4 Continuous Random Variables and Probability Distributions

4.3 The Normal distribution

Copyright © Cengage Learning. All rights reserved.

The Normal Distribution

The normal distribution is the most important one in all of probability and statistics. Many numerical populations have distributions that can be fit very closely by an appropriate normal curve.

Examples include heights, weights, and other physical characteristics (the famous 1903 *Biometrika* article “On the Laws of Inheritance in Man” discussed many examples of this sort), measurement errors in scientific experiments, anthropometric measurements on fossils, reaction times in psychological experiments, measurements of intelligence and aptitude, scores on various tests, and numerous economic measures and indicators.

The Normal Distribution

Definition

A continuous rv X is said to have a **normal distribution** with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

คำนวณยาก ใช้ table

$$\boxed{f(x;\,\mu,\,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty} \qquad (4.3)$$

Again e denotes the base of the natural logarithm system and equals approximately 2.71828, and π represents the familiar mathematical constant with approximate value 3.14159.

The Normal Distribution

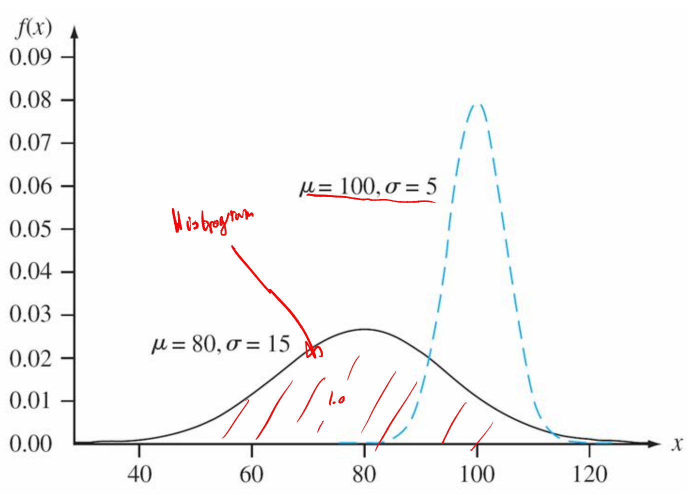
The statement that X is normally distributed with parameters μ and σ^2 is often abbreviated $X \sim N(\mu, \sigma^2)$.

Clearly $f(x; \mu, \sigma) \geq 0$, but a somewhat complicated calculus argument must be used to verify that $\int_{-\infty}^{\infty} f(x; \mu, \sigma)\, dx = 1$.

It can be shown that $E(X) = \mu$ and $V(X) = \sigma^2$, so the parameters are the mean and the standard deviation of X.

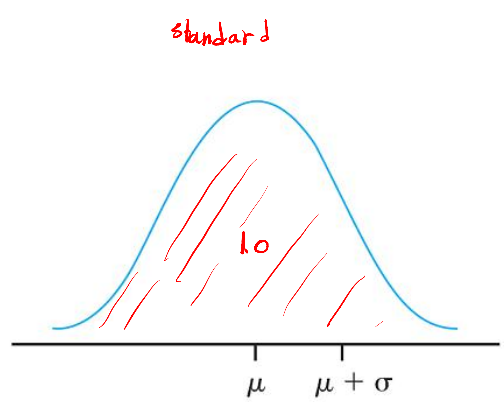
The Normal Distribution

Figure 4.13 presents graphs of $f(x; \mu, \sigma)$ for several different (μ, σ) pairs.



Two different normal density curves

Figure 4.13(a)



Visualizing μ and σ for a normal distribution

Figure 4.13(b)

The Normal Distribution

Each density curve is symmetric about μ and bell-shaped, so the center of the bell (point of symmetry) is both the mean of the distribution and the median.

The value of σ is the distance from μ to the inflection points of the curve (the points at which the curve changes from turning downward to turning upward).

The Normal Distribution

Large values of σ yield graphs that are quite spread out about μ, whereas small values of σ yield graphs with a high peak above μ and most of the area under the graph quite close to μ.

Thus a large σ implies that a value of X far from μ may well be observed, whereas such a value is quite unlikely when σ is small.

The Standard Normal Distribution

The Standard Normal Distribution

The computation of $P(a \leq X \leq b)$ when X is a normal rv with parameters μ and σ requires evaluating

$$\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\,dx \qquad (4.4)$$

None of the standard integration techniques can be used to accomplish this. Instead, for $\mu = 0$ and $\sigma = 1$, Expression (4.4) has been calculated using numerical techniques and tabulated for certain values of a and b.

This table can also be used to compute probabilities for any other values of μ and σ under consideration.

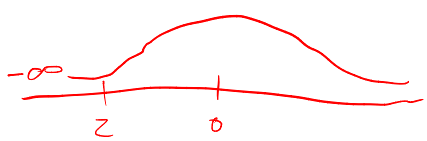
The Standard Normal Distribution

Definition
The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution.**
A random variable having a standard normal distribution is called a **standard normal random variable** and will be denoted by Z. The pdf of Z is

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty$$

The graph of $f(z; 0, 1)$ is called the *standard normal* (or z) curve. Its inflection points are at 1 and –1. The cdf of Z is $\boxed{P(Z \leq z) = \int_{-\infty}^{z} f(y; 0, 1)\, dy,}$ which we will denote by

CDF



$$\Phi(z).$$

The Standard Normal Distribution

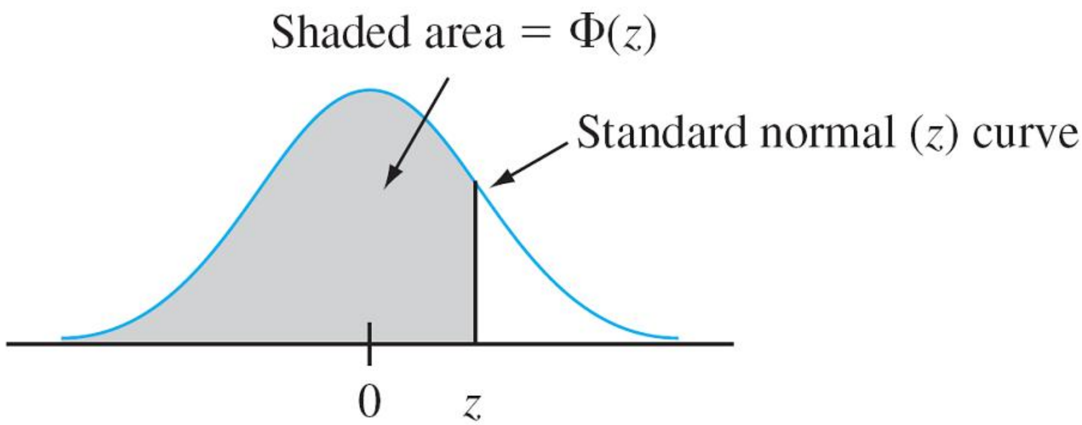
The standard normal distribution almost never serves as a model for a naturally arising population.

Instead, it is a reference distribution from which information about other normal distributions can be obtained.

Appendix Table A.3 gives $\Phi(z) = P(Z \leq z)$, the area under the standard normal density curve to the left of z, for z = –3.49, –3.48,..., 3.48, 3.49.

The Standard Normal Distribution

Figure 4.14 illustrates the type of cumulative area (probability) tabulated in Table A.3. From this table, various other probabilities involving *Z* can be calculated.



Standard normal cumulative areas tabulated in Appendix Table A.3

Figure 4.14

Example 13

Let's determine the following standard normal probabilities:

(a) $P(Z \leq 1.25)$, 0.8944 (b) $P(Z > 1.25)$, 1 − 0.8944

(c) $P(Z \leq -1.25)$, and (d) $P(-.38 \leq Z \leq 1.25)$.

0.1056 0.8944 − 0.352

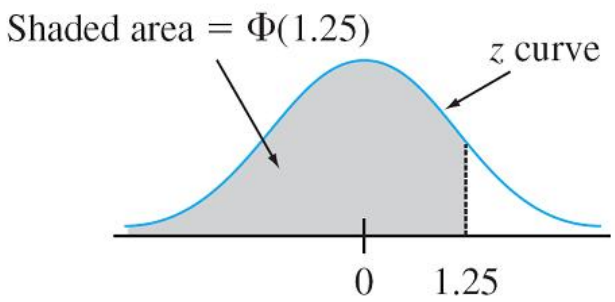
a. $P(Z \leq 1.25) = \Phi(1.25)$, a probability that is tabulated in Appendix Table A.3 at the intersection of the row marked 1.2 and the column marked .05.

The number there is .8944, so $P(Z \leq 1.25) = .8944$.

Example 13

cont'd

Figure 4.15(a) illustrates this probability.



Normal curve areas (probabilities) for Example 13

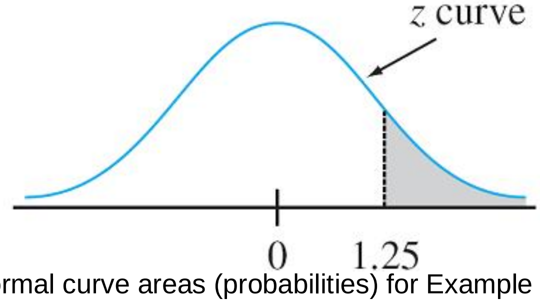
Figure 4.15(a)

b. $P(Z > 1.25) = 1 - P(Z \leq 1.25) = 1 - \Phi(1.25)$, the area under the z curve to the right of 1.25 (an upper-tail area). Then $\Phi(1.25) = .8944$ implies that $P(Z > 1.25) = .1056$.

Example 13

cont'd

Since Z is a continuous rv, $P(Z \geq 1.25) = .1056$. See Figure 4.15(b).



Normal curve areas (probabilities) for Example 13

Figure 4.15(b)

c. $P(Z \leq -1.25) = \Phi(-1.25)$, a lower-tail area. Directly from Appendix Table A.3, $\Phi(-1.25) = .1056$.

By symmetry of the z curve, this is the same answer as in part (b).

Example 13

cont'd

d. $P(-.38 \le Z \le 1.25)$ is the area under the standard normal curve above the interval whose left endpoint is –.38 and whose right endpoint is 1.25.

From Section 4.2, if X is a continuous rv with cdf $F(x)$, then $P(a \le X \le b) = F(b) - F(a)$.

Thus $P(-.38 \le Z \le 1.25) = \Phi(1.25) - \Phi(-.38)$

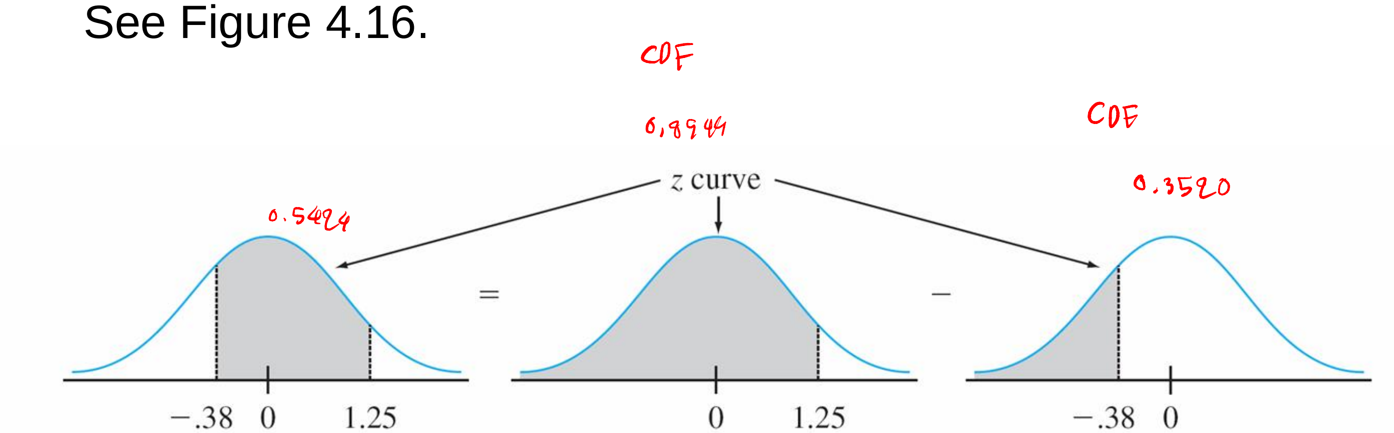
$= .8944 - .3520$

$= .5424$

Example 13

cont'd

See Figure 4.16.



$P(-.38 \le Z \le 1.25)$ as the difference between two cumulative areas

Figure 4.16

Percentiles of the Standard Normal Distribution

Percentiles of the Standard Normal Distribution

For any p between 0 and 1, Appendix Table A.3 can be used to obtain the $(100p)$th percentile of the standard normal distribution.

Example 14

99 %. STD Normal

The 99th percentile of the standard normal distribution is that value on the horizontal axis such that the area under the *z* curve to the left of the value is .9900. ← a.ฝั่งซ้าย

$= 100p$

Appendix Table A.3 gives for fixed *z* the area under the standard normal curve to the left of *z*, whereas here we have the area and want the value of *z*. This is the “inverse” problem to $P(Z \leq z) = ?$

so the table is used in an inverse fashion: Find in the middle of the table .9900; the row and column in which it lies identify the 99th *z* percentile.

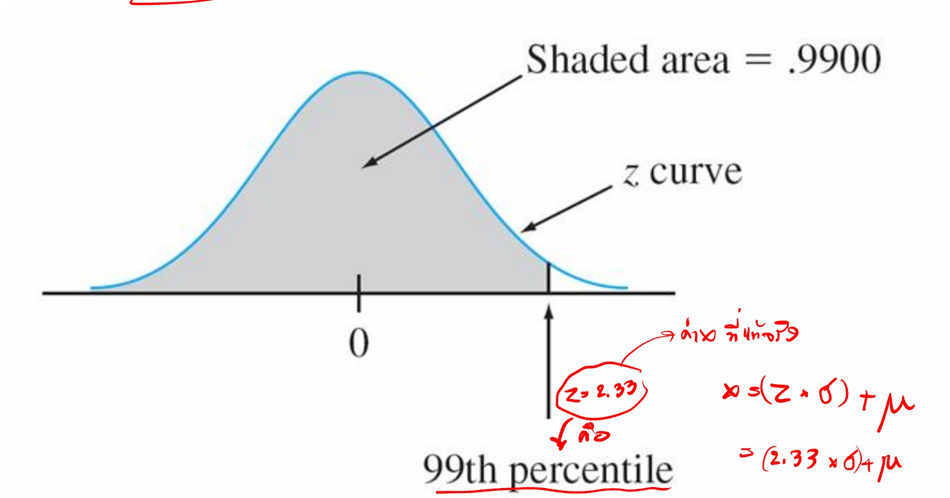
Example 14

cont’d

Here .9901 lies at the intersection of the row marked 2.3 and column marked .03, so the 99th percentile is (approximately) $z = 2.33$. (See Figure 4.17.)

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

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



$z = 2.33$ → ค่าx ที่แท้จริง

คือ

$x = (z \times \sigma) + \mu$

$= (2.33 \times \sigma) + \mu$

Finding the 99th percentile

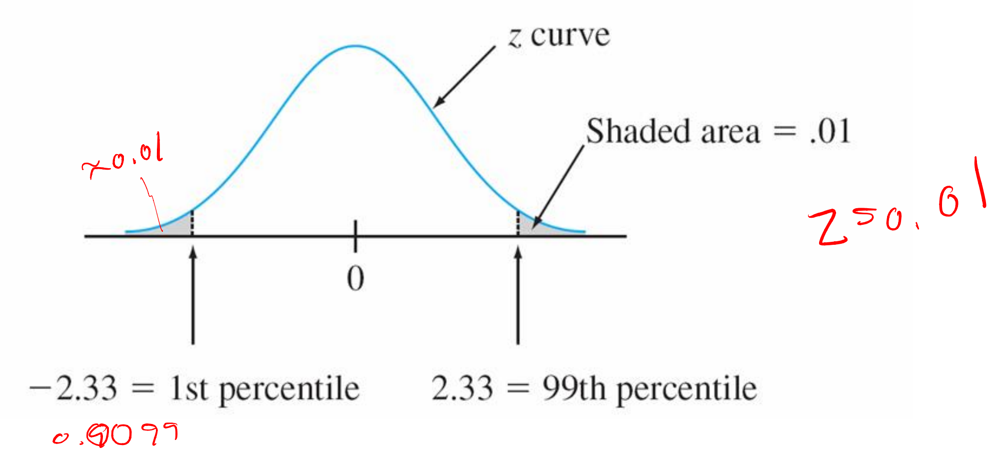
Figure 4.17

Example 14

cont'd

ทางซ้าย 1 %

By symmetry, the first percentile is as far below 0 as the 99th is above 0, so equals –2.33 (1% lies below the first and also above the 99th). (See Figure 4.18.)



The relationship between the 1st and 99th percentiles

Figure 4.18

Percentiles of the Standard Normal Distribution

In general, the (100p)th percentile is identified by the row and column of Appendix Table A.3 in which the entry p is found (e.g., the 67th percentile is obtained by finding .6700 in the body of the table, which gives z = .44).

If p does not appear, the number closest to it is often used, although linear interpolation gives a more accurate answer.

Percentiles of the Standard Normal Distribution

For example, to find the 95th percentile, we look for .9500 inside the table.

Although .9500 does not appear, both .9495 and .9505 do, corresponding to z = 1.64 and 1.65, respectively.

Since .9500 is halfway between the two probabilities that do appear, we will use 1.645 as the 95th percentile and –1.645 as the 5th percentile.

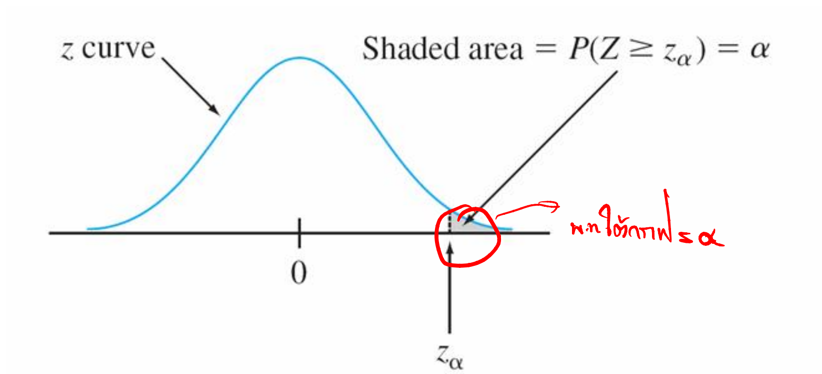
z_α Notation for z Critical Values

z_α Notation for *z* Critical Values

In statistical inference, we will need the values on the horizontal *z* axis that capture certain small tail areas under the standard normal curve.

Notation

z_α will denote the value on the *z* axis for which α of the area under the *z* curve lies to the right of z_α. (See Figure 4.19.)



z_α notation Illustrated

Figure 4.19

z_α Notation for *z* Critical Values

For example, $z_{.10}$ captures upper-tail area .10, and $z_{.01}$ captures upper-tail area .01.

Since α of the area under the *z* curve lies to the right of z_α, $1 - \alpha$ of the area lies to its left. Thus z_α *is the* $100(1 - \alpha)$*th percentile of the standard normal distribution*.

By symmetry the area under the standard normal curve to the left of $-z_\alpha$ is also α. The z_α´s are usually referred to as ***z* critical values.**

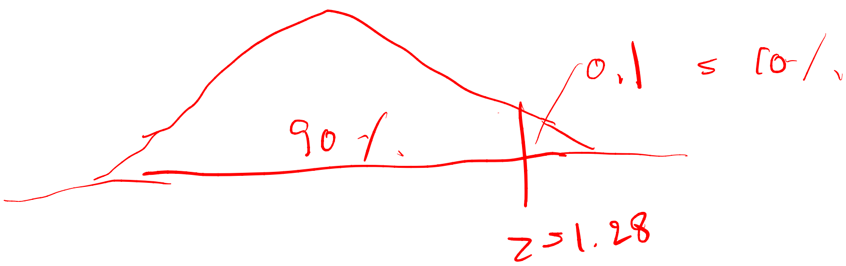
z_α Notation for z Critical Values

Table 4.1 lists the most useful z percentiles and z_α values.

Percentile	90	95	97.5	99	99.5	99.9	99.95
α (tail area)	.1	.05	.025	.01	.005	.001	.0005
$z_\alpha = 100(1-\alpha)$th percentile	1.28	1.645	1.96	2.33	2.58	3.08	3.27

Standard Normal Percentiles and Critical Values

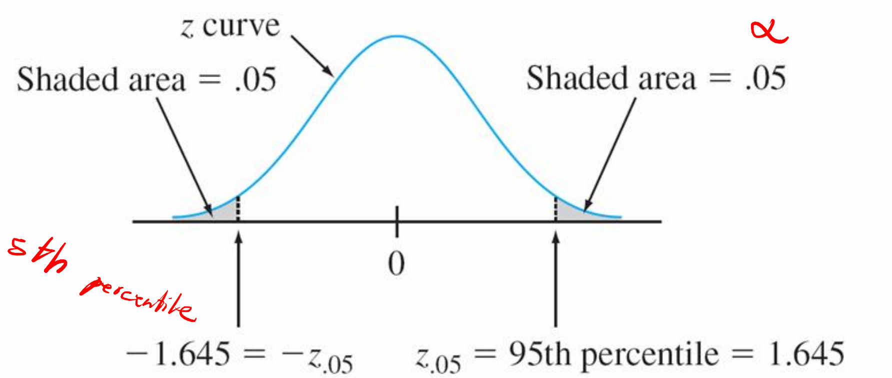
Table 4.1



Example 15

$z_{.05}$ is the 100(1 – .05)th = 95th percentile of the standard normal distribution, so $z_{.05}$ = 1.645.

The area under the standard normal curve to the left of $-z_{.05}$ is also .05. (See Figure 4.20.)



Finding $z_{.05}$

Figure 4.20

Nonstandard Normal Distributions

Nonstandard Normal Distributions

When $X \sim N(\mu, \sigma^2)$, probabilities involving X are computed by "standardizing." The **standardized variable** is $(X - \mu)/\sigma$.

Subtracting μ shifts the mean from μ to zero, and then dividing by σ scales the variable so that the standard deviation is 1 rather than σ.

Proposition
If X has a normal distribution with mean μ and standard deviation σ, then

$$Z = \frac{X - \mu}{\sigma}$$

Nonstandard Normal Distributions

has a standard normal distribution. Thus

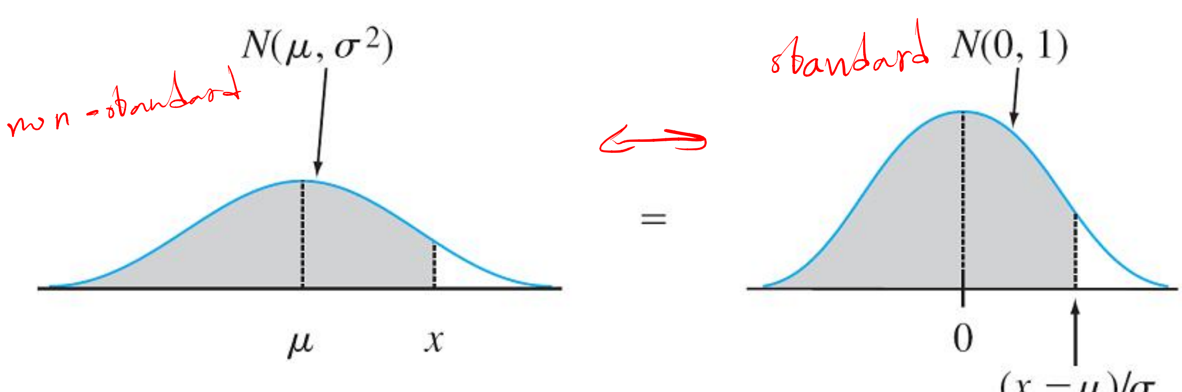
$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$ เปิดตาราง

$$P(X \le a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad P(X \ge b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Nonstandard Normal Distributions

The key idea of the proposition is that by standardizing, any probability involving *X* can be expressed as a probability involving a standard normal rv *Z*, so that Appendix Table A.3 can be used. This is illustrated in Figure 4.21.



Equality of nonstandard and standard normal curve areas

Figure 4.21

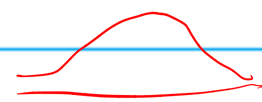
Nonstandard Normal Distributions

The proposition can be proved by writing the cdf of
$Z = (X - \mu)/\sigma$ as

$$P(Z \le z) = P(X \le \sigma z + \mu) = \int_{-\infty}^{\sigma z+\mu} f(x; \mu, \sigma)dx$$

Using a result from calculus, this integral can be differentiated with respect to z to yield the desired pdf $f(z; 0, 1)$.

Example 16



ระยะตอบสนองเป็น normal

The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions. ลดค.เร็ว

The article “Fast-Rise Brake Lamp as a Collision-Prevention Device” (*Ergonomics*, 1993: 391–395) suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 sec and standard deviation of .46 sec.

μ σ

Example 16

cont'd

What is the probability that reaction time is between 1.00 sec and 1.75 sec? If we let *X* denote reaction time, then standardizing gives

$$1.00 \leq X \leq 1.75$$

if and only if

$$\left(\frac{1.00 - 1.25}{.46} \leq \frac{X - 1.25}{.46} \leq \frac{1.75 - 1.25}{.46}\right)$$

Thus

$$P(1.00 \leq X \leq 1.75) = P\left(\underbrace{\frac{1.00 - 1.25}{.46}}_{-0.59} \leq Z \leq \underbrace{\frac{1.75 - 1.25}{.46}}_{1.09}\right)$$

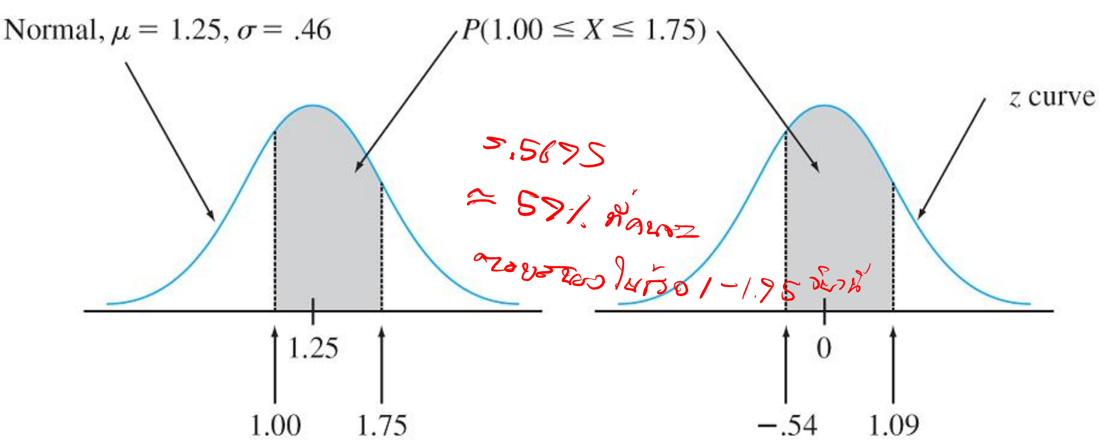
Example 16

cont'd

$$= P(-.54 \leq Z \leq 1.09) = \Phi(1.09) - \Phi(-.54)$$

$$= .8621 - .2946 = .5675$$

This is illustrated in Figure 4.22



0.5675

≈ 57% ที่คนจะ

ตอบสนองในช่วง 1–1.75 วินาที

Normal curves for Example 16

Figure 4.22

Example 16

cont'd

Similarly, if we view 2 sec as a critically long reaction time, the probability that actual reaction time will exceed this value is

$$P(X > 2) = P\left(Z > \frac{2 - 1.25}{.46}\right) = P(Z > 1.63) = 1 - \Phi(1.63) = .0516$$

0.9484

$\simeq 5\%$

มีโอกาส[illegible] 5%

Percentiles of an Arbitrary Normal Distribution

Percentiles of an Arbitrary Normal Distribution

หา.ไอ้กราฟ ร.ค.น่าจะเป็น

The (100*p*)th percentile of a normal distribution with mean μ and standard deviation σ is easily related to the (100*p*)th percentile of the standard normal distribution.

Proposition

$$\overset{x}{\begin{matrix}(100p)\text{th percentile} \\ \text{for normal } (\mu, \sigma)\end{matrix}} = \mu + \overset{z}{\begin{bmatrix}(100p)\text{th for} \\ \text{standard normal}\end{bmatrix}} \cdot \sigma$$

Another way of saying this is that if *z* is the desired percentile for the standard normal distribution, then the desired percentile for the normal (μ, σ) distribution is *z* standard deviations from μ.

Example 18

น้ำกลั่น บรรจุ ขวด

The amount of distilled water dispensed by a certain machine is normally distributed with mean value 64 oz and standard deviation .78 oz.

μ σ

น้ำล้นขวด

What container size *c* will ensure that overflow occurs only .5% of the time? If *X* denotes the amount dispensed, the desired condition is that $P(X > c) = .005$, or, equivalently, that $\boxed{P(X \leq c) = .995.}$

$z = 2.58$

ขนาดของขวด

Thus *c* is the 99.5th percentile of the normal distribution with $\mu = 64$ and $\sigma = .78$.

Example 18

cont'd

The 99.5th percentile of the standard normal distribution is 2.58, so

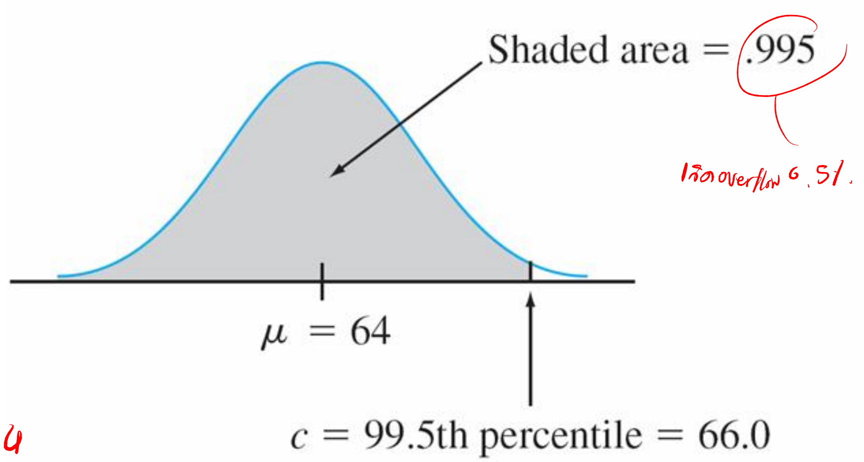
$$c = \eta(.995) = 64 + (2.58)(.78) = 64 + 2.0 = 66 \text{ oz}$$

This is illustrated in Figure 4.23.

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

$2.58 = \frac{x - 64}{0.78}$

$x = 66.0124$



เกิด overflow 0.5%

Distribution of amount dispensed for Example 18

Figure 4.23

The Normal Distribution and Discrete Populations

The Normal Distribution and Discrete Populations

The normal distribution is often used as an approximation to the distribution of values in a discrete population.

In such situations, extra care should be taken to ensure that probabilities are computed in an accurate manner.

Example 19

IQ in a particular population (as measured by a standard test) is known to be approximately normally distributed with $\mu = 100$ and $\sigma = 15$.

What is the probability that a randomly selected individual has an IQ of at least 125?

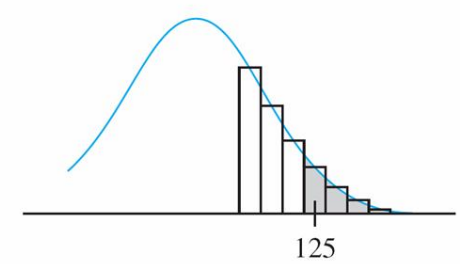
Letting X = the IQ of a randomly chosen person, we wish $P(X \geq 125)$.

The temptation here is to standardize $X \geq 125$ as in previous examples. However, the IQ population distribution is actually discrete, since IQs are integer-valued.

Example 19

cont'd

So the normal curve is an approximation to a discrete probability histogram, as pictured in Figure 4.24.



A normal approximation to a discrete distribution

Figure 4.24

Discrete

The rectangles of the histogram are *centered* at integers, so IQs of at least 125 correspond to rectangles beginning at 124.5, as shaded in Figure 4.24.

Example 19

cont'd

Thus we really want the area under the approximating normal curve to the right of 124.5.

$\frac{124.5-100}{15} = Z$

Standardizing this value gives $P(Z \geq 1.63) = .0516$, whereas standardizing 125 results in $P(Z \geq 1.67) = .0475$.

$\frac{25}{15}$ $\frac{5}{3}$

The difference is not great, but the answer .0516 is more accurate. Similarly, $P(X = 125)$ would be approximated by the area between 124.5 and 125.5, since the area under the normal curve above the single value 125 is zero.

Example 19

cont'd

The correction for discreteness of the underlying distribution in Example 19 is often called a **continuity correction.**

It is useful in the following application of the normal distribution to the computation of binomial probabilities.

Approximating the Binomial Distribution

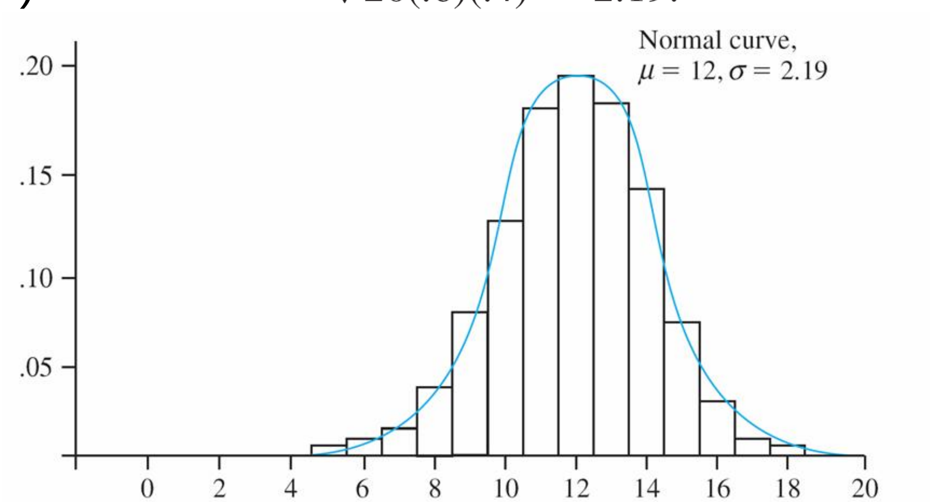
Approximating the Binomial Distribution

Recall that the mean value and standard deviation of a binomial random variable *X* are $\mu_X = np$ and $\sigma_X = \sqrt{npq}$, respectively.

Approximating the Binomial Distribution

จำนวนครั้งการทดลอง

Figure 4.25 displays a binomial probability histogram for the binomial distribution with $n = 20$, $p = .6$, for which $\mu = 20(.6) = 12$ and $\sigma = \sqrt{20(.6)(.4)} = 2.19$.



Binomial probability histogram for $n = 20$, $p = .6$ with normal approximation curve superimposed

Figure 4.25

Approximating the Binomial Distribution

A normal curve with this μ and σ has been superimposed on the probability histogram.

mean = median

$\mu = \tilde{\mu}$

$p = 0.5$

Although the probability histogram is a bit skewed (because $p \neq .5$), the normal curve gives a very good approximation, especially in the middle part of the picture.

The area of any rectangle (probability of any particular X value) except those in the extreme tails can be accurately approximated by the corresponding normal curve area.

Approximating the Binomial Distribution

For example, [illegible]

$P(X = 10) = B(10; 20, .6) – B(9; 20, .6) = .117$,

whereas the area under the normal curve between 9.5 and 10.5 is $P(–1.14 \leq Z \leq –.68) = .1212$.

More generally, as long as the binomial probability histogram is not too skewed, binomial probabilities can be well approximated by normal curve areas.

It is then customary to say that X has approximately a normal distribution.

Approximating the Binomial Distribution

Proposition

Let X be a binomial rv based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$.

In particular, for x = a possible value of X,

$$P(X \leq x) = B(x, n, p) \approx \boxed{\begin{pmatrix} \text{area under the normal curve} \\ \text{to the left of } x + .5 \end{pmatrix}}$$

ปกติ

$$= \Phi\left(\frac{x + .5 - np}{\sqrt{npq}}\right)$$

Approximating the Binomial Distribution

n ต้องเยอะ ถึงจะดี

In practice, the approximation is adequate provided that both $np \geq 10$ and $n\overset{q}{\cancel{p}} \geq 10$, since there is then enough symmetry in the underlying binomial distribution.

A direct proof of this result is quite difficult. In the next chapter we'll see that it is a consequence of a more general result called the Central Limit Theorem. CLT

In all honesty, this approximation is not so important for probability calculation as it once was.

This is because software can now calculate binomial probabilities exactly for quite large values of n.

Example 20

p=0.25

Suppose that 25% of all students at a large public university receive financial aid.

n

Let X be the number of students in a random sample of size 50 who receive financial aid, so that $p = .25$.
Then $\mu = 12.5$ and $\sigma = 3.06$.

np $\sqrt{npq}$

Since $np = 50(.25) = 12.5 \geq 10$ and $np = 37.5 \geq 10$, the approximation can safely be applied.

Example 20

cont'd

The probability that at most 10 students receive aid is

x n p

$$P(X \leq 10) = B(10; 50, .25) \approx \Phi\left(\frac{10 + .5 - 12.5}{3.06}\right)$$

$$= \Phi(-.65) = .2578$$

Similarly, the probability that between 5 and 15 (inclusive) of the selected students receive aid is

$$P(5 \leq X \leq 15) = B(15; 50, .25) - B(4; 50, .25)$$

+0.5 +0.5

$$\approx \Phi\left(\frac{15.5 - 12.5}{3.06}\right) - \Phi\left(\frac{4.5 - 12.5}{3.06}\right) = .8320$$

Example 20

cont'd

The exact probabilities are .2622 and .8348, respectively, so the approximations are quite good.

In the last calculation, the probability $P(5 \leq X \leq 15)$ is being approximated by the area under the normal curve between 4.5 and 15.5—the continuity correction is used for both the upper and lower limits.