) = 0.95

Pr(-2.576 < X < 2.576) = 0.99

Figure 5.9 Empirical properties of the standard normal distribution



The cumulative-distribution function (cdf) for a standard normal distribution is denoted by $\phi(x) = Pr(X \leq x)$ where X follows an N(0,1) distribution.

The symbol ~ is used as shorthand for the phase "is distributed as." Thus, X ~ N(0,1) means that the random variable X is distributed as N(0,1) distribution.

Figure 5.10 The cdf [$\Phi(x)$] for a standard normal distribution

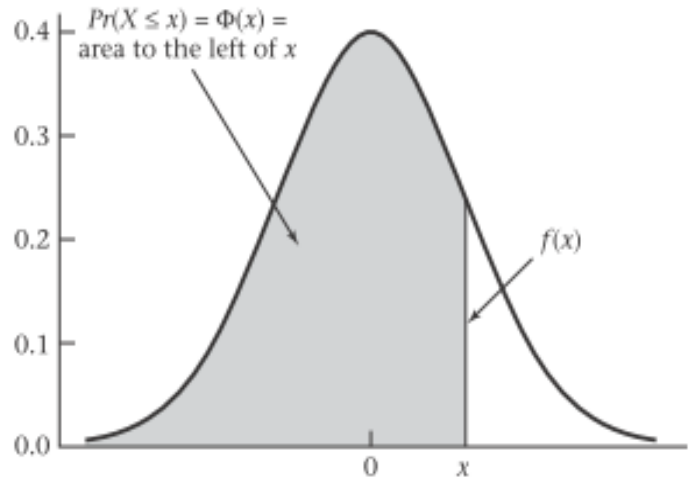
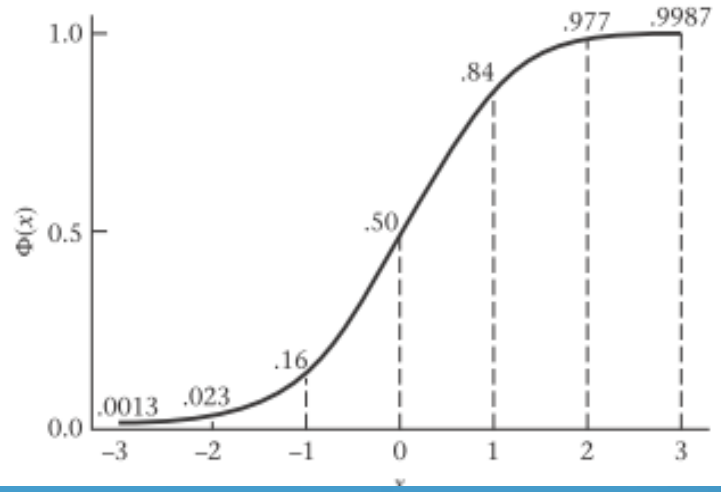


Figure 5.11 The cdf for a standard normal distribution [$\Phi(x)$]

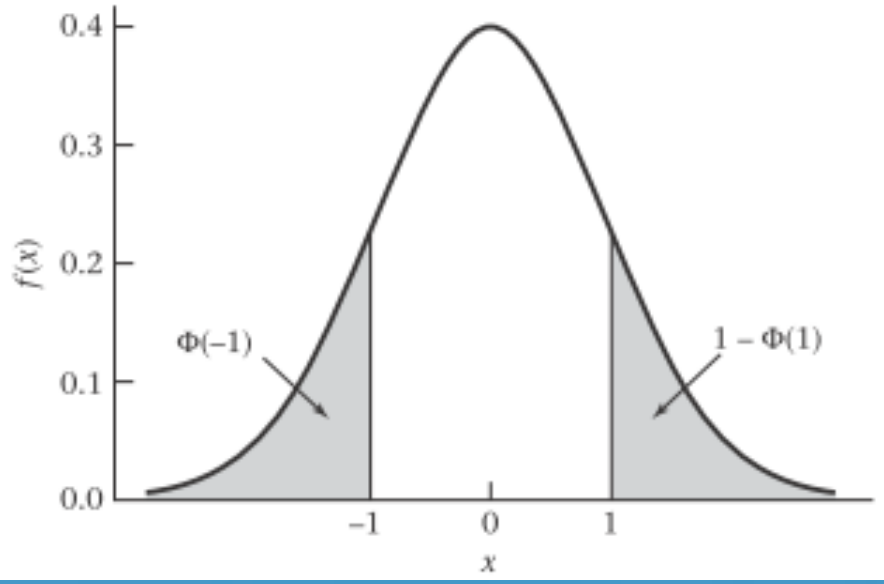


The area to the left of x approaches 0 as x becomes small and approaches 1 as x becomes large.

Symmetry Properties of the Standard Normal Distribution

$$\phi(-x) = Pr(X \leq -x) = Pr(X \geq x) = 1 - pr(X \leq x) = 1 - \phi(x)$$

Figure 5.12 Illustration of the symmetry properties of the normal distribution



A normal range for a biological quantity is often defined by a range within x standard deviations of the mean for some specified value of x.
The probability of a value falling in this range is given by $Pr(-x \leq X \leq x)$ for a standard normal distribution.

EXAMPLE 5.12 Calculate $Pr(X \le -1.96)$ assuming $X \sim N(0,1)$.

Solution: $Pr(X \le -1.96) = Pr(X \ge 1.96) = .0250$ from column B of Table 3.

Furthermore, for any numbers a, b we have $Pr(a \le X \le b) = Pr(X \le b) - Pr(X \le a)$ and thus, we can evaluate $Pr(a \le X \le b)$ for any a, b from Table 3.

EXAMPLE 5.13 Compute $Pr(-1 \le X \le 1.5)$ assuming $X \sim N(0,1)$.

Solution:
$$Pr(-1 \le X \le 1.5) = Pr(X \le 1.5) - Pr(X \le -1)$$
$$= Pr(X \le 1.5) - Pr(X \ge 1) = .9332 - .1587$$
$$= .7745$$

EXAMPLE 5.14

Pulmonary Disease Forced vital capacity (FVC), a standard measure of pulmonary function, is the volume of air a person can expel in 6 seconds. Current research looks at potential risk factors, such as cigarette smoking, air pollution, indoor allergens, or the type of stove used in the home, that may affect FVC in grade-school children. One problem is that age, gender, and height affect pulmonary function, and these variables must be corrected for before considering other risk factors. One way to make these adjustments for a particular child is to find the mean μ and standard deviation σ for children of the same age (in 1-year age groups), gender, and height (in 2-in. height groups) from large national surveys and compute a **standardized FVC**, which is defined as $(X - \mu)/\sigma$, where X is the original FVC. The standardized FVC then approximately follows an $N(0,1)$ distribution, if the distribution of the original FVC values was bell-shaped. Suppose a child is considered in poor pulmonary health if his or her standardized FVC < -1.5. What percentage of children are in poor pulmonary health?

Solution: $Pr(X < -1.5) = Pr(X > 1.5) = .0668$

Thus, about 7% of children are in poor pulmonary health.

EXAMPLE 5.15

Pulmonary Disease Suppose a child is considered to have normal lung growth if his or her standardized FVC is within 1.5 standard deviations of the mean. What proportion of children are within the normal range?

Solution: Compute $Pr(-1.5 \leq X \leq 1.5)$. Under 1.50 in Table 3, column D gives this quantity as .8664. Thus, about 87% of children have normal lung growth, according to this definition.

Finally, column C of Table 3 provides the area under the standard normal density from 0 to x because these areas occasionally prove useful in work on statistical inference.

EXAMPLE 5.16

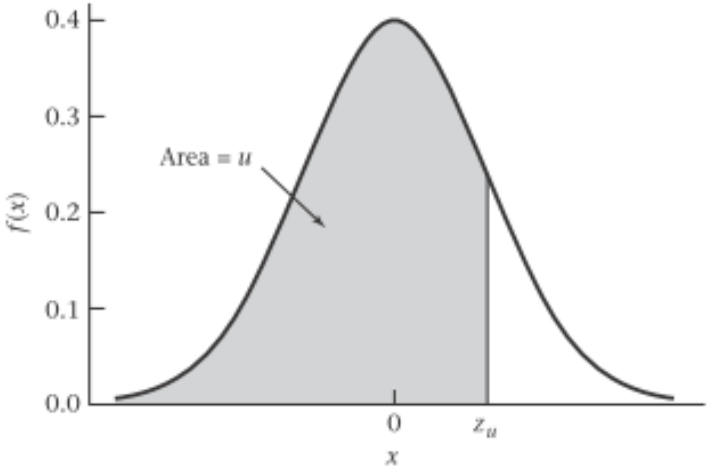
Find the area under the standard normal density from 0 to 1.45.

Solution: Refer to column C of Table 3 under 1.45. The appropriate area is given by .4265.

Using Electronic Tables for the Normal Distribution

In Excel 2007, the function NORMDIST(x) provides the cdf for a standard normal distribution for any value of x.

Figure 5.13 Graphic display of the $(100 \times u)$th percentile of a standard normal distribution (z_u)



The $(100 \times u)$th percentile of a standard normal distribution is denoted by z_u

$\Pr(X < z_u) = u$ *where* $X \sim N(0,1)$

The function z_u is sometimes referred to as the inverse normal function.

To evaluate z_u we determine the area u in the normal tables and then find the value z_u that corresponds to this area.

If $u < 0.5$, then we use the symmetry properties of the normal distribution to obtain $z_u = -z_{1-u}$, where z_{1-u} can be obtained from the normal table.

EXAMPLE 5.18

Compute $z_{.975}$, $z_{.95}$, $z_{.5}$, and $z_{.025}$.

Solution: From Table 3 we have

$$\Phi(1.96) = .975$$

$$\Phi(1.645) = .95$$

$$\Phi(0) = .5$$

$$\Phi(-1.96) = 1 - \Phi(1.96) = 1 - .975 = .025$$

Thus, $z_{.975} = 1.96$

$z_{.95} = 1.645$

$z_{.5} = 0$

$z_{.025} = -1.96$

where for $z_{.95}$ we interpolate between 1.64 and 1.65 to obtain 1.645.

5.5 Conversion from an $N(\mu,\sigma^2)$ Distribution to an $N(0,1)$ distribution

If $X \sim N(\mu,\sigma^2)$, then what is $Pr(a < X < b)$ for any a,b?

Consider the random variable $Z = (X - \mu)/\sigma$.

If $X \sim N(\mu,\sigma^2)$ and $Z = (X - \mu)/\sigma$, then $Z \sim n(0,1)$

Evaluation of Probabilities for Any Normal Distribution via Standardization

If $X \sim N(\mu,\sigma^2)$ and $Z = (X - \mu)/\sigma$

This is known as standardization of a normal variable.

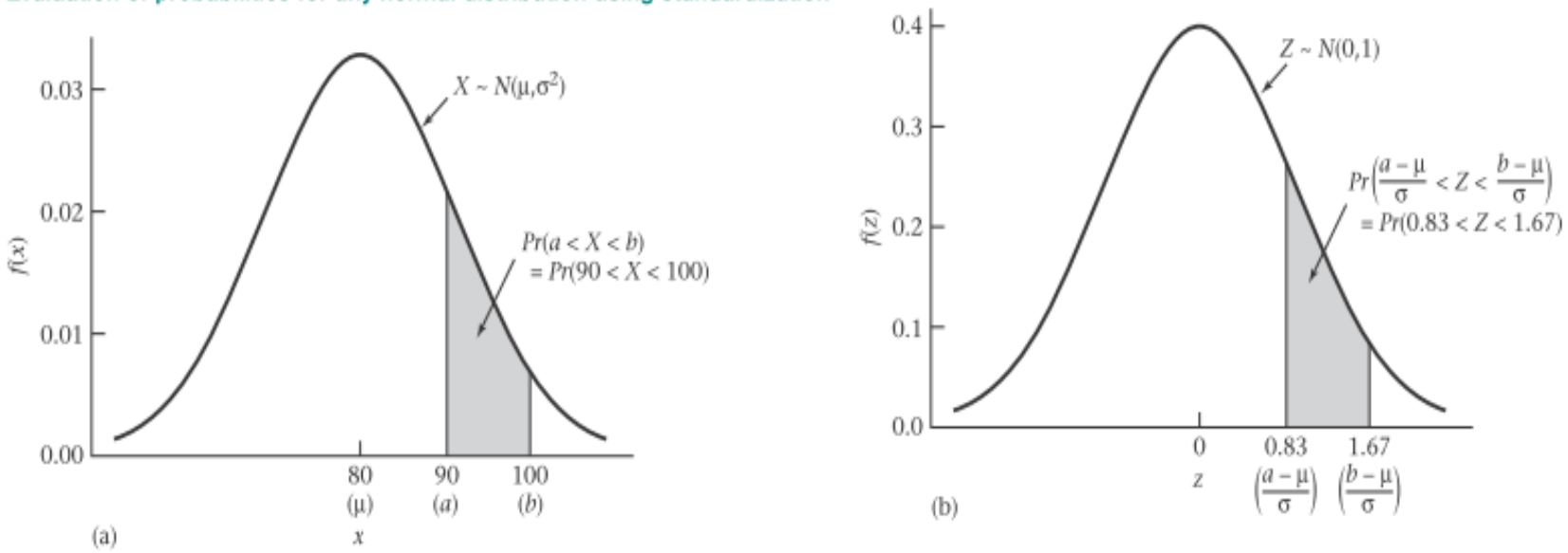
$$Pr(a < X < b) = Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi[(b-\mu)/\sigma] - \Phi[(a-\mu)/\sigma]$$

For any probability expression concerning normal random variables of the form $Pr(a < X < b)$, the population mean μ is subtracted from each boundary point and divided by the standard deviation σ to obtain an equivalent probability expression for the standard normal random variable Z,

$$Pr[(a - \mu)/\sigma < Z < (b - \mu)/\sigma]$$

The standard normal tables are then used to evaluate this latter probability.

Figure 5.14 Evaluation of probabilities for any normal distribution using standardization



EXAMPLE 5.20

Solution: The probability of being a mild hypertensive among the group of 35- to 44-year-old men can now be calculated.

$$Pr(90 < X < 100) = Pr\left(\frac{90-80}{12} < Z < \frac{100-80}{12}\right)$$

$$= Pr(0.833 < Z < 1.667) = \Phi(1.667) - \Phi(0.833)$$

$$= .9522 - .7977 = .155$$

Thus, about 15.5% of this population will have mild hypertension.

EXAMPLE 5.21

Botany Suppose tree diameters of a certain species of tree from some defined forest area are assumed to be normally distributed with mean = 8 in. and standard deviation = 2 in. Find the probability of a tree having an unusually large diameter, which is defined as >12 in.

Solution: We have $X \sim N(8,4)$ and require

$$Pr(X > 12) = 1 - Pr(X < 12) = 1 - Pr\left(Z < \frac{12-8}{2}\right)$$

$$= 1 - Pr(Z < 2.0) = 1 - .977 = .023$$

Thus, 2.3% of trees from this area have an unusually large diameter.

The pth percentile of a general normal distribution (x) can also be written in terms of the percentiles of a standard normal distribution as follows: $x = \mu + z_p\sigma$

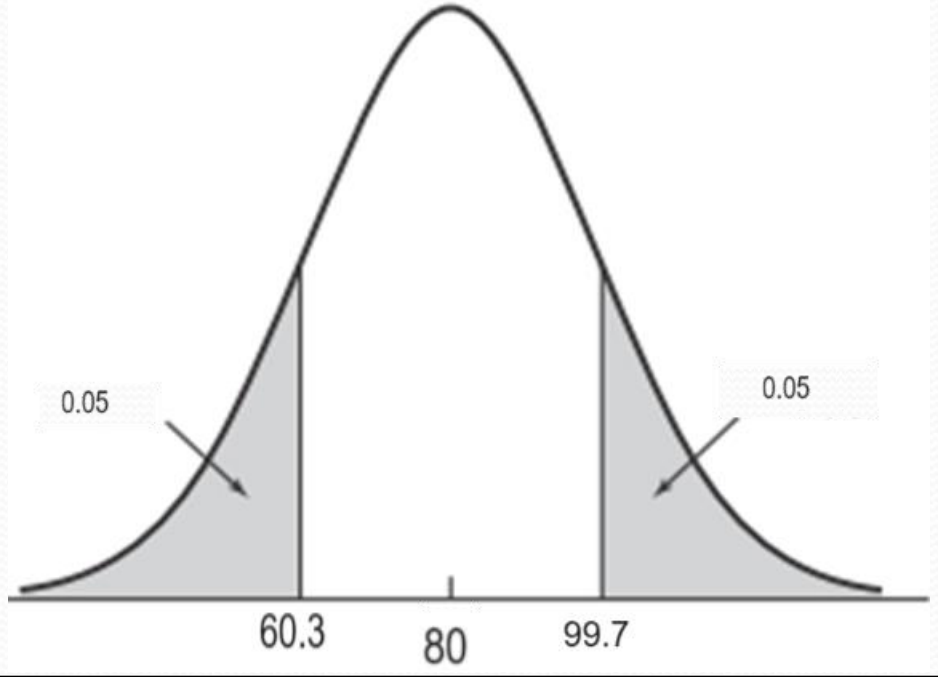
EXAMPLE 5.24

Hypertension Suppose the distribution of DBP in 35- to 44-year-old men is normally distributed with mean = 80 mm Hg and variance = 144 mm Hg. Find the upper and lower fifth percentiles of this distribution.

Solution: We could do this either using Table 3 (Appendix) or using a computer program. If we use Table 3 and we denote the upper and lower 5th percentiles by $x_{.05}$ and $x_{.95}$, respectively, then from Equation 5.7 we have

$$x_{.05} = 80 + z_{.05}(12)$$
$$= 80 - 1.645(12) = 60.3 \text{ mm Hg}$$

$$x_{.95} = 80 + z_{.95}(12)$$
$$= 80 + 1.645(12) = 99.7 \text{ mm Hg}$$



Summary

In this chapter, we discussed

- Continuous random variables
- Concepts of expected value, variance, and cumulative distribution for continuous random variables
- Normal distribution, which is the most important continuous distribution
- The two parameters: mean μ and variance σ^2
- Normal tables, which are used when working with standard normal distribution
- Electronic tables can be used to evaluate areas and/or percentiles for any normal distribution

The End