

Chapter 5 Random Variables

A **random variable** is a variable that takes on numerical outcomes defined over a sample space of a random experiment.

A random variable has a probability distribution.

A random variable can be denoted by X (upper-case) and a possible numerical outcome is x (lower-case).

Example: The random variable X is age in years of a UBC student. The possible outcomes are:

$$x = 15, 16, 17, 18, \dots \text{ etc.}$$

Types of random variables:

- A **discrete random variable** has a countable number of values (typically integer numbers).

Example 1: age in years of a UBC student

Example 2: categorical variables.

For example, the random variable X represents gender.

The possible values can be assigned the codes:

$$x = 0 \quad \text{male}$$

$$x = 1 \quad \text{female}$$

- A **continuous random variable** can take any numerical value in an interval of the real number line.

Examples: income, stock market prices, interest rates, consumer price index, etc.

Chapter 5.2 Discrete Random Variables

For a discrete random variable X the **probability distribution function** is:

$$P(x) = P(X=x) \quad \text{for all possible values of } x.$$

Example: The random variable X is the number resulting from the throw of a six-sided dice.

The probability distribution function is:

x	1	2	3	4	5	6
$P(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

That is, $P(x) = \frac{1}{6}$ for $x = 1, 2, 3, 4, 5, 6$

The probability distribution function for a discrete random variable has the properties:

- $0 \leq P(x) \leq 1$ for all possible values of x .
- $\sum_x P(x) = 1$
↑
summation over all possible values of x .

The **cumulative probability function** is defined as:

$$F(a) = P(X \leq a) \quad \text{for all possible values of } a.$$

This can be calculated from the probability distribution function as:

$$F(a) = \sum_{x \leq a} P(x)$$

↑

summation over all possible values of x that are less than or equal to a .

Example: For the dice throwing experiment:

$$F(1) = P(X \leq 1) = P(X = 1) = \frac{1}{6}$$

$$\begin{aligned} F(2) &= P(X \leq 2) = P(X = 1) + P(X = 2) \\ &= \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \end{aligned}$$

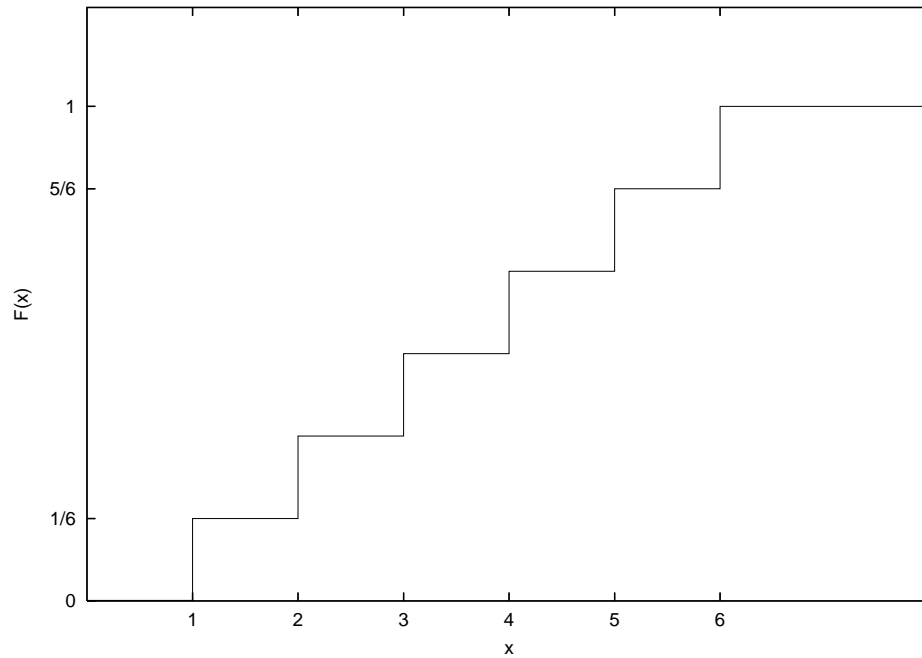
$$\begin{aligned} F(3) &= P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) \\ &= \frac{1}{2} \end{aligned}$$

$$F(4) = P(X \leq 4) = \frac{2}{3}$$

$$F(5) = P(X \leq 5) = \frac{5}{6}$$

$$F(6) = P(X \leq 6) = 1$$

Graph of the cumulative probability function for the dice throwing experiment.



The graph illustrates that, for a discrete random variable, the cumulative probability function is a step function that begins at 0 and ends at 1.

The cumulative probability function for a discrete random variable has the properties:

- $0 \leq F(\mathbf{a}) \leq 1$ for all possible values of \mathbf{a} .
- For two numbers \mathbf{a} , \mathbf{b} with $\mathbf{a} < \mathbf{b}$ then

$$F(\mathbf{a}) \leq F(\mathbf{b})$$

- $P(\mathbf{X} > \mathbf{a}) = 1 - P(\mathbf{X} \leq \mathbf{a})$
 $= 1 - F(\mathbf{a})$

Example: Exercise 5.14, page 140.

The random variable \mathbf{X} is the number of flights delayed per hour at an international airport.

The probability distribution function and cumulative probability function are:

\mathbf{x}	$\mathbf{P(x)}$	$\mathbf{F(x)}$
0	0.10	0.10
1	0.08	0.18
2	0.07	0.25
3	0.15	0.40
4	0.12	0.52
5	0.08	0.60
6	0.10	0.70
7	0.12	0.82
8	0.08	0.90
9	0.10	1.00

What is the probability of five or more delayed flights in a given hour ?

$$\begin{aligned}P(\mathbf{X} \geq 5) &= 1 - P(\mathbf{X} \leq 4) \\ &= 1 - F(4) \\ &= 1 - 0.52 \\ &= 0.48\end{aligned}$$

What is the probability of three through seven (inclusive) delayed flights in a given hour ?

$$\begin{aligned}P(3 \leq \mathbf{X} \leq 7) &= P(\mathbf{X} \leq 7) - P(\mathbf{X} \leq 2) \\ &= F(7) - F(2) \\ &= 0.82 - 0.25 \\ &= 0.57\end{aligned}$$

Chapter 5.3 Mean and Variance of Discrete Random Variables

Summary measures of the information in the probability distribution are of interest. Recall, the mean is the measure of central location for a data set of numeric observations. For a random variable, the **expected value** is the corresponding measure of central location.

For a discrete random variable X , the expected value is defined as:

$$E(X) = \sum_x xP(x)$$

This is a weighted average of all possible outcomes where the weights are the probabilities.

The expected value of a random variable is also called its **mean** and is denoted by:

$$\mu_X = E(X)$$

↑

Greek letter mu

Example: The random variable \mathbf{X} is the sum of the two numbers shown on a throw of two dice. The probability distribution function of \mathbf{X} is:

x	2	3	4	5	6	7
$\mathbf{P(x)}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$
	8	9	10	11	12	
	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

The expected value is calculated as:

$$\begin{aligned}
 \mathbf{E(X)} &= 2\left(\frac{1}{36}\right) + 3\left(\frac{2}{36}\right) + 4\left(\frac{3}{36}\right) + \dots + 11\left(\frac{2}{36}\right) + 12\left(\frac{1}{36}\right) \\
 &= 7
 \end{aligned}$$

The expected value can be viewed as the long-run average value that a random variable would take over a “large” number of trials of the random experiment.

This can be illustrated for the dice-toss experiment.

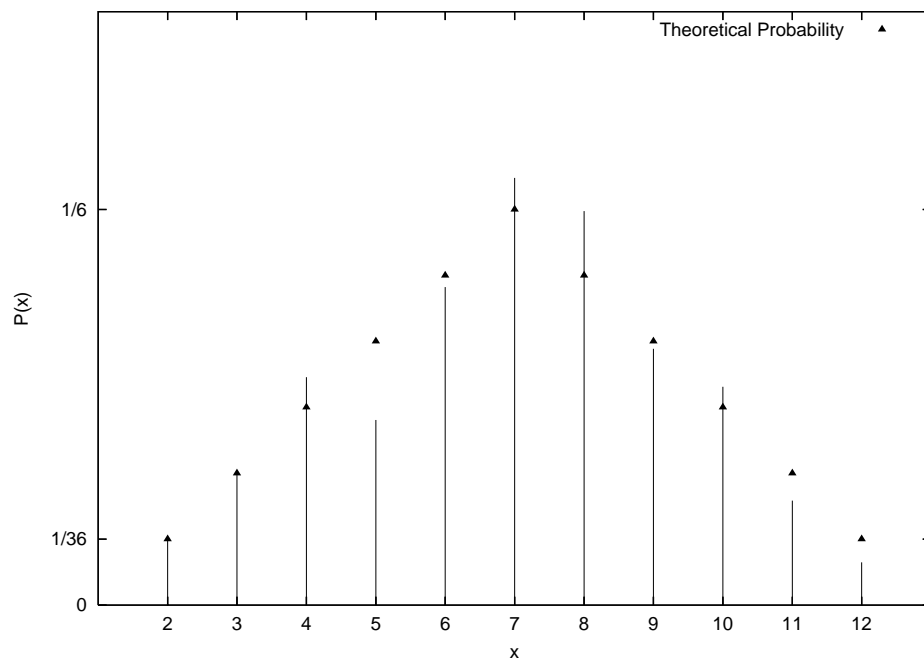
A computer program was used to simulate the throw of two dice.

At each throw, the sum of the two dice faces was recorded.

This was repeated 500 times.

The graph shows the relative frequencies of each outcome.

This gives the empirical probability distribution function.



The average of the 500 observed outcomes was 6.982, close to the proposed expected value of 7.0.

Suppose $g(\mathbf{X})$ is some function of the random variable \mathbf{X} . Then:

$$\mathbf{E}[g(\mathbf{X})] = \sum_{\mathbf{x}} g(\mathbf{x}) P(\mathbf{x})$$

Example: For $g(\mathbf{X}) = \mathbf{X}^2$ the expectation is calculated as:

$$\mathbf{E}(\mathbf{X}^2) = \sum_{\mathbf{x}} \mathbf{x}^2 P(\mathbf{x})$$

A measure of dispersion for a random variable \mathbf{X} is the **variance** defined as:

$$\mathbf{Var}(\mathbf{X}) = \mathbf{E}[(\mathbf{X} - \mu_{\mathbf{X}})^2]$$

where $\mu_{\mathbf{X}} = \mathbf{E}(\mathbf{X})$

$$= \sum_{\mathbf{x}} (\mathbf{x} - \mu_{\mathbf{X}})^2 P(\mathbf{x})$$

The variance of a random variable \mathbf{X} is denoted by the symbol $\sigma_{\mathbf{X}}^2$ (sigma-squared):

$$\sigma_{\mathbf{X}}^2 = \mathbf{Var}(\mathbf{X})$$

The variance can be expressed in an alternative way:

$$\begin{aligned}\text{Var}(X) &= \sum_x (x - \mu_X)^2 P(x) \\ &= \sum_x (x^2 - 2\mu_X x + \mu_X^2) P(x) \\ &= \sum_x x^2 P(x) - 2\mu_X \sum_x x P(x) + \mu_X^2 \sum_x P(x) \\ &= \sum_x x^2 P(x) - 2\mu_X \cdot \mu_X + \mu_X^2 \quad (1) \\ &= \sum_x x^2 P(x) - \mu_X^2 \\ &= E(X^2) - \mu_X^2\end{aligned}$$

The **standard deviation** of a random variable X is defined as:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\text{Var}(X)} > 0$$

Note: A positive variance, and, therefore, positive standard deviation, assumes at least two distinct outcomes.

Consider two random variables X and Y with means:

$$\mu_X = E(X), \quad \mu_Y = E(Y)$$

and variances:

$$\sigma_X^2 = \text{Var}(X), \quad \sigma_Y^2 = \text{Var}(Y)$$

The two random variables may have the same mean but substantial differences in the variance. Suppose that:

$$\mu_X = \mu_Y \quad \text{and}$$

$$\sigma_Y^2 > \sigma_X^2$$

This suggests that outcomes different from the mean are more likely for random variable Y than for random variable X .

❖ Useful Results for Expected Value and Variance

Let \mathbf{a} and \mathbf{b} be any constant fixed numbers.

- $\mathbf{E(a) = a}$
- $\mathbf{E(a + b X) = a + b E(X) = a + b \mu_X}$

This result can be shown:

$$\begin{aligned}\mathbf{E(a + b X)} &= \sum_x (\mathbf{a + b x})P(x) \\ &= \mathbf{a} \sum_x P(x) + \mathbf{b} \sum_x xP(x) \\ &= \mathbf{a + b E(X)}\end{aligned}$$

- $\mathbf{Var(a) = 0}$
- $\mathbf{Var(a + b X) = b^2 Var(X)}$

This result can be shown:

$$\begin{aligned}\mathbf{Var(a + b X)} &= \mathbf{E\{[a + b X - E(a + b X)]^2\}} \\ &= \mathbf{E\{[a + b X - (a + b \mu_X)]^2\}} \\ &= \mathbf{E\{(b X - b \mu_X)^2\}} \\ &= \mathbf{E[b^2 (X - \mu_X)^2\}} \\ &= \mathbf{b^2 E[(X - \mu_X)^2\}} \\ &= \mathbf{b^2 Var(X)}\end{aligned}$$

Example: Exercise 5.20, page 148.

A production process gives variation for the number of paper clips per package.

Let the random variable X be the number of paper clips in a package. The probability distribution function and cumulative probability function are given as:

x	$P(x)$	$F(x)$
47	0.04	0.04
48	0.13	0.17
49	0.21	0.38
50	0.29	0.67
51	0.20	0.87
52	0.10	0.97
53	0.03	1.00

Selected questions and answers.

(d) Two packages are chosen at random. Find the probability that at least one of them contains at least 50 paper clips.

With the assumption of independence the answer is obtained as:

$$\begin{aligned}1 - P(\text{neither contains 50 or more}) &= 1 - P(X \leq 49)^2 \\ &= 1 - F(49)^2 \\ &= 1 - 0.38^2 \\ &= 0.8556\end{aligned}$$

(e) Find the mean and standard deviation of the number paper clips per package.

The mean is calculated as:

$$\begin{aligned} E(X) &= \sum_x xP(x) \\ &= (47)(.04) + (48)(.13) + (49)(.21) + (50)(.29) + \\ &\quad (51)(.20) + (52)(.10) + (53)(.03) \\ &= 49.9 \end{aligned}$$

To calculate the variance, first calculate:

$$\begin{aligned} E(X^2) &= \sum_x x^2 P(x) \\ &= 47^2(.04) + 48^2(.13) + 49^2(.21) + 50^2(.29) + \\ &\quad 51^2(.20) + 52^2(.10) + 53^2(.03) \\ &= 2491.96 \end{aligned}$$

The variance is found as:

$$\begin{aligned} \text{Var}(X) &= E(X^2) - \mu_X^2 \\ &= 2491.96 - 49.9^2 \\ &= 1.95 \end{aligned}$$

The standard deviation is calculated as:

$$\sigma_X = \sqrt{1.95} = 1.396$$

(f) The cost (in cents) of producing a package of paper clips is the random variable:

$$C = 16 + 2X$$

The price of a package of paper clips is \$1.50. Therefore, profit (in cents) per package is the random variable:

$$\begin{aligned} P &= 150 - C \\ &= 134 - 2X \end{aligned}$$

Find the mean and standard deviation of profit per package.

The mean is:

$$\begin{aligned} E(P) &= E(134 - 2X) \\ &= 134 - 2E(X) \\ &= 134 - (2)(49.9) \\ &= 34.2 \text{ cents} \end{aligned}$$

The variance is:

$$\begin{aligned} \sigma_P^2 &= \text{Var}(P) = \text{Var}(134 - 2X) \\ &= (-2)^2 \text{Var}(X) \\ &= 4\sigma_X^2 \end{aligned}$$

This gives the standard deviation:

$$\sigma_P = \sqrt{4\sigma_X^2} = 2\sigma_X = (2)(1.396) = 2.79 \text{ cents}$$

Chapter 5.4 Binomial Distribution

A special application of a discrete probability distribution is the binomial distribution.

To start, introduce the random variable X_B that takes two outcomes:

$$x = 1 \quad \text{“success”}$$

$$x = 0 \quad \text{“failure”}$$

The probability distribution function of X_B is:

$$P(X_B = 1) = p \quad \text{for } 0 < p < 1 \quad (\text{the probability of success})$$

$$P(X_B = 0) = 1 - p$$

This is known as the Bernoulli distribution.

The mean and variance are calculated as:

$$\begin{aligned} \mu_{X_B} = E(X_B) &= \sum_x x P(x) \\ &= 1 \cdot p + 0 \cdot (1 - p) \\ &= p \end{aligned}$$

$$\begin{aligned} \text{Var}(X_B) &= E(X_B^2) - \mu_{X_B}^2 \\ &= \sum_x x^2 P(x) - p^2 \\ &= (1)(1) \cdot p + (0)(0) \cdot (1 - p) - p^2 \\ &= p - p^2 \\ &= p(1 - p) \end{aligned}$$

Now consider that a random experiment with the outcome of success or failure is repeated n times.

Each trial produces success or failure with probabilities p and $(1-p)$ respectively. Assume independence so that the result of one trial does not influence the result of any other trial.

Let the random variable X be the number of successes in n trials.

The probability distribution function of X is defined as:

$$P(x) = P(x \text{ successes in } n \text{ independent trials})$$

$$\text{for } x = 0, 1, 2, \dots, n$$

This is known as the **binomial distribution**.

A calculation formula for the probabilities can be obtained as follows.

In n independent trials, the probability of x successes and $(n-x)$ failures is:

$$p^x (1 - p)^{n-x}$$

The number of combinations of x successes in n trials is:

$$C_x^n = \frac{n!}{x!(n-x)!}$$

Therefore, the probability distribution function for the binomial distribution is:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n$$

Note: a calculation rule is $0! = 1$

Examples of application of the binomial distribution are:

$$\begin{aligned} P(X = 0) &= P(\text{no successes in } n \text{ trials}) \\ &= (1 - p)^n \end{aligned}$$

$$\begin{aligned} P(X = 1) &= P(\text{one success in } n \text{ trials}) \\ &= n p (1 - p)^{n-1} \end{aligned}$$

The cumulative probability function of the binomial distribution may have useful application and is calculated as:

$$F(x) = P(X \leq x) \quad \text{for } x = 0, 1, 2, \dots, n$$

For example,

$$\begin{aligned} F(1) = P(X \leq 1) &= P(\text{at most one success in } n \text{ trials}) \\ &= P(X = 0) + P(X = 1) \end{aligned}$$

Example: Exercise 5.39, page 156.

A company installs new heating furnaces. For any installation, the probability of a return visit for a repair is 0.15.

Six installations are made in a given week.

Assume independence of outcomes for these installations.

Let the random variable X be the number of return visits.

X follows a binomial distribution with:

$$p = 0.15 \quad \text{and} \quad 1 - p = 0.85$$

Find the probability that a return visit will be needed in more than one of the installations.

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - [(0.85)^6 + (6)(0.15)(0.85)^5] \end{aligned}$$

At this point, it is clear that the numerical calculations can be tedious.

Numerical answers for binomial distribution probabilities can be obtained with Microsoft Excel. Select Insert Function BINOMDIST.

The general usage is: $\text{BINOMDIST}(x, n, p, \text{cumulative})$

where $\text{cumulative} = 0$ for the probability distribution function,
 $\text{cumulative} = 1$ for the cumulative probability function

For this exercise the calculation of $\mathbf{P(X \leq 1)}$ was found with

$$\text{BINOMDIST}(1, 6, 0.15, 1) = 0.7765$$

Therefore, the answer is:

$$\mathbf{P(X > 1) = 1 - P(X \leq 1) = 1 - 0.7765 = 0.2235}$$

Note: an important assumption of the binomial distribution is independent trials.

Chapter 5.7 Jointly Distributed Random Variables

Economic relationships between variables are of interest.

Let \mathbf{X} and \mathbf{Y} be a pair of discrete random variables such that:

\mathbf{X} has numerical outcomes x , and

\mathbf{Y} has numerical outcomes y .

The **joint probability function** is:

$$\mathbf{P}_{\mathbf{X},\mathbf{Y}}(x, y) = \mathbf{P}(\mathbf{X} = x \text{ and } \mathbf{Y} = y) \quad \text{for all pairs } (x, y)$$

A joint probability function has the properties:

- $0 \leq \mathbf{P}_{\mathbf{X},\mathbf{Y}}(x, y) \leq 1$ for all pairs (x, y)
- $\sum_x \sum_y \mathbf{P}_{\mathbf{X},\mathbf{Y}}(x, y) = 1$

The probability function of \mathbf{X} is obtained by summing the joint probabilities:

$$\mathbf{P}_{\mathbf{X}}(\mathbf{x}) = \sum_y \mathbf{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \quad \text{for all possible values of } \mathbf{x}.$$

↑
summation over all possible values of \mathbf{y} .

This is called the **marginal probability function** of \mathbf{X} .

Similarly, the marginal probability function of \mathbf{Y} is constructed as:

$$\mathbf{P}_{\mathbf{Y}}(\mathbf{y}) = \sum_x \mathbf{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \quad \text{for all possible values of } \mathbf{y}.$$

The **conditional probability function** of \mathbf{Y} given that $\mathbf{X} = \mathbf{x}$ is:

$$\mathbf{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}) = \frac{\mathbf{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{\mathbf{P}_{\mathbf{X}}(\mathbf{x})} \quad \text{for all possible values of } \mathbf{y}.$$

Similarly, the conditional probability function of \mathbf{X} given that $\mathbf{Y} = \mathbf{y}$ is:

$$\mathbf{P}_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} | \mathbf{y}) = \frac{\mathbf{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{\mathbf{P}_{\mathbf{Y}}(\mathbf{y})} \quad \text{for all possible values of } \mathbf{x}.$$

For the conditional probability function:

$$\sum_y \mathbf{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}) = \mathbf{1} \quad \text{and} \quad \sum_x \mathbf{P}_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} | \mathbf{y}) = \mathbf{1}$$

The random variables \mathbf{X} and \mathbf{Y} are **independent** if and only if:

$$\mathbf{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = \mathbf{P}_{\mathbf{X}}(\mathbf{x})\mathbf{P}_{\mathbf{Y}}(\mathbf{y}) \quad \text{for all pairs } (\mathbf{x}, \mathbf{y})$$

If random variables \mathbf{X} and \mathbf{Y} are independent then:

$$\begin{aligned} \mathbf{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) &= \frac{\mathbf{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{\mathbf{P}_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{\mathbf{P}_{\mathbf{X}}(\mathbf{x})\mathbf{P}_{\mathbf{Y}}(\mathbf{y})}{\mathbf{P}_{\mathbf{X}}(\mathbf{x})} \\ &= \mathbf{P}_{\mathbf{Y}}(\mathbf{y}) \end{aligned}$$

That is, the conditional probability function of \mathbf{Y} , given that the random variable \mathbf{X} takes the value x , is identical to the marginal probability function of \mathbf{Y} , for all possible values of y .

Example: Adapted from Exercise 5.83, page 180

A survey by a real estate agent has collected information on apartment rentals. Consider the discrete random variables:

X volume of inquiries by renters. The possible values are:

$x = 0$ little interest

$x = 1$ moderate interest

$x = 2$ strong interest

Y number of lines in a newspaper ad. Possible values are:

$y = 3, 4, 5$

The joint probability function is:

Y	X		
	0	1	2
3	0.09	0.14	0.07
4	0.07	0.23	0.16
5	0.03	0.10	0.11

Questions and Answers

- Find the probability function of X .

$$P_X(0) = 0.09 + 0.07 + 0.03 = 0.19$$

$$P_X(1) = 0.14 + 0.23 + 0.10 = 0.47$$

$$P_X(2) = 0.07 + 0.16 + 0.11 = 0.34$$

Note: the probabilities sum to one.

- Find the mean of X .

$$\mu_X = E(X) = \sum_x x P_X(x)$$

$$= (0)(0.19) + (1)(0.47) + (2)(0.34) = 1.15$$

- For the random variable Y , find the probability function and mean.

$$P_Y(3) = 0.30, P_Y(4) = 0.46, \text{ and } P_Y(5) = 0.24$$

$$\mu_Y = E(Y) = \sum_y y P_Y(y)$$

$$= (3)(0.30) + (4)(0.46) + (5)(0.24) = 3.94$$

- Find the conditional probability function for Y given $X=0$.

$$P_{Y|X}(3|0) = \frac{P_{X,Y}(0,3)}{P_X(0)} = \frac{0.09}{0.19} = 0.4737$$

$$P_{Y|X}(4|0) = \frac{P_{X,Y}(0,4)}{P_X(0)} = \frac{0.07}{0.19} = 0.3684$$

$$P_{Y|X}(5|0) = \frac{P_{X,Y}(0,5)}{P_X(0)} = \frac{0.03}{0.19} = 0.1579$$

Note: the probabilities sum to one.

- Are X and Y independent?

Recall that independence requires:

$$P_{X,Y}(x,y) = P_X(x)P_Y(y) \quad \text{for all pairs } (x,y)$$

For the values $X=0$ and $Y=3$, the joint probability is:

$$P_{X,Y}(0,3) = 0.09$$

The product of the marginal probabilities is:

$$P_X(0)P_Y(3) = (0.19)(0.30) = 0.057$$

It is clear that $P_{X,Y}(0,3) \neq P_X(0)P_Y(3)$

Therefore, the two random variables are not independent.

Let $g(X, Y)$ be a function of the discrete random variables X and Y . The **expected value** of this function is defined as:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) P_{X, Y}(x, y)$$

A property of expectation is:

$$E(X + Y) = E(X) + E(Y)$$

This result can be shown:

$$\begin{aligned} E(X + Y) &= \sum_x \sum_y (x + y) P_{X, Y}(x, y) \\ &= \sum_x \sum_y [x P_{X, Y}(x, y) + y P_{X, Y}(x, y)] \\ &= \sum_x x \sum_y P_{X, Y}(x, y) + \sum_y y \sum_x P_{X, Y}(x, y) \\ &= \sum_x x P_X(x) + \sum_y y P_Y(y) \\ &= E(X) + E(Y) \end{aligned}$$

For constant fixed numbers a and b a rule is:

$$E(aX + bY) = aE(X) + bE(Y)$$

A general result is that for K random variables X_1, X_2, \dots, X_K with means $\mu_1, \mu_2, \dots, \mu_K$ the expected value of their sum is:

$$E(X_1 + X_2 + \dots + X_K) = \mu_1 + \mu_2 + \dots + \mu_K$$

A measure of a linear relationship between two random variables is of interest.

For random variables X and Y with means μ_X and μ_Y the **covariance** between X and Y is defined as:

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) P_{X,Y}(x, y)\end{aligned}$$

An equivalent expression can be stated:

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[(XY - X\mu_Y - \mu_X Y + \mu_X \mu_Y)] \\ &= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y\end{aligned}$$

where

$$E(XY) = \sum_x \sum_y xy P_{X,Y}(x, y)$$

If the random variables X and Y are independent then:

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy P_{X,Y}(x,y) \\ &= \sum_x \sum_y xy P_X(x) P_Y(y) \\ &= \left[\sum_x x P_X(x) \right] \left[\sum_y y P_Y(y) \right] \\ &= \mu_X \mu_Y \end{aligned}$$

It follows that independence gives:

$$\mathbf{Cov(X, Y) = E(XY) - \mu_X \mu_Y = 0}$$

- However, if zero covariance is established, this does **not** guarantee that the random variables are independent. Covariance is designed to measure the possibility of a linear relationship. Nonlinear relationships between variables may give dependencies even though the covariance is zero.

Also note that, in general, for random variables with non-zero covariance:

$$\mathbf{E(XY) \neq E(X)E(Y)}$$

Covariance gives an indication of the sign (positive or negative) of a linear relationship between random variables.

A measure of the strength of a linear relationship between random variables X and Y is the **correlation** defined as:

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

↑

Greek letter rho

where σ_X and σ_Y are the standard deviations of the random variables.

A result is: $-1 \leq \rho \leq 1$

A value of $\rho = 0$ indicates that the random variables are **uncorrelated**.

Problem: If two random variables are uncorrelated, are they independent ?

Example: the real estate agent exercise Continued.

Earlier in the lecture notes, an exercise introduced the joint probability function:

Y	X		
	0	1	2
3	0.09	0.14	0.07
4	0.07	0.23	0.16
5	0.03	0.10	0.11

To find the covariance between the random variables X and Y first calculate:

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy P_{X,Y}(x,y) \\ &= (0)(3)(0.09) + (0)(4)(0.07) + (0)(5)(0.03) + \\ &\quad (1)(3)(0.14) + (1)(4)(0.23) + (1)(5)(0.10) + \\ &\quad (2)(3)(0.07) + (2)(4)(0.16) + (2)(5)(0.11) \\ &= 4.64 \end{aligned}$$

The covariance is:

$$\begin{aligned} \text{Cov}(X, Y) &= E(XY) - \mu_X \mu_Y \\ &= 4.64 - (1.15)(3.94) \\ &= 0.109 \end{aligned}$$

The variances of the two random variables are calculated as:

$$\begin{aligned}\sigma_X^2 &= E(X^2) - \mu_X^2 \\ &= (0)(0.19) + (1)(0.47) + (4)(0.34) - (1.15)(1.15) \\ &= 0.5075\end{aligned}$$

$$\begin{aligned}\sigma_Y^2 &= E(Y^2) - \mu_Y^2 \\ &= (9)(0.30) + (16)(0.46) + (25)(0.24) - (3.94)(3.94) \\ &= 0.5364\end{aligned}$$

The correlation between X and Y is:

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{0.109}{\sqrt{(0.5075)(0.5364)}} = 0.2089$$

That is, there is a positive correlation between the number of lines in a newspaper ad and the volume of inquiries about the apartment rental.

❖ Useful Results for the Variance of a Linear Combination of Random Variables

For two random variables X and Y a result is:

$$\mathbf{Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)}$$

This result can be shown:

$$\begin{aligned}\mathbf{Var(X + Y)} &= \mathbf{E[\{X + Y - E(X + Y)\}^2]} \\ &= \mathbf{E[\{(X - E(X)) + (Y - E(Y))\}^2]} \\ &= \mathbf{E[(X - E(X))^2 + (Y - E(Y))^2 + 2(X - E(X))(Y - E(Y))]} \\ &= \mathbf{E[(X - E(X))^2] + E[(Y - E(Y))^2] + 2E[(X - E(X))(Y - E(Y))]} \\ &= \mathbf{Var(X) + Var(Y) + 2Cov(X, Y)}\end{aligned}$$

Another result is:

$$\mathbf{Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)}$$

When X and Y are independent then the covariance is zero and:

$$\mathbf{Var(X + Y) = Var(X - Y) = Var(X) + Var(Y)}$$

For constant fixed numbers a and b , a general rule is:

$$\mathbf{Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2a b Cov(X, Y)}$$

Example: Portfolio Analysis, Example 5.18 page 177.

The job of a financial advisor may be to recommend a mix of stocks or a portfolio for investment purposes.

Consider the random variables:

X_1 price of one share of stock for company A

X_2 price of one share of stock for company B

X_3 price of one share of stock for company C

Means and variances are:

$$E(X_1) = \$ 53 \quad \text{Var}(X_1) = 31.3$$

$$E(X_2) = \$ 55 \quad \text{Var}(X_2) = 125$$

$$E(X_3) = \$ 55 \quad \text{Var}(X_3) = 125$$

and covariances are:

$$\text{Cov}(X_1, X_2) = 59.17 \quad \text{and} \quad \text{Cov}(X_1, X_3) = -59.17$$

Two alternative portfolios are represented by the random variables:

$$W_1 = 5X_1 + 10X_2 \quad \text{and}$$

$$W_2 = 5X_1 + 10X_3$$

The mean of the first portfolio is calculated as:

$$\begin{aligned} E(W_1) &= E(5X_1 + 10X_2) \\ &= 5E(X_1) + 10E(X_2) \\ &= (5)(53) + (10)(55) \\ &= \$815 \end{aligned}$$

The portfolio W_2 has the identical mean as the portfolio W_1 .

The performance of the two portfolios can be compared by comparing their variances.

$$\begin{aligned} \text{Var}(W_1) &= \text{Var}(5X_1 + 10X_2) \\ &= 5^2 \text{Var}(X_1) + 10^2 \text{Var}(X_2) + 2(5)(10) \cdot \text{Cov}(X_1, X_2) \\ &= (25)(31.3) + (100)(125) + (100)(59.17) \\ &= 19,199.5 \end{aligned}$$

In contrast:

$$\begin{aligned} \text{Var}(W_2) &= 5^2 \text{Var}(X_1) + 10^2 \text{Var}(X_3) + 2(5)(10) \cdot \text{Cov}(X_1, X_3) \\ &= (25)(31.3) + (100)(125) + (100) \cdot (-59.17) \\ &= 7,365.5 \end{aligned}$$

The effect of the negative covariance is to reduce the variance and hence to reduce the risk of the portfolio.

Chapter 6 Continuous Random Variables

A continuous random variable can take any numerical value in some interval. Assigning probabilities to individual values is not possible. Probabilities can be measured in a given range.

For a continuous random variable X with a numerical value of interest x the **cumulative distribution function (CDF)** is denoted by:

$$\begin{aligned} F(x) &= P(X \leq x) && \text{with } P(X = x) = 0 \\ &= P(X < x) \end{aligned}$$

For two numerical values a and b , with $a < b$, the probability that the outcome is in a range is:

$$\begin{aligned} P(a < X < b) &= P(a \leq X \leq b) \\ &= P(X < b) - P(X < a) \\ &= F(b) - F(a) \end{aligned}$$

The **probability density function (PDF)** is given by:

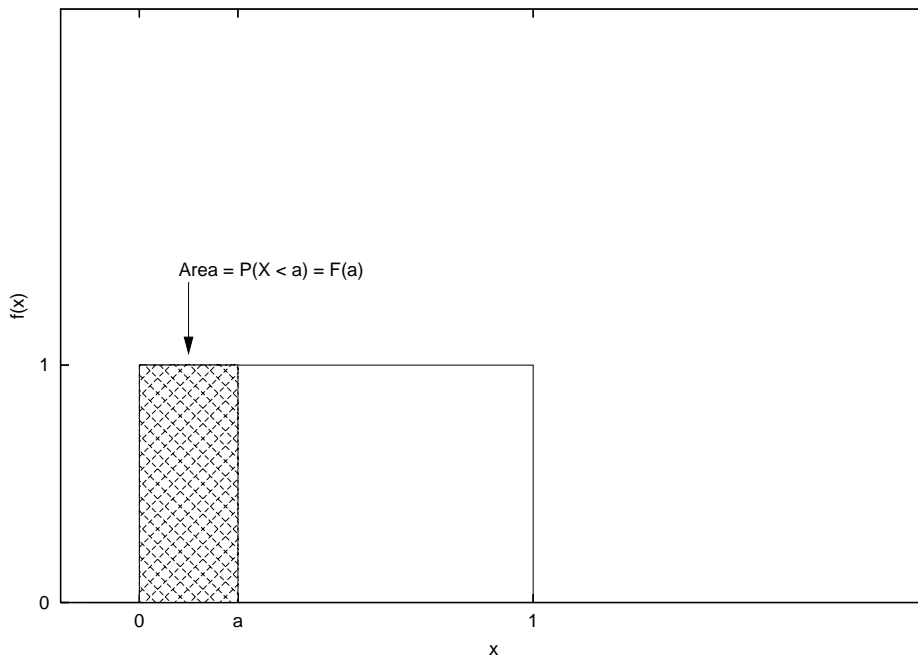
$$f(x) > 0 \quad \text{for all values of } x.$$

The properties of a probability density function can be illustrated with a special distribution called the **uniform distribution**.

The uniform distribution over the interval $[0, 1]$ has the PDF:

$$f(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

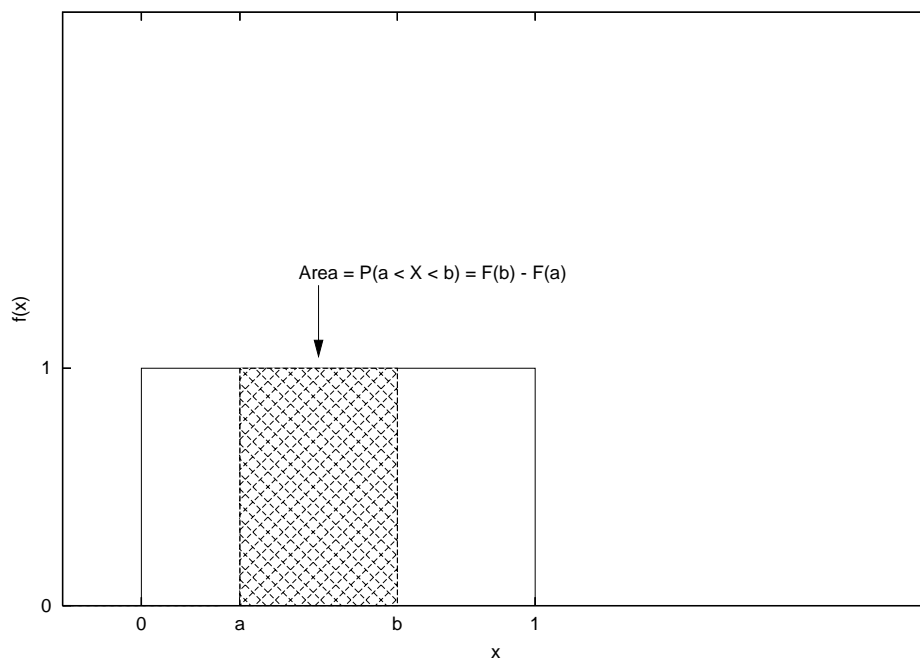
A graph of the probability density function is below.



The important properties of the PDF are:

- the total area under the PDF is equal to one.
- the area under the PDF to the left of the value **a** is **F(a)**.

The next graph illustrates that the PDF can also be used to find a range probability.

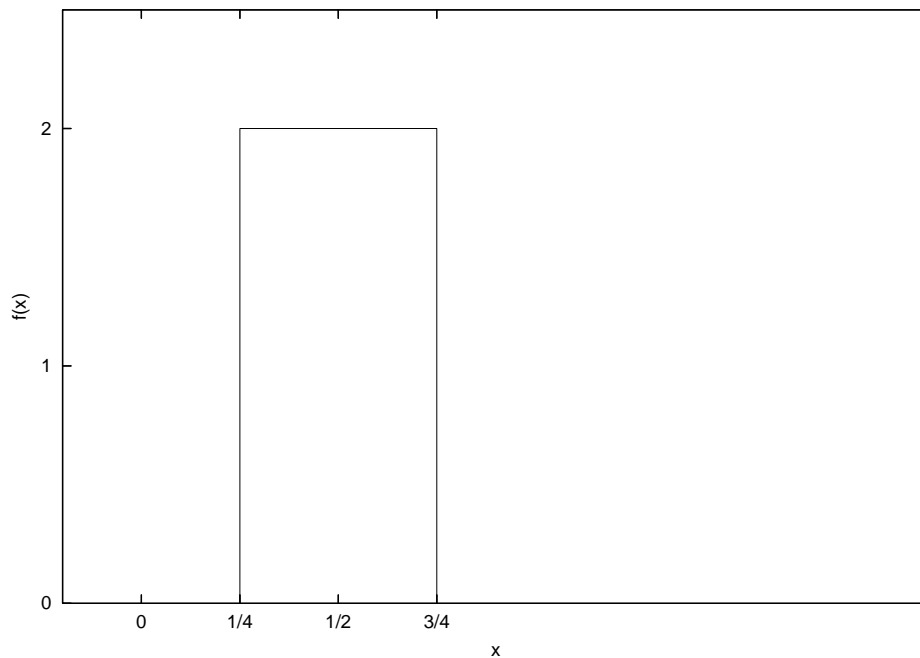


The range probability **$P(a < X < b)$** is the area under the PDF between the values **a** and **b**.

In general, the uniform distribution over the interval $[x_{\min}, x_{\max}]$ has the PDF:

$$f(x) = \begin{cases} \frac{1}{x_{\max} - x_{\min}} & \text{for } x_{\min} < x < x_{\max} \\ 0 & \text{otherwise} \end{cases}$$

For example, consider the uniform distribution over the interval $[\frac{1}{4}, \frac{3}{4}]$. A graph of the probability density function is below:



Again, note that the total area under the PDF is equal to one.

By comparing the graphs of the PDFs for the uniform distribution over the interval $[0, 1]$ and the uniform distribution over $[\frac{1}{4}, \frac{3}{4}]$ it can be seen that both are centered at $\frac{1}{2}$.

However, the two distributions have different dispersion. That is, the PDF for the uniform distribution over $[\frac{1}{4}, \frac{3}{4}]$ has a higher peak to suggest smaller dispersion.

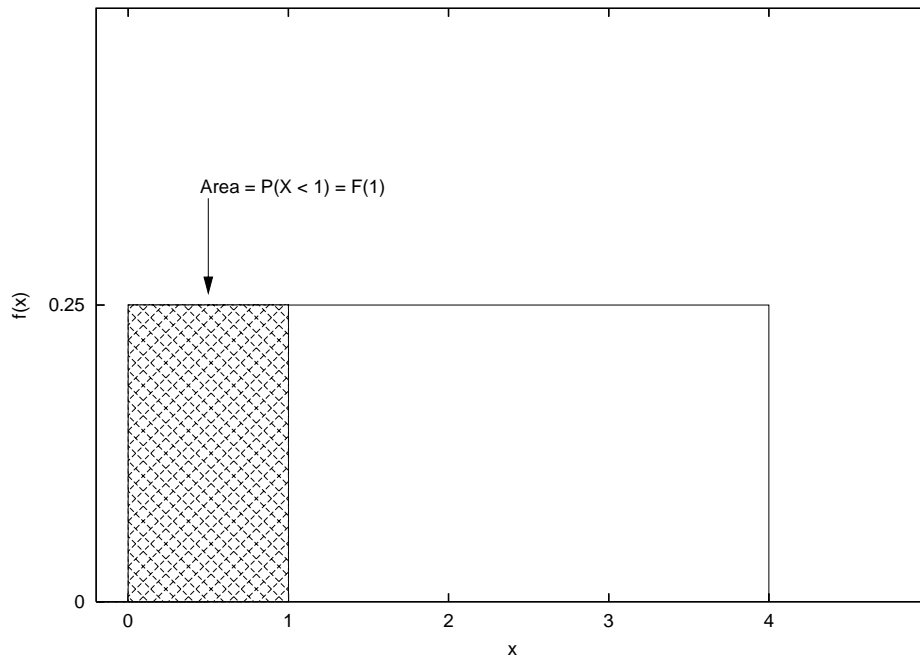
Example: Exercise 6.6, page 193

An emergency rescue team operates on a 4-mile stretch of river. Let the random variable X be the distance (in miles) of an emergency from the northernmost point of this stretch of river. X follows a uniform distribution over the interval $[0, 4]$ with PDF:

$$f(x) = \begin{cases} 0.25 & \text{for } 0 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

Selected questions and answers:

- (c) Find the probability that a given emergency arises within one mile of the northernmost point of this stretch of river. A graph of the PDF is shown:



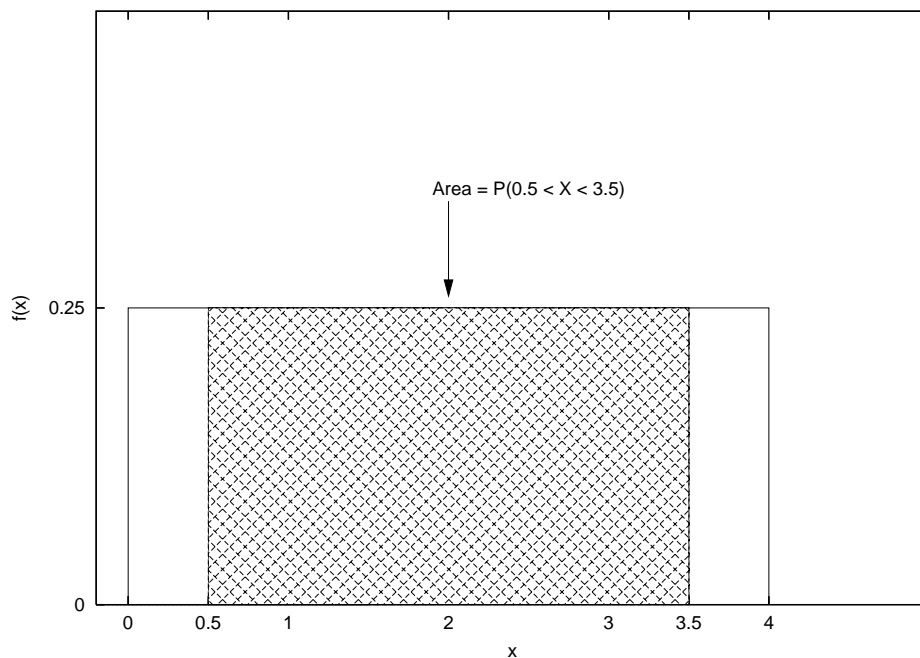
The area of a box is calculated as: (height)·(width).

The answer is: $P(X < 1) = F(1) = (0.25)(1 - 0) = 0.25$

- (d) The rescue team's base is at the mid-point of this stretch of river. Find the probability that a given emergency arises more than 1.5 miles from this base.

First, calculate the probability that an emergency arises within 1.5 miles from the base.

A graph of the PDF is shown:



The answer is:

$$\mathbf{P(0.5 < X < 3.5) = (0.25)(3.5 - 0.5) = 0.75}$$

Also note:

$$\mathbf{P(0.5 < X < 3.5) = F(3.5) - F(0.5)}$$

Therefore, the probability that an emergency is outside the 1.5 mile limit is:

$$\mathbf{1 - P(0.5 < X < 3.5) = 1 - 0.75 = 0.25}$$

Summary information about a probability distribution is provided by the mean and variance.

$E(\mathbf{X})$ is the **expected value** of a random variable \mathbf{X} .

The expected value can be viewed as the average of the observed values from a “large” number of trials of a random experiment.

The **mean** of a random variable \mathbf{X} is denoted by:

$$\mu_{\mathbf{X}} = E(\mathbf{X})$$

A measure of dispersion is the **variance**:

$$\begin{aligned}\sigma_{\mathbf{X}}^2 &= \text{Var}(\mathbf{X}) = E[(\mathbf{X} - \mu_{\mathbf{X}})^2] \\ &= E(\mathbf{X}^2) - \mu_{\mathbf{X}}^2\end{aligned}$$

The **standard deviation** of a random variable \mathbf{X} is defined as:

$$\sigma_{\mathbf{X}} = \sqrt{\sigma_{\mathbf{X}}^2} = \sqrt{\text{Var}(\mathbf{X})} > 0$$

Recall the rules introduced for discrete random variables.
That is, for constant fixed numbers \mathbf{a} and \mathbf{b} :

$$\mathbf{E(a + b X)} = \mathbf{a + b E(X)} = \mathbf{a + b \mu_X} \quad \text{and}$$

$$\mathbf{Var(a + b X)} = \mathbf{b^2 Var(X)}$$

As a special case, the **standardized random variable** is defined as:

$$\mathbf{Z} = \frac{\mathbf{X - \mu_X}}{\sigma_X}$$

The properties of \mathbf{Z} are:

$$\mathbf{E(Z)} = \mathbf{E\left[\frac{X - \mu_X}{\sigma_X}\right]} = \frac{\mathbf{1}}{\sigma_X} \mathbf{E(X - \mu_X)} = \mathbf{0} \quad \text{and}$$

$$\mathbf{Var(Z)} = \mathbf{Var\left[\frac{X - \mu_X}{\sigma_X}\right]} = \frac{\mathbf{1}}{\sigma_X^2} \mathbf{Var(X)} = \mathbf{1}$$

That is, the standardized random variable \mathbf{Z} has mean $\mathbf{0}$ and variance $\mathbf{1}$.

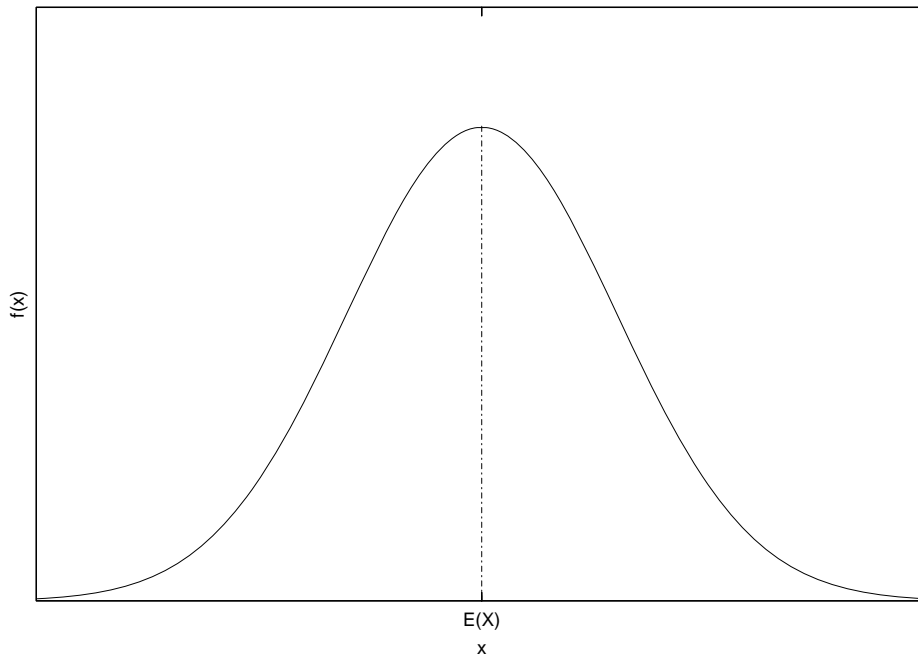
Chapter 6.3 The Normal Distribution

The continuous random variable that follows the **normal distribution** has some popularity in applied work.

The probability density function (PDF) for a normally distributed random variable X with mean μ_X and variance σ_X^2 is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{1}{2\sigma_X^2}(x - \mu_X)^2\right) \quad \text{for } -\infty < x < \infty$$

Graph of the PDF for a Normal Distribution



The shape of the PDF is a symmetric, bell-shaped curve centered on the mean μ_X .

Note: the total area under the PDF curve is equal to one.

To state that a random variable X follows a normal distribution summarized by the parameters mean μ_X and variance σ_X^2 the notation is:

$$X \sim N(\mu_X, \sigma_X^2)$$

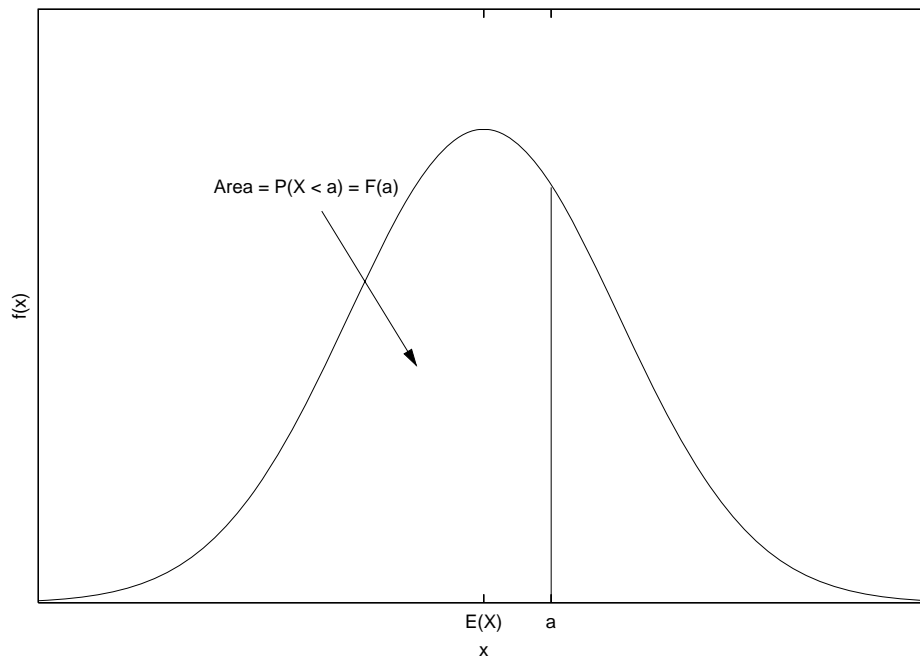
↑

“is distributed as”

The cumulative distribution function (CDF) is:

$$F(x) = P(X \leq x)$$

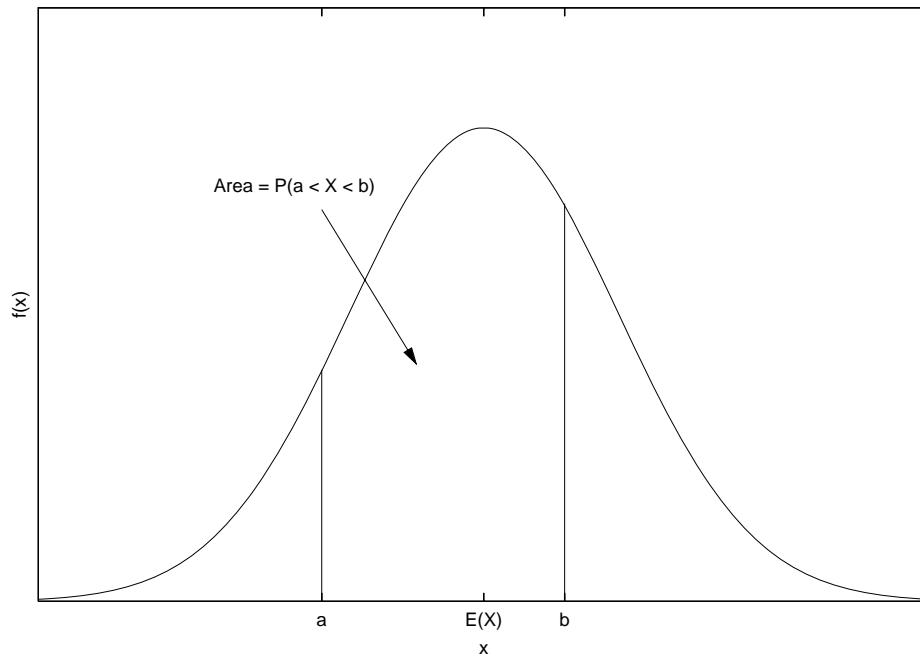
The graph shows that the area under the PDF to the left of the value a is the cumulative probability $F(a)$.



For two values **a** and **b** the range probability is calculated from the CDF as:

$$P(\mathbf{a} < \mathbf{X} < \mathbf{b}) = \mathbf{F}(\mathbf{b}) - \mathbf{F}(\mathbf{a})$$

The next graph shows that the area under the PDF between the values **a** and **b** is the range probability.



A practical problem is that, for the normal distribution, there is no mathematical formula for computing cumulative probabilities.

A quick solution is that computer software offers high accuracy methods for calculating probabilities.

With Microsoft Excel normal distribution probabilities can be obtained by selecting Insert Function NORMDIST.

The general usage is: $\text{NORMDIST}(x, \mu_X, \sigma_X, \text{cumulative})$

where cumulative = 0 for the PDF,
 cumulative = 1 for the CDF

Working with the Normal Distribution

Before the days of high speed laptop computers, applied workers used statistical tables (printed in the Appendix to statistical textbooks) to look-up normal distribution probabilities.

Working with the statistical tables can be useful as a learning exercise as it gives emphasis to understanding the properties of the normal distribution. Therefore, as a check on the calculations that can be obtained with Microsoft Excel, the use of the normal distribution tables will be described here.

It can be noted that probabilities depend on the setting of μ_X and σ_X , the mean and standard deviation of the random variable. However, it turns out that probabilities for the **standard normal** random variable Z with mean 0 and variance 1 can be used to calculate probabilities for any other normal distribution.

The course textbook prints two versions of a table for the cumulative distribution function for the standard normal random variable:

$$Z = \frac{X - \mu_X}{\sigma_X} \sim N(0, 1)$$

The textbook tables are: (1) Appendix Table 1, pages 841-2.

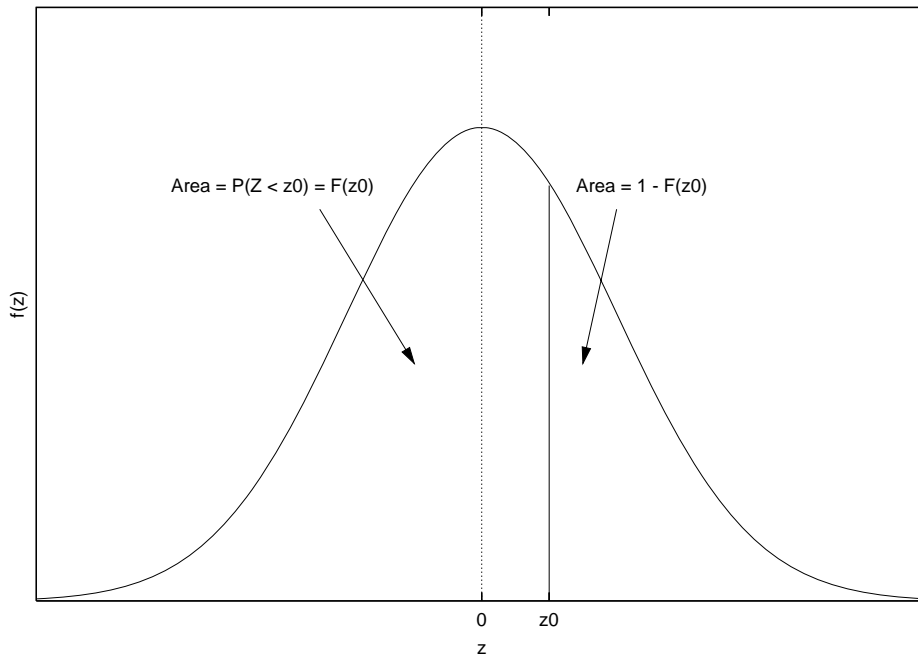
(2) Inside front cover

Appendix Table 1 will be used in the work here.

How is the table read ?

A graph is useful.

Probability Density Function (PDF) of Z



For a value of interest z_0 the table gives the cumulative probability:

$$F(z_0) = P(Z \leq z_0)$$

The table lists values for $z_0 \geq 0$ only.

From **symmetry** of the normal distribution:

$$F(-z_0) = P(Z \leq -z_0)$$

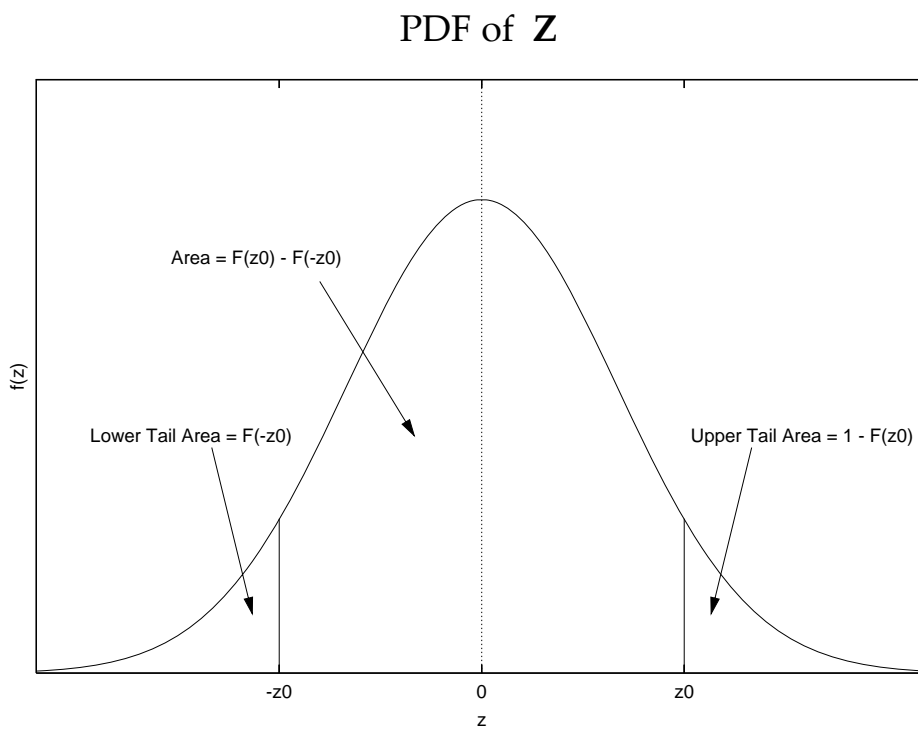
$$= P(Z \geq z_0)$$

$$= 1 - F(z_0)$$

A result for a range probability with symmetric upper and lower values can be stated. For some value z_0 :

$$\begin{aligned} P(-z_0 \leq Z \leq z_0) &= P(Z \leq z_0) - P(Z \leq -z_0) \\ &= F(z_0) - [1 - F(z_0)] \\ &= 2F(z_0) - 1 \end{aligned}$$

This is shown with a graph.



By symmetry of the normal distribution the area in the “lower tail” is identical to the area in the “upper tail.”

Now suppose the random variable to work with is:

$$X \sim N(\mu, \sigma^2)$$

For two numerical values a and b , with $a < b$, a probability of interest is:

$$P(a < X < b)$$

This probability statement can be transformed to a probability statement about the standard normal random variable Z .

This is done as follows:

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) \\ &= P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) \\ &= F\left(\frac{b - \mu}{\sigma}\right) - F\left(\frac{a - \mu}{\sigma}\right) \end{aligned}$$

Appendix Table 1 of the textbook can now be used to look-up the cumulative probabilities for the standard normal distribution.

Example: Exercise 6.22, page 208.

Let the continuous random variable X be the amount of money spent on textbooks by a student in September of the academic year. It is known that:

$$X \sim N(\mu, \sigma^2) \quad \text{with} \quad \mu = \$380 \quad \text{and} \quad \sigma = \$50$$

Questions and Answers

(a) Find $P(X < 400)$.

This gives the probability that a randomly chosen student will spend less than \$400 on textbooks in September.

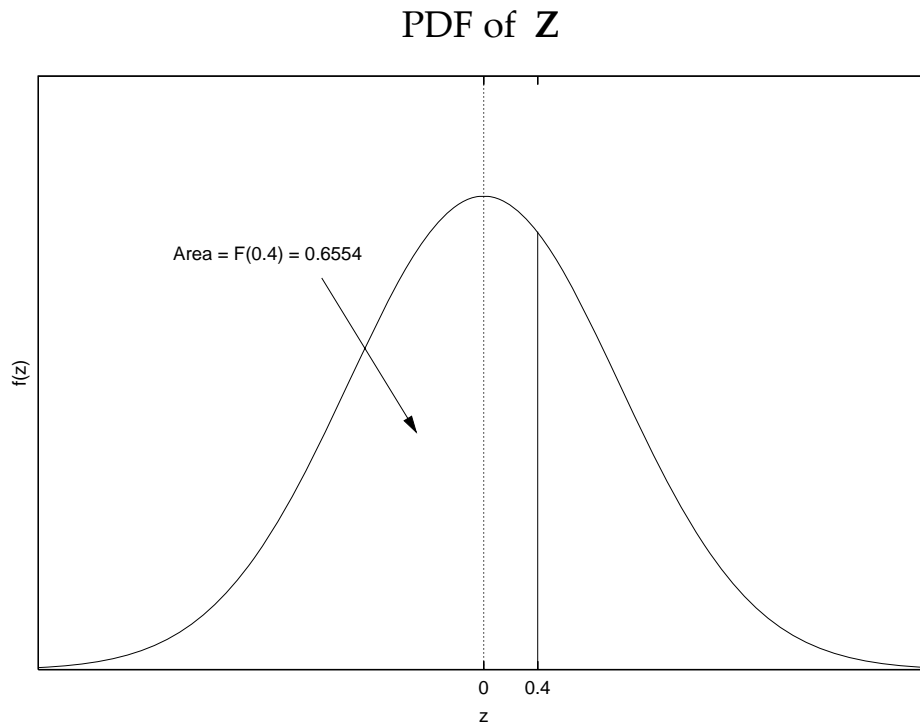
First state the probability as a probability about the standard normal variable Z :

$$\begin{aligned} P(X < 400) &= P\left(\frac{X - \mu}{\sigma} < \frac{400 - \mu}{\sigma}\right) \\ &= P\left(Z < \frac{400 - 380}{50}\right) \\ &= P(Z < 0.4) \\ &= F(0.4) \end{aligned}$$

Now look-up the answer in Appendix Table 1 of the textbook. The table gives: $F(0.4) = 0.6554$

Therefore, $P(X < 400) = 0.6554$

A graph gives a helpful illustration of the use of the statistical tables for this problem.



Now check the answer with Microsoft Excel by selecting Insert Function:

$\text{NORMDIST}(x, \mu_X, \sigma_X, \text{cumulative})$

Enter the values:

$\text{NORMDIST}(400, 380, 50, 1)$

This returns the probability: 0.6554

(b) & (c) Find $\mathbf{P(X > 360)}$.

This gives the probability that a randomly chosen student will spend more than \$360 on textbooks in September.

Express the problem in the form of a probability statement about the standard normal variable \mathbf{Z} :

$$\begin{aligned}\mathbf{P(X > 360)} &= \mathbf{P\left(\frac{X - \mu}{\sigma} > \frac{360 - \mu}{\sigma}\right)} \\ &= \mathbf{P\left(Z > \frac{360 - 380}{50}\right)} \\ &= \mathbf{P(Z > -0.4)} \\ &= \mathbf{P(Z < 0.4)} && \text{by symmetry} \\ &= \mathbf{F(0.4)}\end{aligned}$$

This is identical to the probability calculated for part (a).

That is, $\mathbf{P(X > 360) = 0.6554}$

The answers in parts (a) and (b) are the same because the normal distribution is symmetric about the mean.

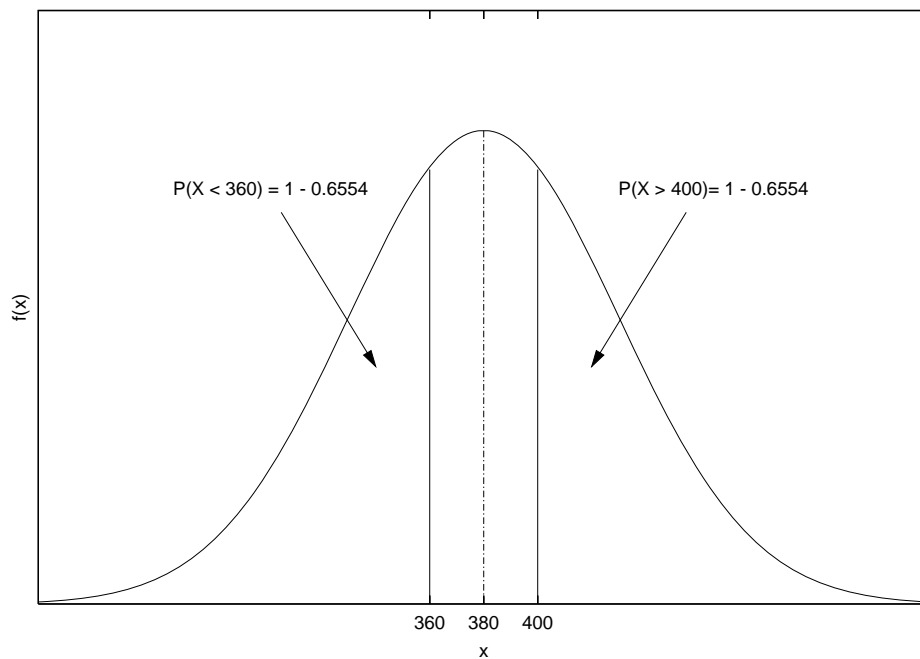
The graph below demonstrates that because of symmetry about the mean:

$$P(X > 360) = P(X < 400)$$

Also,

$$P(X < 360) = P(X > 400)$$

PDF of $X \sim N(380, 50^2)$



(d) Find $\mathbf{P(300 < X < 400)}$.

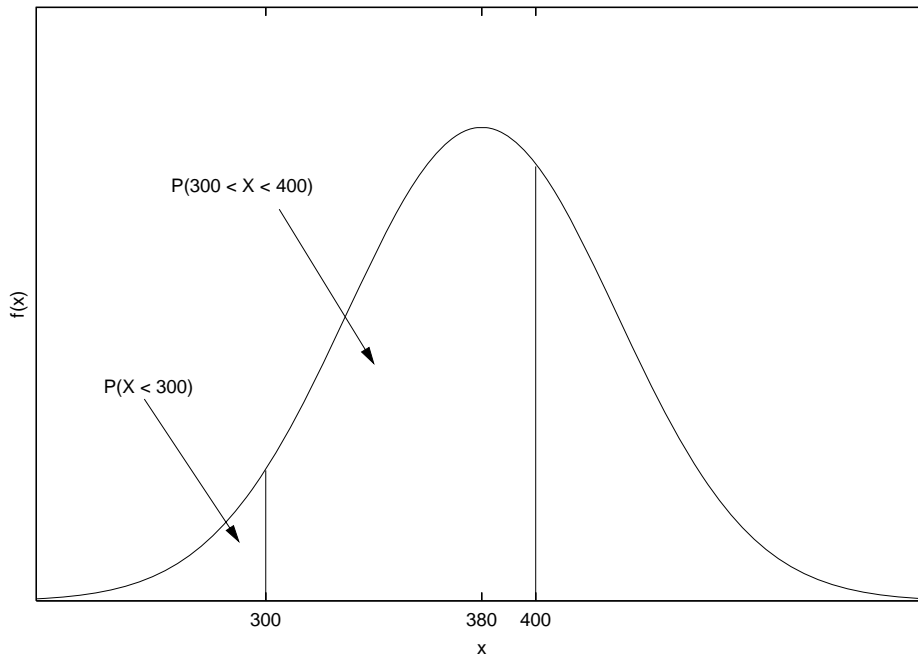
This gives the probability that a randomly chosen student will spend between \$300 and \$400 on textbooks in September.

The range probability is calculated as:

$$\mathbf{P(300 < X < 400) = P(X < 400) - P(X < 300)}$$

A graph gives a helpful picture of the calculations.

PDF of $\mathbf{X \sim N(380, 50^2)}$



From the previous calculations: $\mathbf{P(X < 400) = 0.6554}$

Now find:

$$\begin{aligned}\mathbf{P(X < 300)} &= \mathbf{P\left(\frac{X - \mu}{\sigma} < \frac{300 - \mu}{\sigma}\right)} \\ &= \mathbf{P\left(Z < \frac{300 - 380}{50}\right)} \\ &= \mathbf{P(Z < -1.6)} \\ &= \mathbf{1 - P(Z < 1.6)} \quad \text{by symmetry} \\ &= \mathbf{1 - F(1.6)}\end{aligned}$$

A look-up in Appendix Table 1 gives: $\mathbf{F(1.6) = 0.9452}$

The answer is:

$$\begin{aligned}\mathbf{P(300 < X < 400)} &= \mathbf{0.6554 - (1 - 0.9452)} \\ &= \mathbf{0.60}\end{aligned}$$

❖ Finding Cutoff Points or Critical Values

A problem that has been presented is: What is the probability that values will occur in some range ?

Another problem is: What numerical value corresponds to a probability of 10% ?

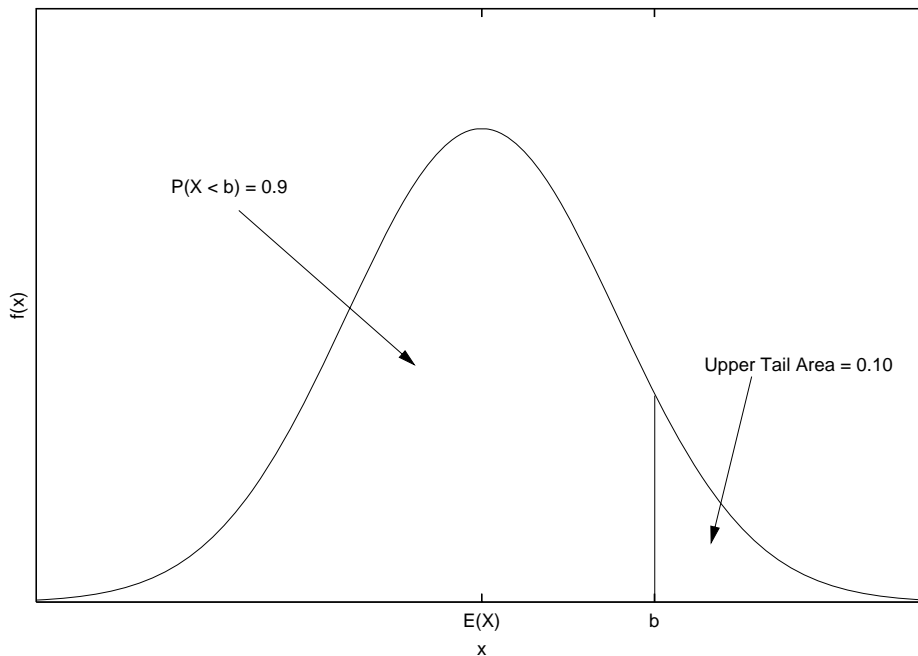
That is, find the value **b** such that:

$$P(X > b) = 0.10$$

where the probability of 10% can be varied to any level of interest.

A graph of the problem is below.

PDF of $X \sim N(\mu, \sigma^2)$



A probability result is: $\mathbf{P(X > b) = 1 - P(X < b)}$

Therefore, as shown in the above graph, the problem is to find the value \mathbf{b} such that:

$$\mathbf{P(X < b) = 0.90}$$

A result is:

$$\begin{aligned}\mathbf{P(X < b)} &= \mathbf{P\left(Z < \frac{b - \mu}{\sigma}\right)} \\ &= \mathbf{F\left(\frac{b - \mu}{\sigma}\right)}\end{aligned}$$

Appendix Table 1 gives $\mathbf{F(1.28) = 0.90}$
(some approximation was used).

Therefore,
$$\frac{\mathbf{b - \mu}}{\sigma} = 1.28$$

Rearranging gives:
$$\mathbf{b = \mu + 1.28 \sigma}$$

The cutoff point (or critical value) \mathbf{b} can be computed with Microsoft Excel with the function:

$$\text{NORMINV}(\text{probability}, \mu_X, \sigma_X)$$

Cutoff points from the standard normal distribution are computed with the function:

$$\text{NORMSINV}(\text{probability})$$

For example, to find the value $\mathbf{z_0}$ such that $\mathbf{F(z_0) = 0.90}$ with Microsoft Excel select Insert Function:

$$\text{NORMSINV}(0.9) \quad \text{or}$$

$$\text{NORMINV}(0.9, 0, 1)$$

Both these functions return the answer $\mathbf{z_0 = 1.2816}$

Example: student textbook buying exercise Continued

- (e) Find a range of dollar spending on textbooks that includes 80% of all students.

Any number of ranges can be found. That is, a variety of values x_0 and x_1 with $x_0 < 380$ and $x_1 > 380$ will satisfy:

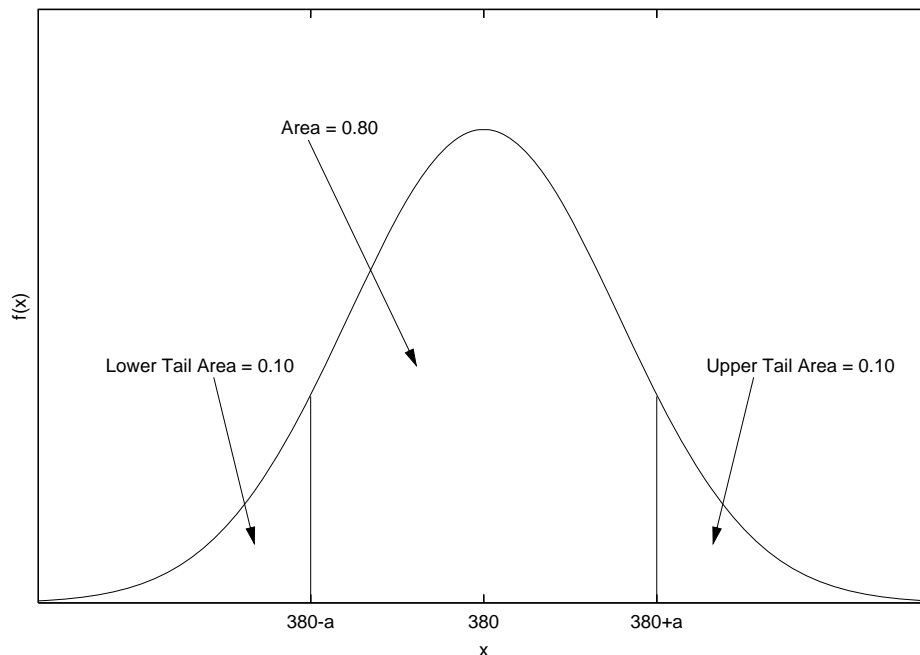
$$P(x_0 < X < x_1) = 0.80$$

The shortest range is centered at the mean \$380.
To calculate this range, find a number a such that:

$$P(380 - a < X < 380 + a) = 0.80$$

This is illustrated with a graph:

PDF of $X \sim N(380, 50^2)$



By inspecting the graph, it can be seen that an equivalent statement of the problem is: find a number **a** such that:

$$\mathbf{P(X < 380 + a) = 0.90}$$

To work with the standard normal distribution consider:

$$\begin{aligned} \mathbf{P(X < 380 + a)} &= \mathbf{P\left(\frac{X - \mu}{\sigma} < \frac{(380 + a) - 380}{50}\right)} \\ &= \mathbf{P\left(Z < \frac{a}{50}\right)} \\ &= \mathbf{F\left(\frac{a}{50}\right)} \end{aligned}$$

Appendix Table 1 gives $\mathbf{F(1.28) = 0.90}$

Therefore, $\frac{\mathbf{a}}{\mathbf{50}} = 1.28$ and

$$\mathbf{a = (1.28)(50) = 64}$$

The range centered at \$380 is:

$$[\$380 - 64, \$380 + 64] = [\$316, \$444]$$

As a check on the calculations, the upper limit can be calculated with Microsoft Excel by using the function:

$$\mathbf{NORMINV(0.9, 380, 50)}$$

Chapter 6.6 Jointly Distributed Continuous Random Variables

Results stated earlier for jointly distributed discrete random variables can be extended to work with continuous random variables.

Let X and Y be two continuous random variables that take numeric values denoted by x and y , respectively.

The joint cumulative distribution function (CDF) is:

$$F_{X,Y}(x,y) = P(X < x \text{ and } Y < y)$$

The marginal distribution functions are:

$$F_X(x) = P(X < x) \quad \text{and} \quad F_Y(y) = P(Y < y)$$

X and Y are statistically independent if and only if:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

A measure of **linear** association is covariance:

$$\begin{aligned}\mathbf{Cov(X, Y)} &= \mathbf{E[(X - \mu_X)(Y - \mu_Y)]} \\ &= \mathbf{E(XY) - \mu_X \mu_Y}\end{aligned}$$

where $\mu_X = \mathbf{E(X)}$ and $\mu_Y = \mathbf{E(Y)}$

If \mathbf{X} and \mathbf{Y} are independent then $\mathbf{Cov(X, Y) = 0}$.

However, zero covariance does not guarantee independence.

\mathbf{X} and \mathbf{Y} may have some complicated non-linear relationship.

- Special Case: If \mathbf{X} and \mathbf{Y} are joint **normally distributed** random variables then zero covariance also gives the result that \mathbf{X} and \mathbf{Y} are independent.

❖ Linear Combinations of Random Variables

For constant fixed numbers \mathbf{a} and \mathbf{b} , a linear combination of random variables \mathbf{X} and \mathbf{Y} is:

$$\mathbf{W} = \mathbf{a}\mathbf{X} + \mathbf{b}\mathbf{Y}$$

The mean of the random variable \mathbf{W} is:

$$\mu_{\mathbf{W}} = \mathbf{E}(\mathbf{W}) = \mathbf{a}\mathbf{E}(\mathbf{X}) + \mathbf{b}\mathbf{E}(\mathbf{Y})$$

The variance of \mathbf{W} is:

$$\sigma_{\mathbf{W}}^2 = \mathbf{Var}(\mathbf{W}) = \mathbf{a}^2 \mathbf{Var}(\mathbf{X}) + \mathbf{b}^2 \mathbf{Var}(\mathbf{Y}) + 2\mathbf{a}\mathbf{b}\mathbf{Cov}(\mathbf{X}, \mathbf{Y})$$

- Special Case: If \mathbf{X} and \mathbf{Y} are joint **normally distributed** random variables then $\mathbf{W} = \mathbf{a}\mathbf{X} + \mathbf{b}\mathbf{Y}$ is also normally distributed with mean and variance as given above. That is,

$$\mathbf{W} \sim \mathbf{N}(\mu_{\mathbf{W}}, \sigma_{\mathbf{W}}^2)$$

Chapter 7 Sampling and Sampling Distributions

A **random sample** is a set of random variables $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ (upper case notation) that are:

- identically distributed.
That is, each of these random variables has mean μ and variance σ^2 ; and
- independently distributed.
That is, $\mathbf{Cov}(\mathbf{X}_i, \mathbf{X}_j) = 0$ for any $i \neq j$.

Typically, the population parameters (such as μ and σ^2) are unknown.

A sample of data are the observed numerical outcomes $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ (lower case notation).
The sample mean can be calculated as:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

Clearly, $\bar{\mathbf{x}}$ will not be identical to the population mean $\boldsymbol{\mu}$.

For a second sample of n observations denote the numerical outcomes as: $\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_n^*$

From this sample the sample mean is: $\bar{\mathbf{x}}^* = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^*$

The two calculated sample means $\bar{\mathbf{x}}$ and $\bar{\mathbf{x}}^*$ will be different numbers and neither will be the same as the population mean $\boldsymbol{\mu}$.

That is, different samples of n observations have different numerical observations and therefore, the calculated sample means are different.

The **sample mean** of the random variables X_1, X_2, \dots, X_n is defined as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

\bar{X} is a linear combination of random variables and, therefore, is also a random variable.

\bar{X} has a probability distribution known as the **sampling distribution**. The sampling distribution of a sample statistic is the probability distribution of the values it could take over all possible samples of size n drawn from the population.

What are the properties of the sampling distribution of \bar{X} ?

First, state the mean:

$$\begin{aligned} E(\bar{X}) &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E(X_i) \\ &= \frac{1}{n} (n\mu) \\ &= \mu \end{aligned}$$

That is, $E(\bar{X}) = \mu$.

This says – for a large number of samples (say 1000 samples), each with n observations, the average of the calculated sample means will approach the population mean μ .

Now state the variance:

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \quad \text{use independence} \\ &= \frac{1}{n^2} (n \sigma^2) \\ &= \sigma^2 / n\end{aligned}$$

Problem: How did the assumption of independence simplify the variance formula ?

That is, $\text{Var}(\bar{X}) = \sigma^2 / n$

This gives the result that as the sample size n increases the variance of the sample mean decreases.

The standard deviation of the sampling distribution of \bar{X} is called the **standard error** of \bar{X} . This is:

$$\text{se}(\bar{X}) = \sigma / \sqrt{n}$$

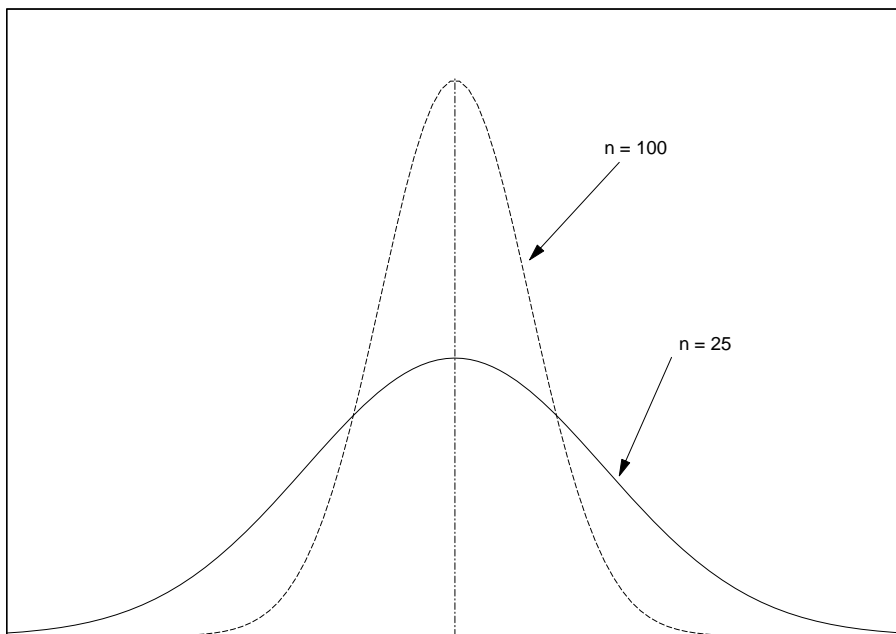
Now introduce the assumption of normality.

Let the random sample $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be a set of normally distributed and independent random variables with mean μ and variance σ^2 .

It follows that $\bar{\mathbf{X}}$ is also normally distributed (recall that an earlier result stated that a linear combination of normally distributed random variables is also normally distributed). That is,

$$\bar{\mathbf{X}} \sim \mathbf{N}\left(\mu, \sigma^2/n\right)$$

PDF of $\bar{\mathbf{X}}$ for $n=25$ and $n=100$



Note: the total area under each curve is equal to one.

The graph on the previous page shows the probability density function of the sampling distribution of the sample mean.

This is centered at μ .

The graph demonstrates that as the sample size n increases, the variance decreases, and the distribution becomes more concentrated around the population mean.

A standardized normal random variable can be stated:

$$Z = \frac{\bar{X} - \mu}{\text{se}(\bar{X})} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

Probability statements about the mean can now be considered.

Example: Exercise 7.17, page 253.

Times spent studying by students in the week before final exams follow a normal distribution with standard deviation 8 hours. A random sample of 4 students was taken in order to estimate the mean study time for the population of all students.

Questions and Answers

- (a) What is the probability that the sample mean exceeds the population mean by more than 2 hours ?

That is, find: $\mathbf{P(\bar{X} > \mu + 2)}$

With $\mathbf{n = 4}$ the standard error of the sample mean $\bar{\mathbf{X}}$ is:

$$\mathbf{se(\bar{X}) = \sigma / \sqrt{n} = 8 / 2 = 4}$$

Write the problem as a probability statement about the standard normal random variable \mathbf{Z} :

$$\begin{aligned} \mathbf{P(\bar{X} > \mu + 2)} &= \mathbf{P\left(\frac{\bar{X} - \mu}{\mathbf{se(\bar{X})}} > \frac{(\mu + 2) - \mu}{\mathbf{se(\bar{X})}}\right)} \\ &= \mathbf{P\left(Z > \frac{2}{4}\right)} \\ &= \mathbf{1 - P(Z < 0.5)} && \text{by symmetry} \\ &= \mathbf{1 - F(0.5)} \end{aligned}$$

A look-up in Appendix Table 1 gives: $\mathbf{F(0.5) = 0.6915}$

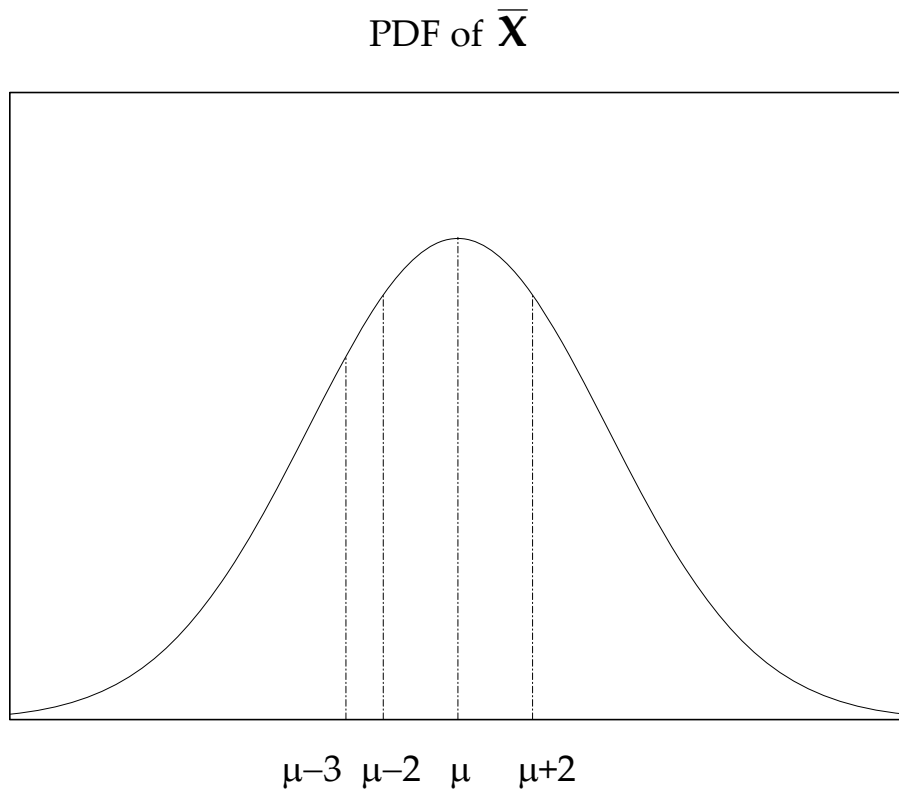
The answer is: $\mathbf{P(\bar{X} > \mu + 2) = 1 - 0.6915 = 0.3085}$

(b) What is the probability that the sample mean is more than 3 hours below the population mean ?

That is, find: $\mathbf{P(\bar{X} < \mu - 3)}$

Note: $\mathbf{P(\bar{X} < \mu - 3) < P(\bar{X} < \mu - 2) = P(\bar{X} > \mu + 2)}$

This is illustrated in the graph below.



Therefore, the answer must be smaller than the probability calculated for part (a).

The solution method follows similar steps to part (a):

$$\begin{aligned} P(\bar{X} < \mu - 3) &= P\left(\frac{\bar{X} - \mu}{se(\bar{X})} < \frac{(\mu - 3) - \mu}{se(\bar{X})}\right) \\ &= P\left(Z < -\frac{3}{4}\right) \\ &= 1 - P(Z < 0.75) \\ &= 1 - F(0.75) \end{aligned}$$

A look-up in Appendix Table 1 gives: $F(0.75) = 0.7734$

The answer is: $P(\bar{X} < \mu - 3) = 1 - 0.7734 = 0.2266$

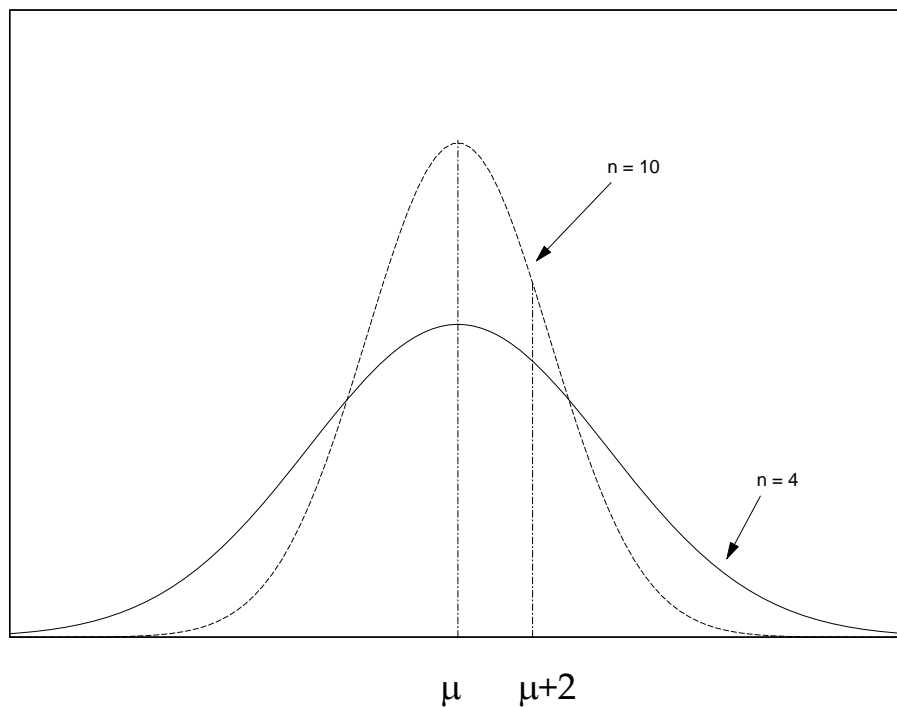
- (d) Suppose that a second (independent) random sample of 10 students was taken. Without doing the calculations, state whether the probabilities in part (a) would be higher, lower, or the same for the second sample.

The standard error of \bar{X} is: $se(\bar{X}) = \sigma / \sqrt{n}$

An increase in the sample size from $n=4$ to $n=10$ gives a smaller standard error. This leads to more concentration about the population mean μ and so $P(\bar{X} > \mu + 2)$ becomes **lower**.

This is illustrated in the graph.

PDF of \bar{X}



The Central Limit Theorem

The normal distribution is a convenient approximation in many applications. The Central Limit Theorem gives a justification for this.

Let the random sample X_1, X_2, \dots, X_n be a set of random variables that are independently and identically distributed with mean μ and variance σ^2 .

The random variables need not follow the normal distribution – they may fit a skewed distribution or any other non-normal distribution.

Consider the sample mean:
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Earlier lecture notes stated the standard error of \bar{X} as:

$$se(\bar{X}) = \sigma / \sqrt{n}$$

Define the standardized random variable:

$$Z = \frac{\bar{X} - \mu}{se(\bar{X})} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

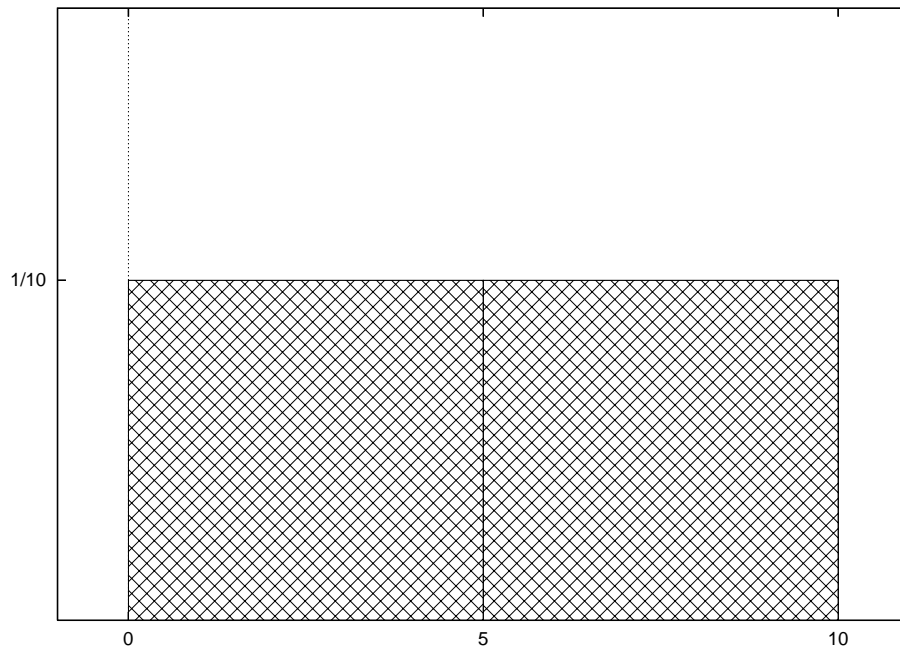
The **Central Limit Theorem** states that as n becomes ‘large’ the distribution of Z approaches the standard normal distribution. That is, $Z \sim N(0,1)$.

Therefore, a random variable that can be viewed as the sum of a ‘large’ number of independently and identically distributed random variables will tend to have a normal distribution.

A computer simulation can be used to demonstrate the Central Limit Theorem.

Consider a random variable that follows the uniform distribution over the interval $[0, 10]$.

A graph of the probability density function is below.

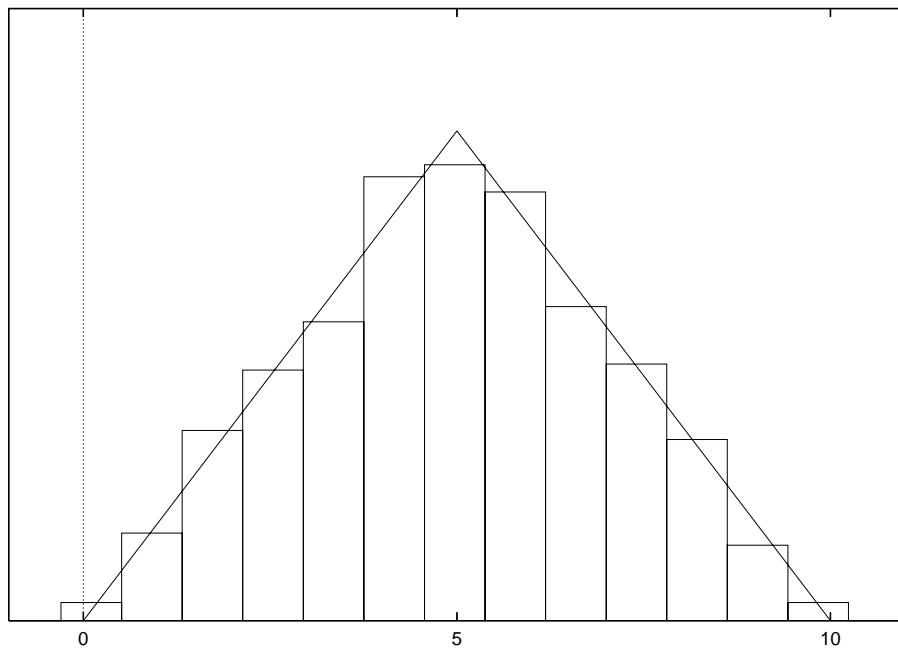


A statistical result is that the average of two ($n=2$) independent uniform random variables has a triangular shape for the probability density function.

To show this result with a computer simulation, select outcomes from two uniform random variables on the interval $[0, 10]$ and calculate the average of the two values. Repeat this 1000 times.

A histogram gives a graph of the frequency distribution of the sample means generated by the computer simulation. This gives a rough picture of the probability density function of the sample mean.

The histogram generated by the computer simulation is shown below. The triangular shape of the theoretical probability density function has also been sketched.

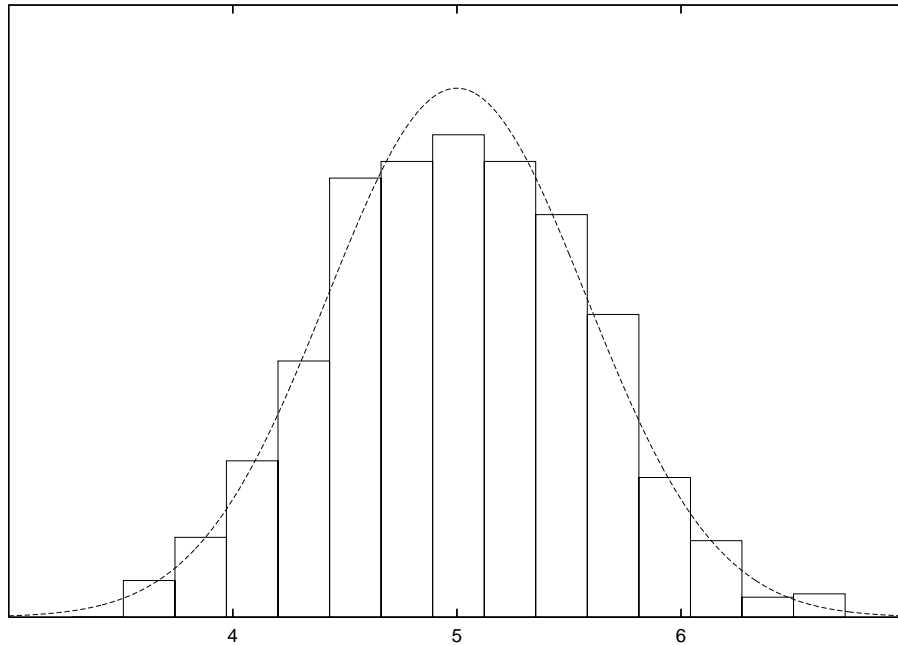


Now consider a sample size of $n=25$.

By the Central Limit Theorem it may be reasonable to assume that the sample mean follows a normal distribution.

To start the computer simulation, select outcomes from 25 uniform random variables on the interval $[0, 10]$ and calculate the average of the values. Repeat this 1000 times.

The histogram generated by the computer simulation is shown below. A normal probability density function is also sketched to reveal that, with $n=25$, the distribution of the sample mean is closely approximated by the normal distribution.



Chapter 7.4 The Sample Variance

Let the random sample X_1, X_2, \dots, X_n be a set of identically distributed and independent random variables with mean μ and variance σ^2 .

The sample mean is defined as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Previous work has studied the properties of the sampling distribution of the sample mean. A familiar result is:

$$\bar{X} \sim N\left(\mu, \sigma^2/n\right)$$

Now consider the sample variance defined as the random variable:

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

What are the properties of the sampling distribution of s_X^2 ?

A result is:

$$E(s_X^2) = \sigma^2$$

This can be shown as follows.

$$\begin{aligned} E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] &= E\left[\sum_{i=1}^n \{(X_i - \mu) - (\bar{X} - \mu)\}^2\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu)^2 - n \cdot (\bar{X} - \mu)^2\right] \\ &= \sum_{i=1}^n E[(X_i - \mu)^2] - n \cdot E[(\bar{X} - \mu)^2] \\ &= n\sigma^2 - n \cdot \frac{\sigma^2}{n} \\ &= (n-1)\sigma^2 \end{aligned}$$

Therefore,

$$\begin{aligned} E(s_X^2) &= \frac{1}{n-1} [(n-1)\sigma^2] \\ &= \sigma^2 \end{aligned}$$

That is, the divisor of $(n-1)$ ensures that s_X^2 gives an unbiased estimation rule for the population parameter σ^2 .

This justifies the use of $(n-1)$ in the divisor.

Another statistical result about probability distributions is needed.

Let Z_1, Z_2, \dots, Z_m be a set of independent standard normal random variables. Define the random variable:

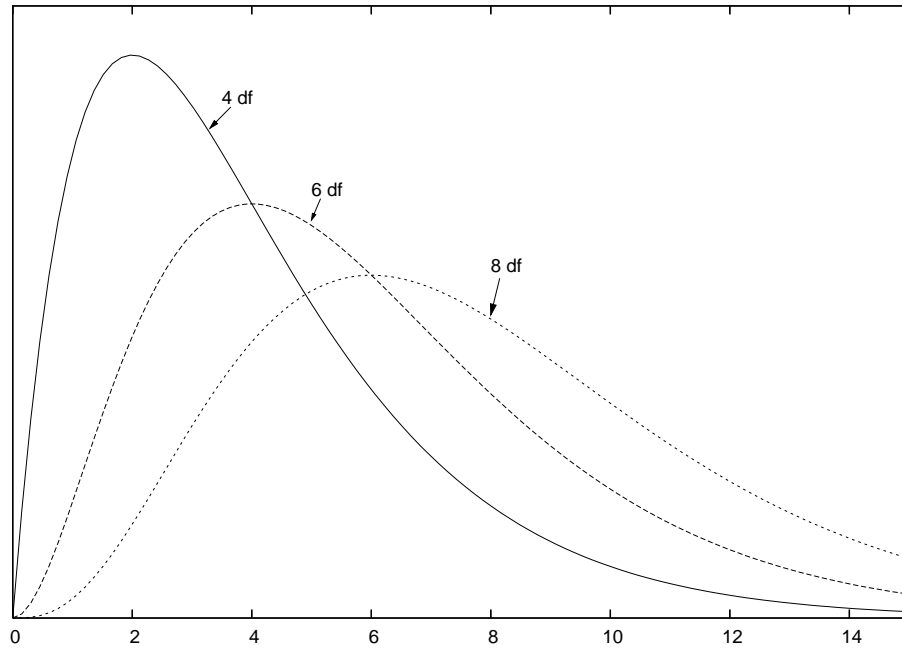
$$C = \sum_{i=1}^m Z_i^2$$

C has a χ^2 (chi-square – pronounced ki-square) distribution with m degrees of freedom.

Properties of the probability density function for the chi-square distribution are:

- defined only for values ≥ 0 ,
- a skewed shape that depends on the degrees of freedom.

PDF for the χ^2 distribution with 4, 6 and 8 degrees of freedom (df).



The critical values or cut-off points of the chi-square distribution are listed in Appendix Table 7, page 869 of the textbook.

Assume that X_1, X_2, \dots, X_n are normally distributed random variables.

Define the random variable:

$$\frac{(n-1)s_X^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$$

This random variable has a chi-square distribution with $(n - 1)$ degrees of freedom.

Probability statements about the variance can now be made.

Example: Exercise 7.71 (c), page 270 of the textbook.

Let the random variable X be the time to complete a tax form.

Assume that X follows a normal distribution with mean $\mu = 100$ minutes and standard deviation $\sigma = 30$.

A random sample of $n = 9$ tax filers is taken.

Find a value a such that:

$$\mathbf{P}(s_X < a) = 0.05$$

This is equivalent to finding a value a such that:

$$\mathbf{P}(s_X^2 < a^2) = 0.05$$

The probability can be stated:

$$\begin{aligned}\mathbf{P}(s_X^2 < a^2) &= \mathbf{P}\left(\frac{(n-1)s_X^2}{\sigma^2} < \frac{(n-1)a^2}{\sigma^2}\right) \\ &= \mathbf{P}\left(\chi_{(8)}^2 < \frac{8a^2}{(30)(30)}\right) \\ &= 0.05\end{aligned}$$

From Appendix Table 7: $\mathbf{P}(\chi_{(8)}^2 < 2.73) = 0.05$

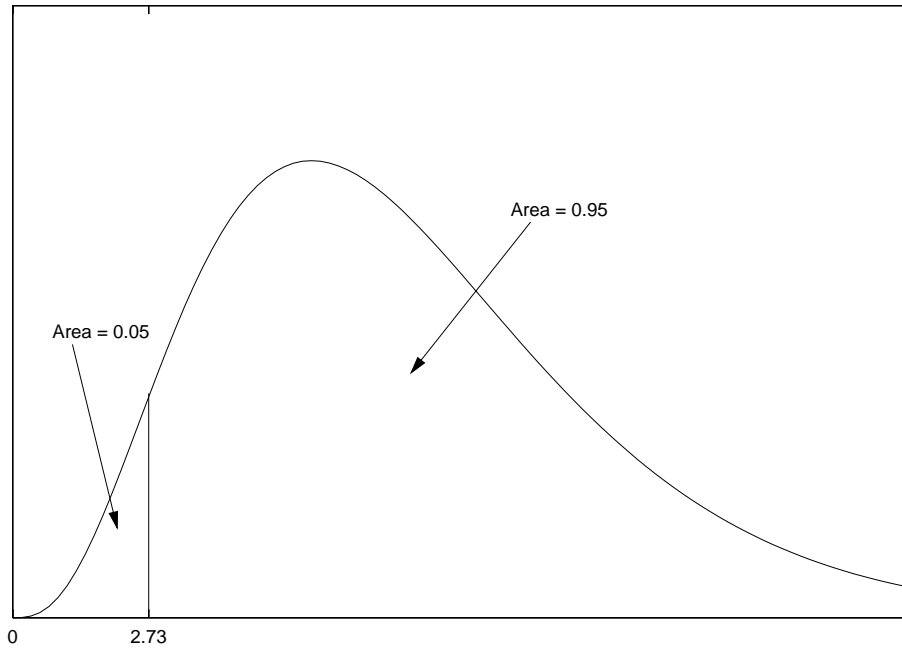
Therefore,
$$\frac{8a^2}{(30)(30)} = 2.73$$

Solving gives:
$$a = 30 \cdot \sqrt{\frac{2.73}{8}} = 17.52$$

That is, the probability is 0.05 that the sample standard deviation of time taken to complete the tax form is less than 17.52 minutes.

The graph below illustrates $P(\chi_{(8)}^2 < 2.73) = 0.05$

PDF of $\chi_{(8)}^2$



Chapter 8.1 Point Estimation

Let the random sample $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be a set of random variables that are independently and identically distributed.

Population characteristics are summarized by parameters – the true values are typically unknown.

For example, the population mean is denoted by μ .

An estimation rule can be specified for a parameter of interest. This estimation rule is called a **point estimator**.

For example, a point estimator for the population mean μ is:

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$$

An estimator is a random variable that is a function of the sample information. An estimator has a probability distribution called the sampling distribution.

An applied study works with a data set.

The numeric observations are: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$.

The estimation rule given by the estimator $\bar{\mathbf{X}}$ can be used to calculate a **point estimate** of the population mean $\boldsymbol{\mu}$:

$$\bar{\mathbf{x}} = \frac{1}{\mathbf{n}} \sum_{i=1}^{\mathbf{n}} \mathbf{x}_i$$

- An important distinction is made between an estimator and an estimate.

A point estimator is a random variable.

A point estimate is a numeric outcome.

Different samples of data will have different numeric observations and, therefore, will result in different point estimates of the population parameter.

❖ Properties of Point Estimators

Denote θ (the Greek letter theta) as a population parameter to be estimated (as a special case this may be the population mean μ).

Let $\hat{\theta}$ (theta-hat) be a point estimator of θ .
 $\hat{\theta}$ is a function of the sample information:

$$\hat{\theta} = f(X_1, X_2, \dots, X_n)$$

This estimator is a random variable with a sampling distribution.

$\hat{\theta}$ is said to be an **unbiased estimator** of θ if:

$$E(\hat{\theta}) = \theta$$

The **bias** of an estimator $\hat{\theta}$ is defined as:

$$\mathbf{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

It follows that the bias of an unbiased estimator is zero.

Example: Let X_1, X_2, X_3 be a random sample from a population with mean μ .

That is, $E(X_1) = E(X_2) = E(X_3) = \mu$

Consider two alternative point estimators of μ :

$$\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3) \quad \text{and}$$

$$\bar{X}^W = \frac{1}{6}(X_1 + 4X_2 + X_3)$$

The second estimator is a weighted average of the sample information.

To compare these estimators consider:

$$\begin{aligned} E(\bar{X}) &= \frac{1}{3}[E(X_1) + E(X_2) + E(X_3)] \\ &= \frac{1}{3}(3\mu) = \mu \end{aligned}$$

and

$$\begin{aligned} E(\bar{X}^W) &= \frac{1}{6}[E(X_1) + 4E(X_2) + E(X_3)] \\ &= \frac{1}{6}(\mu + 4\mu + \mu) = \mu \end{aligned}$$

Therefore, both estimators are unbiased estimators of the population mean μ .

The above example demonstrated that there may be several unbiased estimators of a population parameter of interest.

A problem that arises is: how can an estimator be selected from among a number of competing unbiased estimators ?

A suggestion is to choose the estimator with minimum variance.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of the population parameter θ . That is,

$$E(\hat{\theta}_1) = \theta \quad \text{and} \quad E(\hat{\theta}_2) = \theta$$

$\hat{\theta}_1$ is said to be **more efficient** than $\hat{\theta}_2$ if:

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$$

The **relative efficiency** of one estimator with respect to another is the variance ratio:

$$\frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}$$

- If $\hat{\theta}$ is an unbiased estimator of θ , and no other unbiased estimator has smaller variance than $\hat{\theta}$, then $\hat{\theta}$ is said to be the **most efficient** or **minimum variance unbiased estimator** of θ .

Example Continued: For the random sample X_1, X_2, X_3 , introduced above, assume the population variance is σ^2 and also assume independence.

That is, $\text{Var}(X_1) = \text{Var}(X_2) = \text{Var}(X_3) = \sigma^2$ and

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_1, X_3) = \text{Cov}(X_2, X_3) = 0$$

Two unbiased estimators for the population mean were proposed as:

$$\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3) \quad \text{and}$$

$$\bar{X}^W = \frac{1}{6}(X_1 + 4X_2 + X_3)$$

The variance of the first estimator is:

$$\begin{aligned} \text{Var}(\bar{X}) &= \frac{1}{9}[\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)] \\ &= \frac{1}{9}(3\sigma^2) = \sigma^2/3 \end{aligned}$$

The variance of the second estimator is:

$$\begin{aligned} \text{Var}(\bar{X}^W) &= \frac{1}{36}[\text{Var}(X_1) + 16\text{Var}(X_2) + \text{Var}(X_3)] \\ &= \frac{1}{36}(18\sigma^2) = \sigma^2/2 \end{aligned}$$

It can be seen that: $\text{Var}(\bar{X}) < \text{Var}(\bar{X}^W)$

Therefore, \bar{X} is more efficient than \bar{X}^W .

The relative efficiency is:

$$\frac{\text{Var}(\bar{X}^W)}{\text{Var}(\bar{X})} = \frac{\sigma^2/2}{\sigma^2/3} = \frac{3}{2} = 1.5$$

Note: A reporting style for the measure of relative efficiency is to place the higher variance in the numerator.

As a variation, suppose that the random sample X_1, X_2, X_3 have probability distributions with identical population mean μ but unequal population variances. Assume:

$$\text{Var}(X_1) = 4\sigma^2$$

$$\text{Var}(X_2) = \sigma^2 \quad \text{and}$$

$$\text{Var}(X_3) = 4\sigma^2$$

The revised variances for the two competing estimators of the population mean are:

$$\begin{aligned}\text{Var}(\bar{X}) &= \frac{1}{9}[\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)] \\ &= \frac{1}{9}(4\sigma^2 + \sigma^2 + 4\sigma^2) \\ &= \sigma^2\end{aligned}$$

$$\begin{aligned}\text{Var}(\bar{X}^W) &= \frac{1}{36}[\text{Var}(X_1) + 16\text{Var}(X_2) + \text{Var}(X_3)] \\ &= \frac{1}{36}(4\sigma^2 + 16\sigma^2 + 4\sigma^2) \\ &= \frac{24}{36}\sigma^2\end{aligned}$$

The results now show: $\text{Var}(\bar{X}^W) < \text{Var}(\bar{X})$

When the random sample have distributions with population variances unequal then the weighted average \bar{X}^W is more efficient than the sample mean \bar{X} as an estimator of the population mean.

Chapter 8.2 Interval Estimation

Let the random sample X_1, X_2, \dots, X_n be a set of independent and identically distributed random variables with mean μ and variance σ^2 .

Consider θ as a population parameter of interest. The true value of this parameter is unknown.

A point estimator of θ can be proposed as: $\hat{\theta} = f(X_1, X_2, \dots, X_n)$

It may also be informative to find random variables $\hat{\theta}_{\text{low}}$ and $\hat{\theta}_{\text{high}}$ such that:

$$P(\hat{\theta}_{\text{low}} < \theta < \hat{\theta}_{\text{high}}) = 0.9$$

$[\hat{\theta}_{\text{low}}, \hat{\theta}_{\text{high}}]$ is called a 90% **confidence interval estimator** for θ .

In general, find random variables $\hat{\theta}_{\text{low}}$ and $\hat{\theta}_{\text{high}}$ such that:

$$P(\hat{\theta}_{\text{low}} < \theta < \hat{\theta}_{\text{high}}) = 1 - \alpha$$

↑

Greek letter alpha

$1 - \alpha$ is called the **confidence level**.

$[\hat{\theta}_{\text{low}}, \hat{\theta}_{\text{high}}]$ gives a $100(1 - \alpha)\%$ confidence interval estimator for θ .

❖ Interval Estimation for the Population Mean

Results established in previous lecture notes are first reviewed.

A point estimator for the population mean μ is:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

\bar{X} has a sampling distribution with the properties:

$$E(\bar{X}) = \mu \quad (\bar{X} \text{ is an unbiased estimator of } \mu) \text{ and}$$

$$\text{Var}(\bar{X}) = \sigma^2 / n$$

The standard error of \bar{X} is: $se(\bar{X}) = \sigma / \sqrt{n}$

To proceed further, assume that \bar{X} follows a normal distribution.

This assumption is reasonable since:

- If X_1, X_2, \dots, X_n follow a normal distribution then a result is that \bar{X} also has a normal distribution.
- Even if X_1, X_2, \dots, X_n are not normally distributed, by the Central Limit Theorem, \bar{X} will tend to the normal distribution.

A standard normal random variable is:

$$Z = \frac{\bar{X} - \mu}{se(\bar{X})} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

A procedure for interval estimation is now developed.

A critical value z_c can be found so that:

$$P(-z_c < Z < z_c) = 1 - \alpha$$

where $1 - \alpha$ is set to a desired level.

Rearrange the probability statement to obtain:

$$\begin{aligned} P(-z_c < Z < z_c) &= P\left(-z_c < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_c\right) \\ &= P\left(\bar{X} - z_c \sigma/\sqrt{n} < \mu < \bar{X} + z_c \sigma/\sqrt{n}\right) \end{aligned}$$

This gives the $100(1 - \alpha)\%$ confidence interval estimator for the population mean μ as:

$$\left[\bar{X} - z_c \sigma/\sqrt{n}, \bar{X} + z_c \sigma/\sqrt{n} \right]$$

This can be written as:

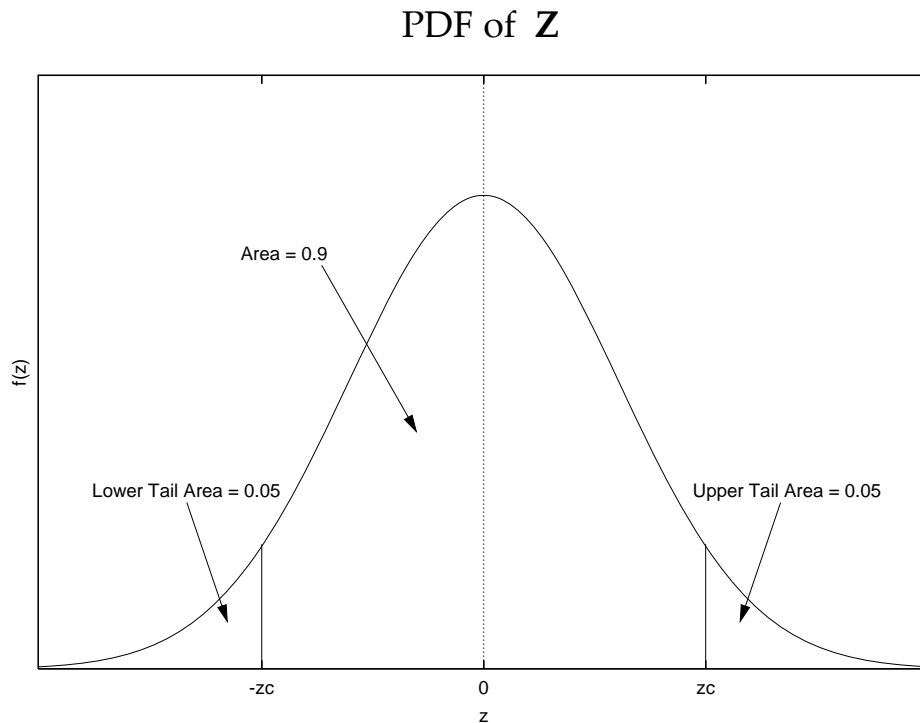
$$\bar{X} \pm z_c \sigma/\sqrt{n}$$

To finish off, the critical value z_c must be set.

For a 90% confidence interval estimator, with $\alpha = 0.10$, find a number z_c such that:

$$P(-z_c < Z < z_c) = 0.90$$

An illustration is below.



Note that the area in each tail is: $\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$

This says find z_c such that:

$$P(Z < z_c) = F(z_c) = 1 - \frac{\alpha}{2} = 0.95$$

Appendix Table 1 can be used to get an answer.

Another option is to use Microsoft Excel.

Select Insert Function: `NORMSINV(0.95)`

This returns the answer: $z_c = 1.645$

The confidence level $1 - \alpha$ can be set to any desired probability.

A table of some popular choices is below.

Confidence level	$\frac{\alpha}{2}$	Microsoft Excel NORMSINV probability	z_c
0.90	0.05	0.95	1.645
0.95	0.025	0.975	1.96
0.99	0.005	0.995	2.576

An application can now proceed.

Collect a data set with numeric observations: x_1, x_2, \dots, x_n .

The point estimate for the population mean μ is the calculated sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Assume that the population standard deviation σ is known from previous research.

A 90% **confidence interval estimate** is calculated as:

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

➤ What is the interpretation of an interval estimate ?

In applied work, the calculated interval estimate is based on one sample of data. It may contain the true parameter μ , or it may not contain μ . Since the true value of μ is unknown, it is impossible to say whether or not the population mean is contained in the interval estimate calculated from the sample of data.

The interpretation of a 90 % confidence interval estimate for the population mean can be explained in the context of repeated sampling.

Different samples of data will give different calculated sample means. Some of the sample means will be less than the true mean μ and some will be greater than this value.

Therefore, different samples will give different interval estimates.

In a 'large' number of samples 90% of the interval estimates will contain the true population mean μ and the other 10% will not contain this value (that is, they either have an upper limit below μ or a lower limit that exceeds μ).

The interpretation of a 90 % confidence interval estimate can be demonstrated with a computer simulation.

An experiment was designed to work as follows:

To start, take a sample of $n = 25$ observations from a population that follows the normal distribution with $\mu = 5$ and $\sigma = 2$. From the sample observations calculate the sample mean and a 90% confidence interval estimate.

Now take another sample of 25 observations and repeat the calculations.

Continue drawing samples from the population.

Stop after interval estimates from 1000 samples have been calculated.

The calculation of the sample mean is based on an unbiased estimation rule. This says that the average of the 1000 calculated sample means should give the true mean 5.

A computer experiment was tried and the results showed the average of the sample means was 5.003 to support the idea of an unbiased estimation rule.

Estimation results for 20 selected samples are given below.

Sample	\bar{x}	90% Confidence Interval Estimate $\bar{x} \pm 1.645(2/\sqrt{25})$	
1	4.94	4.28	5.60
2	5.29	4.63	5.95
3	4.97	4.32	5.63
4	4.15	3.49	4.81 **
5	4.49	3.84	5.15
6	5.06	4.40	5.72
7	5.35	4.69	6.00
8	5.28	4.62	5.93
9	4.83	4.17	5.48
10	5.49	4.83	6.15
11	4.71	4.06	5.37
12	4.71	4.05	5.37
13	5.59	4.94	6.25
14	4.92	4.26	5.58
15	4.75	4.09	5.40
16	5.54	4.89	6.20
17	5.04	4.39	5.70
18	4.72	4.06	5.38
19	4.95	4.29	5.60
20	5.84	5.18	6.50 **

Each interval estimate is centered at the calculated sample mean \bar{x} . All interval estimates have the same width.

The samples marked with ** have interval estimates that do not contain the true population mean $\mu = 5$.

That is, Sample 4 has an upper limit below $\mu = 5$ and Sample 20 has a lower limit that exceeds 5.

It can be seen that the other 18 samples listed all have interval estimates that contain the true population mean 5.

To demonstrate the interpretation of a 90% confidence interval, for the 1000 samples generated for the experiment, about 900 of the calculated interval estimates should contain the true population mean 5 and the remaining interval estimates (about 100) will not contain the true mean (like Sample numbers 4 and 20 in the list printed above).

The computer experiment reported above counted 107 interval estimates that did not contain the population mean $\mu = 5$.

It should be noted that if the experiment was repeated, a different set of 1000 samples would be generated, and therefore the numerical summary of the results would be a bit different.

Now take another look at the calculation for the confidence interval estimate:

$$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}}$$

The width of the interval estimate is: $2 \cdot z_c \frac{\sigma}{\sqrt{n}}$

The width will be affected by:

- the level of α . This sets the value of z_c .
Smaller α leads to a wider confidence interval.
That is, a 99% interval is wider than a 95% interval.
- the variance σ^2 . As σ^2 increases, the confidence interval becomes wider.
- the sample size n . As n increases, the confidence interval becomes narrower.

In general, a wide confidence interval reflects imprecision in the knowledge about the population mean.

Chapter 8.3 Interval Estimation Continued

A 90% confidence interval estimate for the population mean can be calculated as:

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

In practice, the population variance σ^2 is unknown.

With a sample of data, a way to proceed is to calculate a variance estimate as:

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

Then, for the calculation of an interval estimate, replace the unknown σ with the calculated standard deviation s to get the interval estimate for the population mean as:

$$\bar{x} \pm 1.645 \frac{s}{\sqrt{n}}$$

A problem with this is that the confidence level is no longer guaranteed to be 0.90. The interval estimate may now be viewed as an *approximate* 90% interval estimate. Since s is an estimate of the population variance, the critical value 1.645 may be smaller than what will give a correct 90% confidence interval.

It turns out that the quality of the approximation gradually improves with increasing sample size n .

As a rough guideline, with $n > 60$, a good approximation for a 90% confidence interval is given with:

$$\bar{x} \pm 1.645 \frac{s}{\sqrt{n}}$$

Many economic data sets meet this requirement of a sample size exceeding 60 observations.

However, methods are available for the calculation of exact interval estimates. These methods are standard features of computer software designed for the statistical analysis of economic data. Therefore, this is the next topic to discuss.

The problem is to construct a confidence interval for the population mean when the population variance is also unknown.

Some more statistical theory is needed.

For a random sample X_1, X_2, \dots, X_n a familiar result is:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1) \quad (\text{the standard normal random variable})$$

A variance estimator can be stated as:

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Now consider the random variable:

$$t = \frac{\bar{X} - \mu}{s_X / \sqrt{n}}$$

In the denominator, σ is replaced by the estimator s_X .

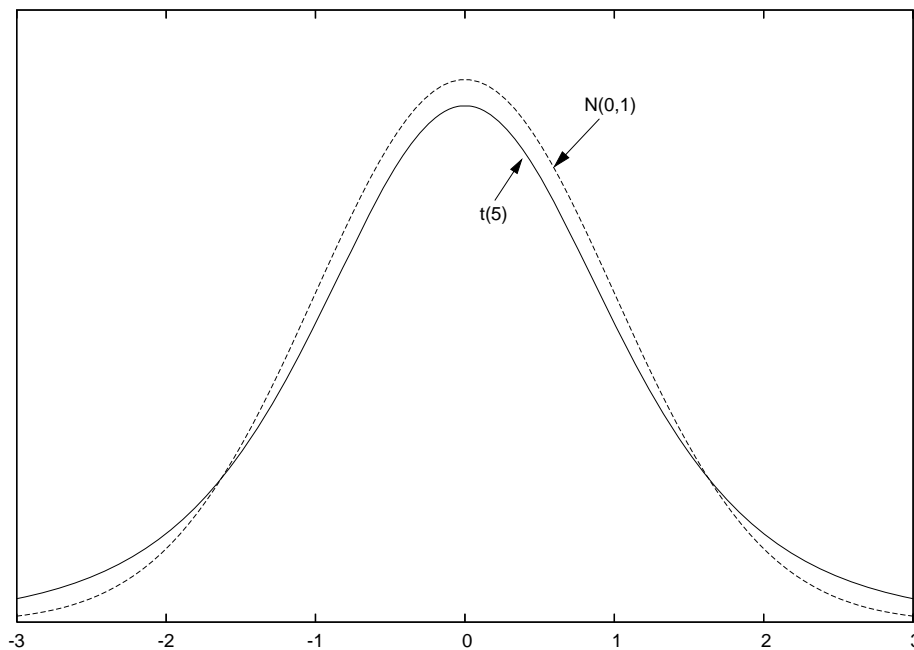
This new random variable has a Student's **t-distribution** with **(n-1) degrees of freedom**.

The degrees of freedom come from the divisor used for the sample variance.

Properties of the probability density function of the t-distribution:

- the shape is determined by the degrees of freedom (**$n-1$**).
- like the standard normal distribution, the shape of the probability density function of the t-distribution is a symmetric curve with mean zero. But the t-distribution has thicker tails compared to the normal distribution.
- as the degrees of freedom increases the t-distribution becomes the same as the standard normal distribution.

The graph below shows a comparison of the probability density function (PDF) of the standard normal distribution and the t-distribution with 5 degrees of freedom.



Note that the t-distribution is less 'peaked' compared to the standard normal distribution.

Let $t_{(m)}$ denote a random variable having a t-distribution with m degrees of freedom. For an upper tail probability $\alpha/2$, a critical value t_c is the number such that:

$$P(t_{(m)} > t_c) = \alpha/2$$

These values are reported in the Appendix Table printed in the inside front cover of the textbook.

(Caution: when reading the tables the numeric value for the upper tail probability must be set correctly).

The t-distribution critical values can also be obtained with Microsoft Excel with the function:

$$TINV(\alpha, \text{degrees_of_freedom})$$

This gives a probability of $\alpha/2$ in each of the upper tail and lower tail.

An application can proceed.

From a data set with n observations calculate the sample mean \bar{x} and sample standard deviation s .

A $100(1 - \alpha)\%$ confidence interval estimate for the population mean is given by:

$$\bar{x} \pm t_c \frac{s}{\sqrt{n}}$$

where t_c is the critical value from the t-distribution with $(n-1)$ degrees of freedom such that:

$$P(t_{(n-1)} > t_c) = \alpha/2$$

Example: Gasoline Consumption of Trucks (Example 8.5 page 292)

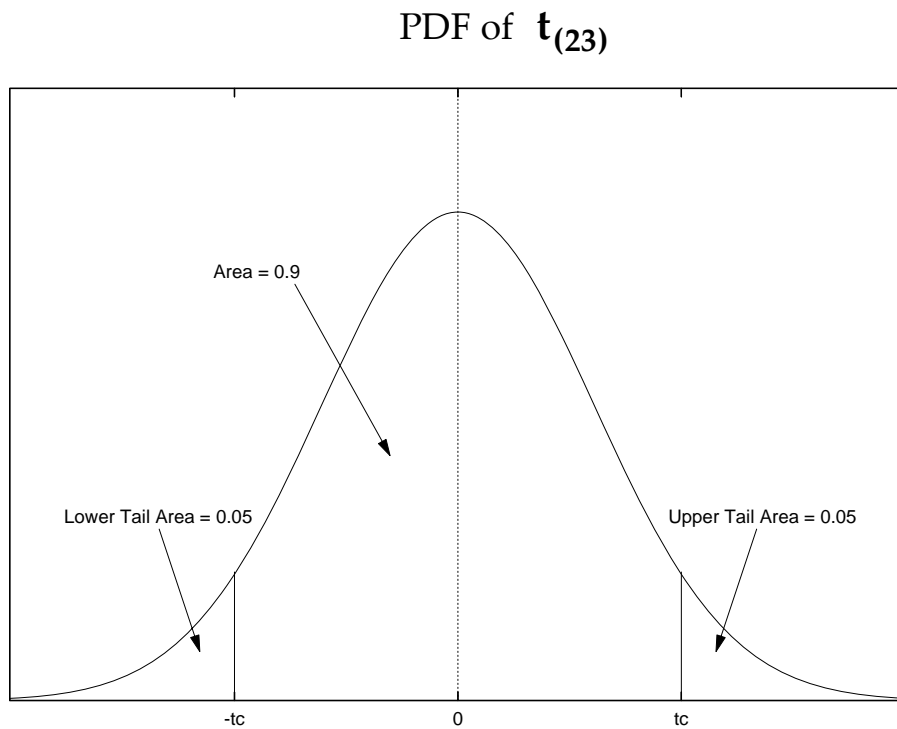
A data set has observations on fuel consumption, in miles per gallon, for 24 trucks. Summary statistics are:

$$\bar{x} = 18.68 \quad \text{and} \quad s = 1.695$$

A 90% confidence interval estimate for the population mean fuel consumption is:

$$18.68 \pm t_c \frac{1.695}{\sqrt{24}}$$

The graph below illustrates the t-distribution critical value.



The Appendix Table for the t-distribution can be used to look-up the critical value t_c . For this example, to correctly use the table, select the degrees of freedom $(n-1) = 23$, and set the upper tail area to $0.10/2 = 0.05$.

Alternatively, use Microsoft Excel. Select Insert Function:

TINV(0.10, 23)

This returns the answer: $t_c = 1.714$

The calculations required for the interval estimate are:

$$18.68 \pm 1.714 \cdot \frac{1.695}{\sqrt{24}}$$

For the given data set, the calculations give a 90% confidence interval estimate for the population mean as:

[18.09 ,19.27]

Chapter 9.1 Confidence Intervals for the Difference Between Two Population Means

Let (X_i, Y_i) for $i = 1, 2, \dots, n$ be a pair of random variables that each follow the normal distribution with population means μ_X and μ_Y .

A data set is collected with the numeric observations:

$$(x_i, y_i) \quad \text{for } i = 1, 2, \dots, n$$

Example: For a survey of households, x_i is the dollar expenditure on fresh fruits and vegetables by household i in a given week and y_i is the dollar expenditure on canned drinks by the same household.

The interest is to develop a confidence interval for the difference in the population means: $\mu_X - \mu_Y$.

An estimation method follows.

From the observed sample calculate the differences:

$$\mathbf{d}_i = \mathbf{x}_i - \mathbf{y}_i \quad \text{for } i = 1, 2, \dots, n$$

Obtain the sample mean and variance of the differences as:

$$\begin{aligned} \bar{\mathbf{d}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{d}_i = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{y}_i) \\ &= \bar{\mathbf{x}} - \bar{\mathbf{y}} \end{aligned}$$

and

$$\begin{aligned} s_{\mathbf{d}}^2 &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{d}_i - \bar{\mathbf{d}})^2 \\ &= s_{\mathbf{x}}^2 + s_{\mathbf{y}}^2 - 2s_{\mathbf{xy}} \end{aligned}$$

where $s_{\mathbf{x}}^2$ and $s_{\mathbf{y}}^2$ are the sample variances from the two variables and $s_{\mathbf{xy}}$ is the sample covariance.

Note that the variance of the differences recognizes the covariance between the two variables.

A $100(1 - \alpha)\%$ confidence interval estimate for the difference in population means $(\mu_X - \mu_Y)$ is given by:

$$\bar{d} \pm t_c \frac{s_d}{\sqrt{n}}$$

where t_c is the critical value from the t-distribution with $(n-1)$ degrees of freedom such that:

$$P(t_{(n-1)} > t_c) = \alpha/2$$

Example: Stock market data for 20 successive business days has been collected. The data set has the observations:

x_1, x_2, \dots, x_n daily percentage returns for a company, and
 y_1, y_2, \dots, y_n daily percentage returns for a market portfolio.

On a given day, stock market prices respond to information about the general economy and therefore, it may be expected that the returns for an individual company may be correlated with the overall performance of the market. That is, a non-zero covariance between the x and y observations is realistic.

The task of interest is to obtain a confidence interval estimate for the difference between the mean return for the company and the mean return for the market portfolio.

The differences are generated as:

$$d_i = x_i - y_i \quad \text{for } i = 1, 2, \dots, 20$$

The sample mean and standard deviation of the differences were calculated as:

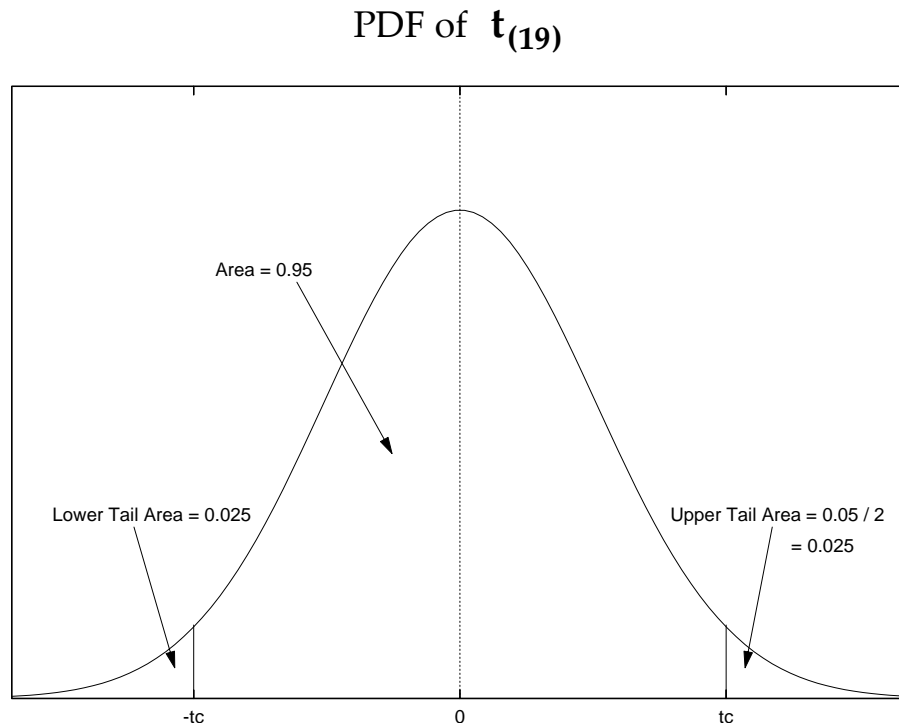
$$\bar{d} = -0.173 \quad \text{and} \quad s_d = 1.391$$

The negative value for \bar{d} says that the sample mean for the company returns is less than the sample mean for the market returns. This does not imply that this is the case for the population – the population means are unknown.

For the purpose of the exercise, set the confidence level to 0.95.
A 95% interval estimate is:

$$\bar{d} \pm t_c \frac{s_d}{\sqrt{n}}$$

An illustration of the t-distribution critical value t_c is below.



To look-up the critical value from the Appendix Table for the t-distribution select the degrees of freedom $(n-1) = 19$, and set the upper tail area to 0.025.

To check the method, with Microsoft Excel select Insert Function:

$$\text{TINV}(0.05, 19)$$

This returns the answer: $t_c = 2.093$

By using the numerical results, the 95% interval estimate for the difference in population mean returns for the company and the market is calculated as:

$$-0.173 \pm 2.093 \cdot \frac{1.391}{\sqrt{20}}$$

This gives the lower and upper limits:

$$[-0.82, 0.48]$$

- Note that the interval contains the value zero (the lower limit is negative and the upper limit is positive). This suggests the possibility that $\mu_X - \mu_Y = 0$ or $\mu_X = \mu_Y$. That is, from the sample of data, there is evidence that the two population means are the same.

Chapter 9.2 More Confidence Intervals for the Difference Between Two Population Means

Another problem of interest is to compare the population means of two *independent* samples.

Consider two independent random samples from normal populations:

- the first sample has n_x observations from a population with mean μ_x .
An estimator of the population mean is the sample mean \bar{X} .
- the second sample has n_y observations from a population with mean μ_y .
An estimator of the population mean is the sample mean \bar{Y} .

Note that the samples can have different sample sizes.

To develop results, assume that the two populations have the same (unknown) variance σ^2 .

The difference between the two sample means $\bar{X} - \bar{Y}$ is a normally distributed random variable with mean $\mu_X - \mu_Y$ and variance:

$$\begin{aligned}\text{Var}(\bar{X} - \bar{Y}) &= \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) \\ &= \frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}\end{aligned}$$

From the numeric data set, the calculated sample means and variances are: \bar{x} , \bar{y} and s_x^2 , s_y^2 .

An estimate of the population variance σ^2 is needed. A method is to **pool** (or combine) the data from the two samples and calculate the **pooled sample variance**:

$$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)}$$

The degrees of freedom associated with the variance calculation is $(n_x + n_y - 2)$.

With this design, a $100(1 - \alpha)\%$ confidence interval estimate for the difference in population means $(\mu_X - \mu_Y)$ is given by:

$$(\bar{x} - \bar{y}) \pm t_c \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

where t_c is the critical value from the t-distribution with $(n_x + n_y - 2)$ degrees of freedom such that:

$$P(t_{(n_x + n_y - 2)} > t_c) = \alpha/2$$

Example: A stock market data base contains daily closing prices for the company Johnson & Johnson for the year 1999. For the first six months (January to June) observations are recorded for 124 trading days. For the period July to December observations are available for 128 trading days.

An exercise of interest is to find a 95% confidence interval estimate for the difference between the population mean closing price in the two sample periods.

Stock market prices adjust rapidly to the arrival of new information. Therefore, it is reasonable to consider that the two samples are independent.

Summary statistics for the closing prices for the two sample periods are:

January to June	$n_x = 124$	$\bar{x} = \$89.96$	$s_x^2 = 32.54$
July to December	$n_y = 128$	$\bar{y} = \$97.98$	$s_y^2 = 21.45$

The pooled variance estimate is: $s^2 = 26.91$

A confidence interval estimate is calculated as:

$$(89.96 - 97.98) \pm t_c \sqrt{\frac{26.91}{124} + \frac{26.91}{128}}$$

A t-distribution critical value t_c for a 95% interval estimate is needed. The degrees of freedom is:

$$(n_x + n_y - 2) = 124 + 128 - 2 = 250$$

The Appendix Table for the t-distribution does not have an entry for 250 degrees of freedom. However, for $\alpha/2 = 0.05/2 = 0.025$,

$$\mathbf{P}(t_{(250)} > 1.96) \cong \mathbf{P}(Z > 1.96) = 0.025$$

where Z is the standard normal random variable.

With Microsoft Excel the Function `TINV(0.05, 250)` returns the answer: $t_c = 1.969$.

Calculations give a 95% interval estimate for the difference in means for the closing prices in the two sample periods as:

$$[-9.31, -6.73]$$

A 99% interval estimate is wider.

With $\alpha = 0.01$ the Microsoft Excel Function $TINV(0.01, 250)$ gives the critical value: $t_c = 2.596$.

For a 99% interval estimate the lower and upper limits are:

$$[-9.72, -6.32]$$

The value zero is outside the range of the calculated interval estimate to suggest that the mean closing price is lower in the first sample period compared to the second sample period.

Chapter 9.5 Sample Size Determination

A wide confidence interval reflects uncertainty about the parameter being estimated. A larger sample size n will give a narrower interval.

Consider a confidence interval for the population mean μ in a situation where the population variance σ^2 is known from previous research.

For a given data set, a $100(1 - \alpha)\%$ interval estimate for the population mean is:

$$\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}} \right]$$

where z_c is the value such that:

$$P(Z < z_c) = F(z_c) = 1 - \frac{\alpha}{2}$$

The width of the interval estimate is: $w = 2 \cdot z_c \frac{\sigma}{\sqrt{n}}$

Suppose a set width w is desired.

What sample size n will guarantee this width ?

Rearranging gives: $\sqrt{n} = 2 \cdot z_c \frac{\sigma}{w}$

By squaring both sides: $n = \left(2 \cdot z_c \frac{\sigma}{w} \right)^2$

Round up to get an integer number for n .

Chapter 10.1 Hypothesis Testing

Interval estimation leads the way to hypothesis testing.

To illustrate the idea – an application may suggest the question:

Is the population mean equal to 5 ?

This can be considered a hypothesis to test.

A statement of the hypothesis must be formed.

The **null hypothesis** is denoted by H_0 (H-naught – pronounced H-not). For example,

$$H_0 : \mu = 5$$

This is tested against the **alternative hypothesis** denoted by H_1 .

The form of the alternative hypothesis may arise from the specific application. Three possible options for the alternative hypothesis are:

$$H_1 : \mu \neq 5 \quad \text{two-sided alternative}$$

$$\left. \begin{array}{l} H_1 : \mu > 5 \\ H_1 : \mu < 5 \end{array} \right\} \text{one-sided alternative}$$

To test the hypothesis, a **test statistic** is computed from the sample data. The decision to take can then be either:

- do **not reject** the null hypothesis, or
- **reject** the null hypothesis in favour of the alternative.

With a sample of data it is always possible that a wrong decision will be made. Two different mistakes are:

- the null hypothesis is true – but the decision is to reject it.
This is called a **Type I error**.
- the null hypothesis is false – but the test does not reject it.
This is called a **Type II error**.

For a test method:

α is the probability of a Type I error, and

β is the probability of a Type II error.

It would be desirable to use a test method that gives a small value for both α and β . But typically, there is some trade-off. By setting a lower value for α this leads to reluctance to reject the null hypothesis and therefore a greater risk of a Type II error and a larger value for β .

For a given level of α , a way to lower β is to increase the sample size n .

How can a decision rule be set ?

A decision rule can be set to give a probability of a Type I error at some fixed level α .

α is called the **significance level** of the test.

Common choices for α are: $\alpha = 0.10, 0.05$ or 0.01 .

Chapter 10.2 Hypothesis Tests of the Mean

Suppose economic theory proposes that the population mean for a variable of interest exceeds the value a .

Does the data support this theory ?

Consider testing the null hypothesis:

$$\mathbf{H_0 : \mu \leq a} \quad \text{or} \quad \mathbf{H_0 : \mu = a}$$

against the alternative hypothesis:

$$\mathbf{H_1 : \mu > a}$$

Note the hypothesis is stated so that if the economic theory is correct the null hypothesis will be rejected. That is, there is strong evidence to support the economic theory.

From a sample of data the calculated sample mean is \bar{x} .

If the sample mean is substantially greater than the value a then the null hypothesis can be rejected.

When the null hypothesis is true, $\mu = a$, and a result is:

$$\frac{\bar{X} - a}{\sigma / \sqrt{n}} \sim N(0, 1)$$

Assume that the population standard deviation σ is known from previous research.

Choose a significance level α . This sets the probability of a Type I error – rejecting a true null hypothesis.

A sensible choice may be $\alpha = 0.05$.

From the sample of data calculate the test statistic:

$$z = \frac{\bar{x} - a}{\sigma / \sqrt{n}}$$

The decision rule is to reject the null hypothesis H_0 if:

$$z > z_c$$

where z_c is the **critical value** that satisfies:

$$P(Z > z_c) = \alpha$$

That is z_c is the value such that the upper tail probability from standard normal distribution is α .

The critical value can be found by inspecting the table for the standard normal distribution given in Appendix Table 1 of the textbook.

Alternatively, with Microsoft Excel, critical values can be found for the common choices of α as follows:

α	Microsoft Excel NORMSINV probability	z_c
0.10	0.90	1.282
0.05	0.95	1.645
0.01	0.99	2.326

Another way to look-up the critical values is to use the Appendix Table for the t-distribution and use the row with the label of ∞ (infinity) for the degrees of freedom.

Example: Evaluating a New Production Process (Example 10.1, page 341 of the textbook).

A manager will switch to a new technology if the production process exceeds 80 units per hour. The manager asks the company statistician to test the null hypothesis:

$$H_0 : \mu \leq 80$$

against the alternative hypothesis:

$$H_1 : \mu > 80$$

If there is strong evidence to reject the null hypothesis then the new technology will be adopted.

Past experience has shown that $\sigma = 8$.

A data set with $n = 25$ for the new technology has a sample mean of:

$$\bar{x} = 83$$

Does this justify adoption of the new technology ?

The calculated test statistic is:

$$z = \frac{83 - 80}{\frac{8}{\sqrt{25}}} = 1.875$$

This can be compared with a critical value – see the table of critical values given earlier.

With a significance level of $\alpha = 0.05$ (5%) it can be seen that:

$$z = 1.875 > z_c = 1.645$$

Therefore, at a 5% level of significance, the conclusion is to reject the null hypothesis. There is evidence that the new technology results in a statistically significant increase in productivity.

Switching to the new technology may be expensive.

By choosing a significance level of $\alpha = 0.01$ (1%) it is now revealed that:

$$z = 1.875 < z_c = 2.326$$

This leads to the conclusion that, at a 1% level of significance, the null hypothesis is **not** rejected. Possibly the company should stay with the current technology.

A question that arises from the above example is:
What is the smallest significance level at which the null hypothesis H_0 can be rejected ?

For the example, this must be something more than 0.01 (since the null hypothesis was not rejected at this level) but less than 0.05 (since the null hypothesis was rejected at this level).

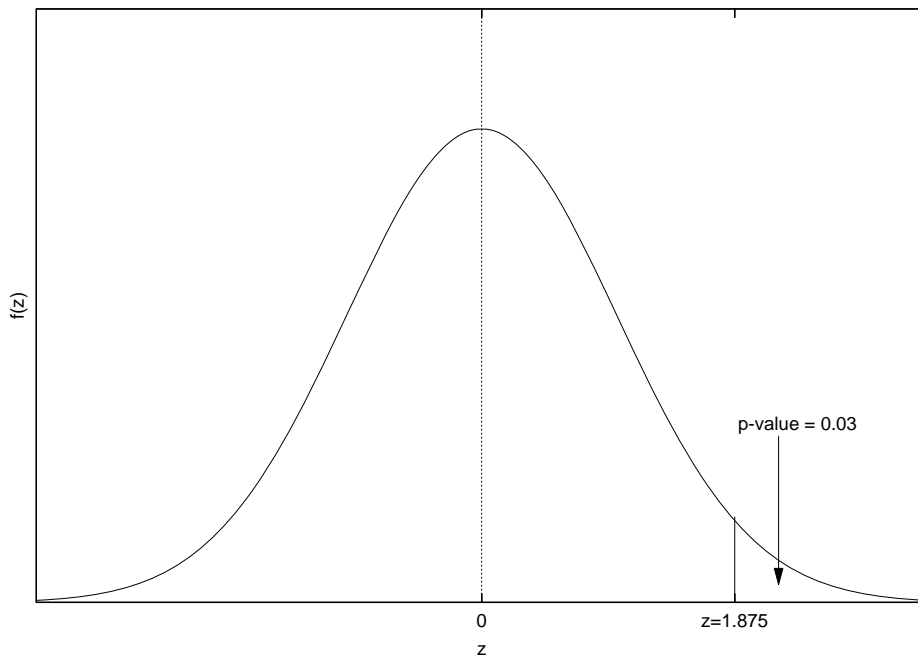
The answer is found as:

$$\begin{aligned} & \text{calculated test statistic} \\ & \quad \downarrow \\ \mathbf{P(Z > 1.875)} &= \mathbf{1 - P(Z < 1.875)} \\ &= \mathbf{1 - F(1.875)} \\ &= \mathbf{1 - 0.9693} \quad \text{look - up in Appendix Table 1} \\ &= \mathbf{0.03} \end{aligned}$$

This probability is called the **p-value** of the test.

For the example, the calculation of the p-value of the test can be illustrated with a graph.

$$\text{PDF of } \frac{\bar{X} - a}{\sigma/\sqrt{n}} = \frac{\bar{X} - 80}{\sigma/\sqrt{n}}$$



With Microsoft Excel the p-value is calculated by selecting Insert Function NORMSDIST(1.875)

This returns the probability 0.9696.

The p-value for the test is $1 - 0.9696 = 0.0304$.

The p-value gives useful information.

For a hypothesis testing application, the computer software can be used to:

- calculate summary statistics for the data set.
- calculate a test statistic for hypothesis testing.
- calculate a p-value for the test.

For a chosen significance level α , the decision rule is to reject the null hypothesis if:

$$\text{p-value} < \alpha$$

This rule gives the same conclusion as presented above.

That is, when $\text{p-value} < \alpha$ the calculated test statistic must be in the rejection region for the test.

Chapter 10.3 Hypothesis Tests of the Mean Continued

The discussion has considered testing the null hypothesis:

$$\mathbf{H_0 : \mu \leq a} \quad \text{or} \quad \mathbf{H_0 : \mu = a}$$

against the one-sided alternative hypothesis: $\mathbf{H_1 : \mu > a}$

The proposal for the test statistic was:

$$\mathbf{z = \frac{\bar{x} - a}{\sigma / \sqrt{n}}}$$

In applied work, replace the unknown population parameter σ with the sample standard deviation s to get the t-test statistic:

$$\mathbf{t = \frac{\bar{x} - a}{s / \sqrt{n}}}$$

For a significance level α , the decision rule is to reject the null hypothesis $\mathbf{H_0}$ if:

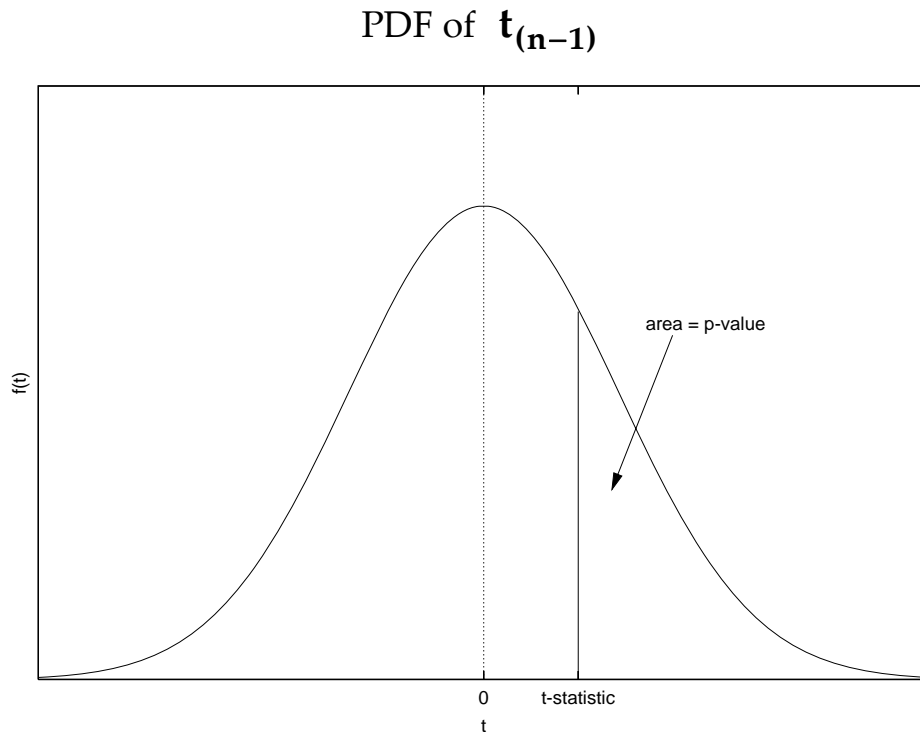
$$\mathbf{t > t_c}$$

where $\mathbf{t_c}$ is the critical value that satisfies:

$$\mathbf{P(t_{(n-1)} > t_c) = \alpha}$$

$\mathbf{t_{(n-1)}}$ is a random variable that has a t-distribution with $(n - 1)$ degrees of freedom. The critical value can be found from the Appendix Table for the t-distribution.

The calculation of a p-value for the test is demonstrated in the graph.



To find a p-value the Appendix Table of the t-distribution does not give enough detail to get any accuracy.

Therefore, the calculation of an exact p-value requires special purpose statistical computing.

The p-value illustrated above can be calculated with Microsoft Excel with the function:

$$\text{TDIST}(t\text{-statistic, degrees_of_freedom, } 1)$$

↑
one-sided alternative

Other forms of the alternative hypothesis have meaningful application.

A problem may suggest testing the null hypothesis:

$$\mathbf{H_0 : \mu \geq a} \quad \text{or} \quad \mathbf{H_0 : \mu = a}$$

against the alternative hypothesis:

$$\mathbf{H_1 : \mu < a}$$

The test procedure uses the same test statistic as given in previous discussion – but the calculation of the p-value is different.

From the data set, calculate the sample mean \bar{x} and sample variance s^2 . From these results, calculate the t-test statistic:

$$\mathbf{t = \frac{\bar{x} - a}{s / \sqrt{n}}}$$

For a significance level α , the decision rule is to reject the null hypothesis H_0 if:

$$t < -t_c$$

where t_c is the critical value that satisfies:

$$P(t_{(n-1)} > t_c) = \alpha$$

The critical value can be found from the Appendix Table for the t-distribution using $(n-1)$ degrees of freedom.

That is, the null hypothesis is rejected for t-test statistic values that are suitably large negative numbers.

A p-value for the test, that gives the smallest significance level at which the null hypothesis can be rejected, is calculated as:

$$\text{p-value} = P(t_{(n-1)} < t)$$

↑
calculated t-test statistic

Note that the calculation of the p-value uses the probability from the t-distribution with $(n-1)$ degrees of freedom that is the area under the probability density function to the *left* of the t-test statistic.

For a negative test statistic this will be the *lower* tail probability.

The null hypothesis is rejected if:

$$\text{p-value} < \alpha$$

Example: Exercise 10.23, page 352 of the textbook

A company selling licenses for a franchise operation claims that, in the first year, the yield on an initial investment is 10%.

Test the company's claim.

How should a hypothesis test be stated ?

If there is strong evidence that the mean return on the investment is below 10% this will give a cautionary warning to a potential investor.

Therefore, test the null hypothesis: $H_0 : \mu = 10$

against the alternative hypothesis: $H_1 : \mu < 10$

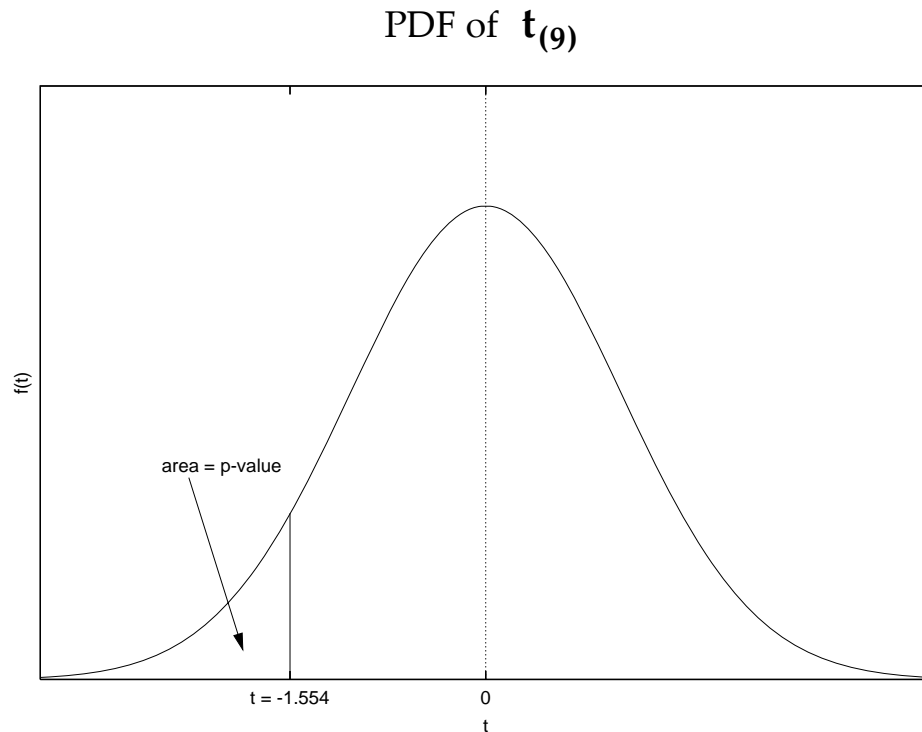
From a sample of $n = 10$ observations, the sample statistics are:

$$\bar{x} = 8.82 \quad \text{and} \quad s = 2.40$$

The t-test statistic is:

$$t = \frac{\bar{x} - a}{\frac{s}{\sqrt{n}}} = \frac{8.82 - 10}{2.40/\sqrt{10}} = -1.554$$

The calculation of the p-value for the test is shown in the graph.



Because of symmetry about zero, the p-value will be identical to the upper tail area, of the probability density function for the t-distribution with $(n-1) = 9$ degrees of freedom, to the right of the value of 1.554.

With Microsoft Excel the probability is calculated with the function:

$$\text{TDIST}(1.554, 9, 1)$$

This returns the answer: p-value = 0.077.

With a significance level of $\alpha = 0.05$, since $p\text{-value} > \alpha$, the null hypothesis is **not** rejected. The data supports the company's claim for the level of return on the investment.

A cautious investor may choose a significance level of $\alpha = 0.10$. Now, $p\text{-value} < \alpha$, and the decision is to reject the null hypothesis. That is, there is some evidence that the company's claim may be over-optimistic.

These conclusions can be confirmed by inspecting the t-distribution values listed in the Appendix Table.

With $(n-1) = 9$ degrees of freedom:

- the 5% critical value is: $-t_c = -1.833$, and
- the 10% critical value is: $-t_c = -1.383$

The calculated t-test statistic of $t = -1.554$ lies between these two values to suggest a p-value somewhere between 5% and 10% – the exact p-value was calculated above.

❖ A Two-Sided Alternative or a Two-Tailed Hypothesis Test

Consider a test of the null hypothesis:

$$\mathbf{H_0 : \mu = a}$$

against the alternative hypothesis:

$$\mathbf{H_1 : \mu \neq a}$$

With a given data set, for a chosen significance level α , the decision rule is to reject the null hypothesis in favour of the alternative if the test statistic:

$$\mathbf{t = \frac{\bar{x} - a}{\frac{s}{\sqrt{n}}}}$$

is such that: $\mathbf{t > t_c}$ or $\mathbf{t < -t_c}$

where $\mathbf{t_c}$ is the critical value that satisfies:

$$\mathbf{P(t_{(n-1)} > t_c) = \alpha/2}$$

That is, the null hypothesis is rejected if: $\mathbf{|t| > t_c}$
↑
absolute value of the t-statistic

A p-value for the test can be calculated as:

$$\text{p-value} = 2 \cdot \mathbf{P(t_{(n-1)} > |t|)}$$

- The p-value for the two-tailed test is double the p-value from a one-tailed test.

Example: Exercise 10.24, page 352 of the textbook

A data set has 9 observations on weight (in ounces) of bottles of shampoo.

The manufacturing process should give a weight of 20 ounces.

Test the null hypothesis:

$H_0: \mu = 20$ the process is operating correctly

against the alternative:

$H_1: \mu \neq 20$ the process is not operating correctly

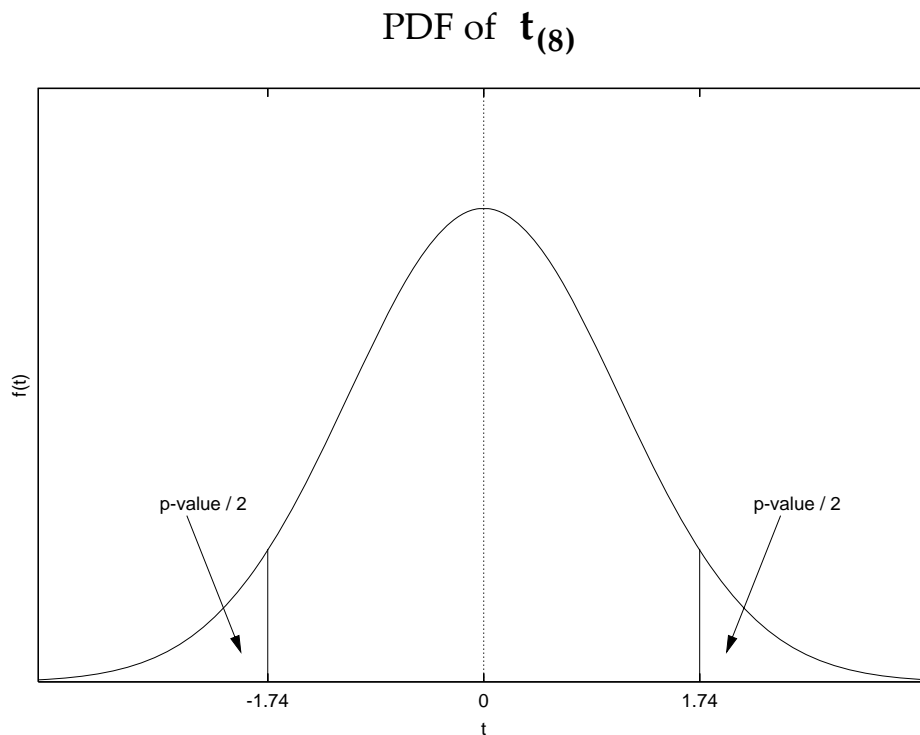
From the data set, the sample statistics are:

$n = 9$, $\bar{x} = 20.356$ (ounces) and $s = 0.6126$

The t-test statistic is:

$$t = \frac{\bar{x} - a}{\frac{s}{\sqrt{n}}} = \frac{20.356 - 20}{0.6126/\sqrt{9}} = 1.74$$

The calculation of a p-value for this 2-sided test is shown in the graph.



With Microsoft Excel find this probability with the function:

TDIST(1.74 , 8, 2)



Two-tailed test

This returns the answer: p-value = 0.12.

Note that the p-value is the sum of the upper and lower tail probabilities. This is equivalent to saying that the p-value is double the upper tail probability.

With a significance level of $\alpha = 0.05$, since the calculated p-value exceeds the chosen significance level, the null hypothesis is not rejected. The evidence in the data set is that the bottles of shampoo are being packaged to a correct standard.

The same conclusion is found by comparing the t-test statistic to the t-distribution critical value.

In the Appendix Table for the t-distribution, with $(n-1) = 8$ degrees of freedom, the value that gives an upper tail area of $\alpha/2 = 0.025$ is:

$$t_c = 2.306$$

The calculated t-test statistic of $t = 1.74$ is less than this critical value and so the null hypothesis is not rejected.

- For the two-tailed hypothesis test, there is a connection to interval estimation. For a test of the null hypothesis $\mathbf{H_0 : \mu = a}$ against a two-sided alternative, the null hypothesis is rejected if the $100(1 - \alpha)\%$ confidence interval estimate for μ does not contain the number \mathbf{a} .

For the shampoo example, a 95% confidence interval estimate for the population mean is calculated as:

$$\bar{x} \pm t_c \frac{s}{\sqrt{n}}$$

Calculations give the lower and upper limits:

$$[19.88, 20.83]$$

The value 20 falls between the lower and upper limits of the 95% confidence interval estimate and, therefore, the null hypothesis that the true population mean is 20 is not rejected when testing against a two-sided alternative.

❖ Summary of hypothesis tests for the population mean

Consider testing the null hypothesis: $\mathbf{H_0 : \mu = a}$

From the data set calculate the test statistic:

$$t = \frac{\bar{x} - a}{s / \sqrt{n}}$$

A p-value for the test can be calculated as follows:

Type of hypothesis	$\mathbf{H_1}$	p-value
One-tail	$\mu > a$	$\mathbf{P(t_{(n-1)} > t)}$
One-tail	$\mu < a$	$\mathbf{P(t_{(n-1)} < t)}$
Two-tail	$\mu \neq a$	$\mathbf{2 \cdot P(t_{(n-1)} > t)}$

For a chosen significance level α , the decision rule is to reject the null hypothesis if:

$$\text{p-value} < \alpha$$

An equivalent test procedure is to use the Appendix Table for the t-distribution to look-up a critical value t_c .

A summary is below.

H_1	t_c	Reject H_0 if:
$\mu > a$	$P(t_{(n-1)} > t_c) = \alpha$	$t > t_c$
$\mu < a$	$P(t_{(n-1)} > t_c) = \alpha$	$t < -t_c$
$\mu \neq a$	$P(t_{(n-1)} > t_c) = \alpha/2$	$ t > t_c$

An additional note is that for 'large' n (say $n > 60$) the random variable $t_{(n-1)}$ can be replaced by the standard normal random variable Z for the purpose of calculating p-values and finding critical values. This comment is of interest if you were stranded without computer software for the calculation of t-distribution p-values but you had access to a table for the standard normal distribution.

Chapter 10.5 Measuring the Power of a Test

An economic problem motivates the statement of a null and alternative hypothesis.

For a numeric data set, a decision rule can lead to the rejection of the null hypothesis. This involves the risk of making an error:

- Type I Error the rejection of a true null hypothesis,
- Type II Error the failure to reject a null hypothesis when the alternative is true.

An approach to hypothesis testing is:

§ choose a significance level α . This sets the probability of a Type I error.

§ establish a decision rule.
This is determined by the significance level chosen for the test.

§ the probability of a Type II error, β , follows.

The **power** of the test is $1 - \beta$.

This is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_1 is true.

How can β , the probability of a Type II error, be determined ?

This will be demonstrated for tests of the mean of a normal population when the population variance is known.

With this set-up, the Appendix Table for the standard normal distribution can be used to look-up required probabilities.

The ideas can be applied to any other hypothesis testing application – but computer software must be used to find probabilities.

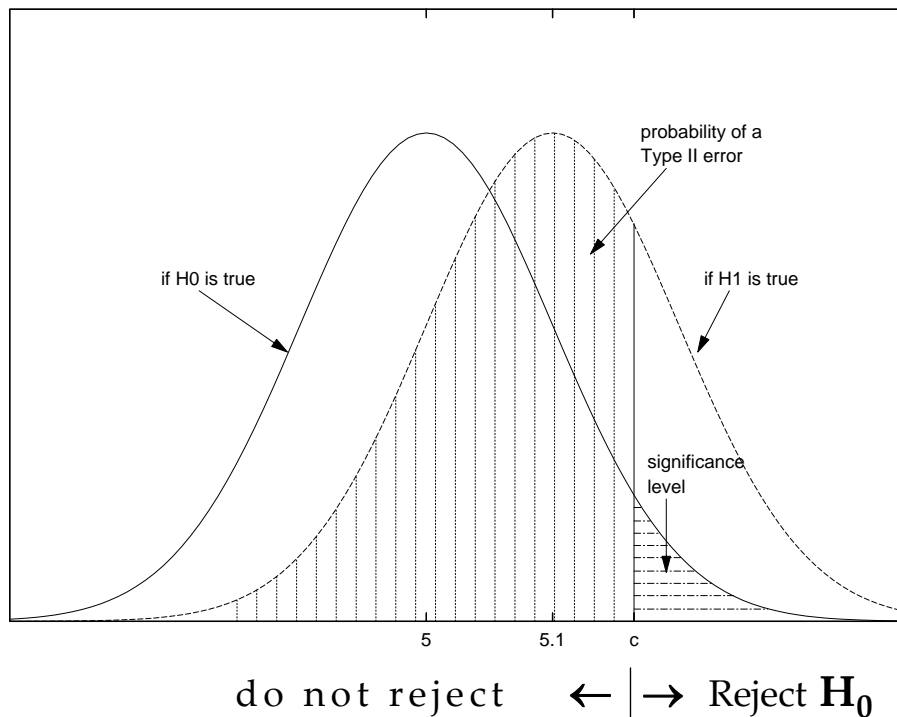
The probability of a Type II error can be illustrated with a picture.

Consider testing $H_0 : \mu = 5$

against the one-sided alternative $H_1 : \mu > 5$

Suppose the true population mean is $\mu = 5.1$.

PDF of \bar{X} with $\mu = 5$ and $\mu = 5.1$



The probability of a Type II error is the probability that the sample mean is below the critical value c when the true population mean is 5.1.

The probability of a Type II error depends on the true value of the population parameter – in this case, the population mean.

In practice, the true value of the population mean is unknown.

From inspection of the above graph, some general results can be stated. By keeping the sample size and population variance the same then:

- the probabilities of a Type I error and Type II error are inversely related.
Lowering the significance level of a test (the probability of a Type I error) increases the chance of a Type II error.
- the closer the true value of the population mean μ to the value stated in the null hypothesis, the greater the probability of a Type II error and the lower the power ($1 - \beta$) of the test.
That is, it is more difficult to detect differences between the null and alternative hypotheses.
For example, in the above graph, the probability density function for the sample mean when the true population mean is $\mu = 5.05$ is shifted to the left compared to the PDF when the mean is $\mu = 5.1$. This increases the the probability of a Type II error.

Example: Exercise 10.39, page 361.

Assume the population standard deviation is: $\sigma = 3$.

From a sample of $n = 9$ the calculated sample mean is: $\bar{x} = 48.2$.

- (a) At a 10% significance level, test
 $H_0: \mu \geq 50$ against $H_1: \mu < 50$

The test statistic is:

$$z = \frac{\bar{x} - 50}{\sigma/\sqrt{n}} = \frac{48.2 - 50}{3/\sqrt{9}} = -1.8$$

The 10% critical value is $z_c = 1.282$

(since the population variance is viewed as known, the critical value is found in the t-distribution Appendix Table in the final row for degrees of freedom ∞).

The results show $z < -z_c$ to give evidence to reject the null hypothesis.

(b) Find the power of a 10% level test when the true mean is 49.

The null hypothesis is rejected for:

$$\frac{\bar{x} - 50}{\sigma/\sqrt{n}} = \frac{\bar{x} - 50}{3/\sqrt{9}} < -1.282$$

or, by rearranging, the null hypothesis is rejected for:

$$\bar{x} < 50 - 1.282 = 48.718$$

The probability of a Type II error is the probability that the sample mean is *not* in the rejection region when the true mean is 49. This is stated as:

$$\beta = P(\bar{X} > 48.718 \mid \mu = 49)$$

This is found as:

$$\begin{aligned}\beta &= P\left(\frac{\bar{X} - 49}{\sigma/\sqrt{n}} > \frac{48.718 - 49}{1}\right) \\ &= P(Z > -0.282) \\ &= P(Z < 0.282) \\ &= \mathbf{0.61} \quad \text{look - up in Appendix Table 1}\end{aligned}$$

The power of the test is:

$$1 - \beta = 1 - 0.61 = \mathbf{0.39}$$

Example: Exercise 10.45, page 361.

Test $H_0 : \mu \geq 32$ against $H_1 : \mu < 32$

Assume the population standard deviation is: $\sigma = 3$.

The decision rule adopted is to reject the null hypothesis in favour of the alternative if the calculated sample mean is such that:

$$\bar{x} < 30.8$$

- (a) With a sample size of $n = 36$, what is the probability of a Type I error, using this decision rule ?

$$\alpha = P(\bar{X} < 30.8 \mid \mu = 32)$$

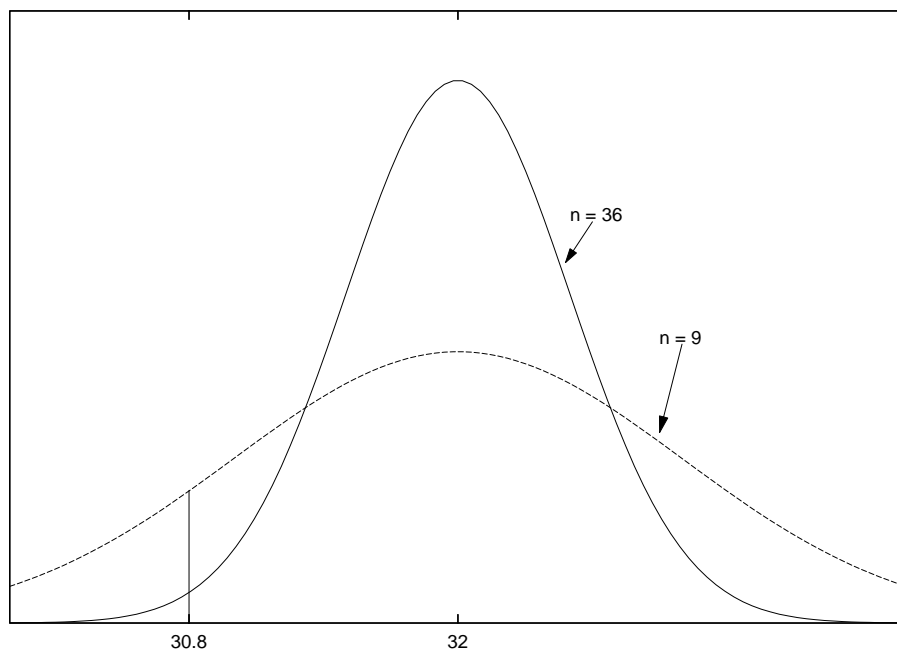
This is the probability that the sample mean is in the rejection region when the null hypothesis is true (the true mean is 32). This is found as:

$$\begin{aligned}\alpha &= P\left(\frac{\bar{X} - 32}{\sigma/\sqrt{n}} < \frac{30.8 - 32}{3/\sqrt{36}}\right) \\ &= P\left(Z < \frac{-1.2}{3/6}\right) \\ &= P(Z < -2.4) \\ &= 1 - P(Z < 2.4) \\ &= 1 - 0.9918 \quad \text{look - up in Appendix Table 1} \\ &= 0.0082\end{aligned}$$

(b) With a sample size of $n = 9$, what is the probability of a Type I error, using this decision rule ?

The calculation is illustrated in the graph below.

PDF of \bar{X} with $\mu = 32$, $\sigma = 3$



Reject $H_0 \leftarrow \mid \rightarrow$ do not reject

With $n = 9$ the probability that the sample mean is in the rejection region (the lower tail) is greater than for the same test based on $n = 36$.

That is, the probability of a Type I error is greater for a smaller sample size.

A numerical answer for the probability can be found by following the calculation steps shown in part (a).

(c) Suppose the true mean is $\mu = 31$.

With a sample size of $n = 36$, what is the probability of a Type II error, using this decision rule ?

$$\beta = P(\bar{X} > 30.8 \mid \mu = 31)$$

This is the probability that the sample mean is *not* in the rejection region when the true mean is 31. This is found as:

$$\begin{aligned}\beta &= P\left(\frac{\bar{X} - 31}{\sigma/\sqrt{n}} > \frac{30.8 - 31}{3/\sqrt{36}}\right) \\ &= P(Z > -0.4) \\ &= P(Z < 0.4) \\ &= 0.6554 \quad \text{look - up in Appendix Table 1}\end{aligned}$$

For a population mean of 31, the power of the test is:

$$1 - \beta = 1 - 0.6554 = 0.3446$$

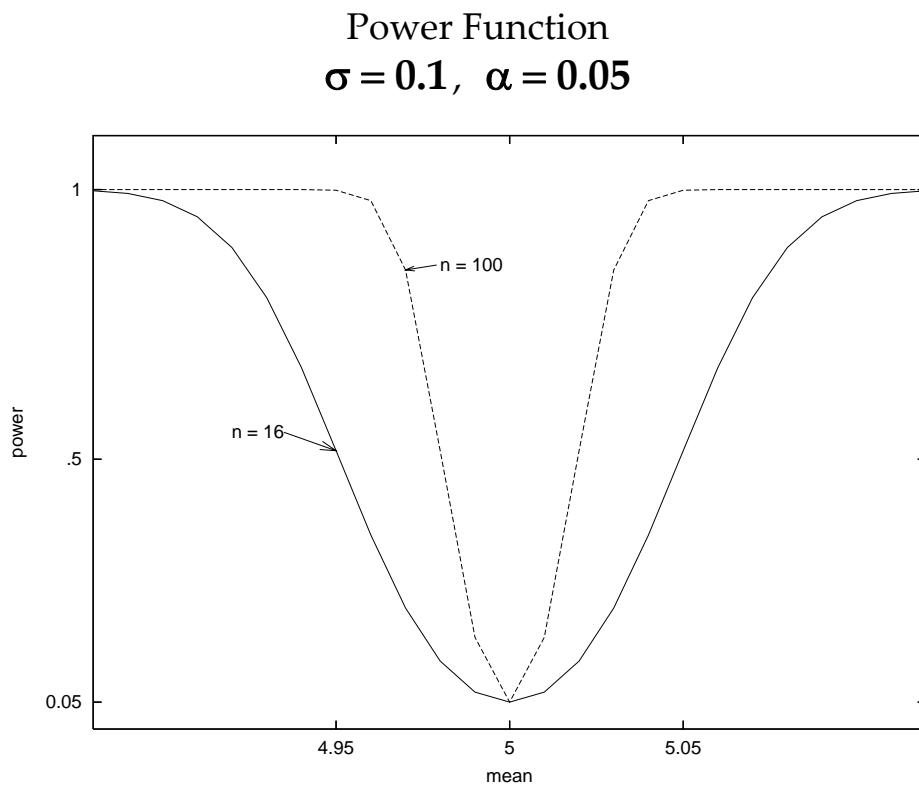
This probability can be calculated for any value of μ (the true mean). The value of β and the power will be different for different values of μ .

Now consider testing the null hypothesis: $H_0 : \mu = 5$

against the two-sided alternative: $H_1 : \mu \neq 5$

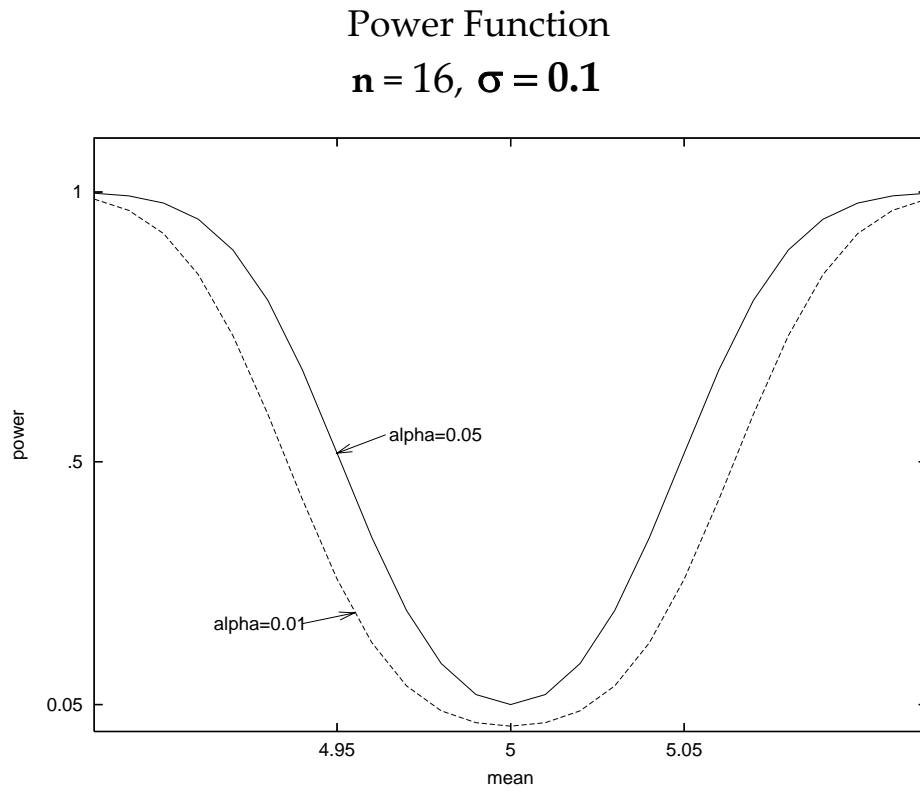
For a given decision rule, the probability of a Type II error can be calculated for different values of the true population mean.

The graph below shows the power function for the two-tailed test with a sample size of $n = 16$ and $n = 100$.



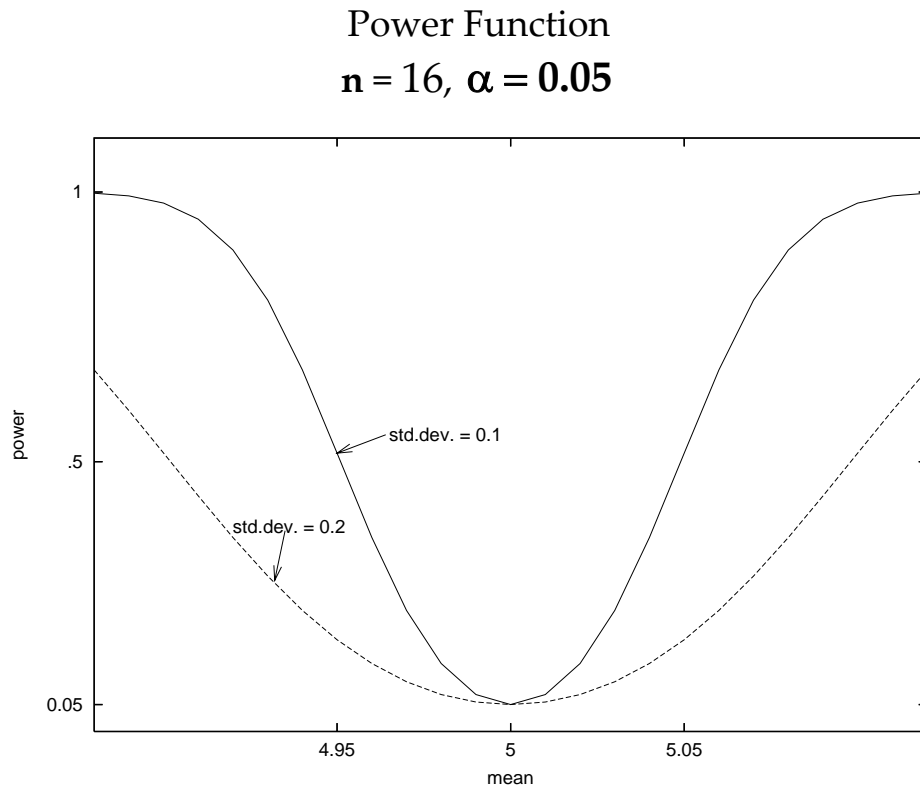
The figure illustrates that an increase in sample size leads to greater power. The figure also shows that the farther the true mean from the hypothesized value of 5 the greater the power of the test. When the true population mean is 5 the probability that the null hypothesis is rejected is 0.05, the significance level of the test.

The next graph shows the power function for significance levels of $\alpha = 0.05$ and $\alpha = 0.01$.



The figure illustrates that a smaller significance level gives lower power.

The next graph shows the power function for standard deviations of $\sigma = 0.1$ and $\sigma = 0.2$.



The figure illustrates that a larger variance gives lower power.

The above discussion has highlighted that the power of any test will depend on:

- the true population parameter
 - the sample size
 - the significance level
 - the population variance
- Failure to reject a false null hypothesis may simply reflect poor quality of the sample information. Small sample size n or high variance leads to low power of a test.

Chapter 11.1 Tests of the Difference Between Two Means

Interval estimation for the difference between two population means was presented in the lecture notes for Chapters 9.1 and 9.2.

An extension to hypothesis testing is of interest for applied work.

For two random samples from populations with means μ_X and μ_Y consider testing the null hypothesis:

$$\mathbf{H_0 : \mu_X = \mu_Y} \quad \text{population means are equal}$$

against the two-sided alternative hypothesis:

$$\mathbf{H_1 : \mu_X \neq \mu_Y} \quad \text{population means are not equal}$$

An equivalent way of expressing the problem is test:

$$\mathbf{H_0 : \mu_X - \mu_Y = 0} \quad \text{against} \quad \mathbf{H_1 : \mu_X - \mu_Y \neq 0}$$

A one-sided alternative hypothesis can also be entertained.

The test method depends on the particular data set.

Separate cases are:

- 'matched pairs' . The numeric observations are:
 $(\mathbf{x}_i, \mathbf{y}_i)$ for $i = 1, 2, \dots, n$.
- independent samples with sample sizes \mathbf{n}_x and \mathbf{n}_y .
That is, the two samples can have different sample sizes.

First, develop results for the matched pairs data set.

Descriptive statistics are denoted by:

$$\begin{aligned} \bar{\mathbf{x}}, \quad \bar{\mathbf{y}} & \quad \text{sample means,} \\ \mathbf{s}_x^2, \quad \mathbf{s}_y^2 & \quad \text{sample variances, and} \\ \mathbf{s}_{xy} & \quad \text{sample covariance.} \end{aligned}$$

From the observations calculate the differences:

$$\mathbf{d}_i = \mathbf{x}_i - \mathbf{y}_i \quad \text{for } i = 1, 2, \dots, n$$

Following the discussion in the lecture notes for Chapter 9.1, the sample mean and variance of the differences can be calculated as:

$$\bar{\mathbf{d}} = \bar{\mathbf{x}} - \bar{\mathbf{y}} \quad \text{and} \quad \mathbf{s}_d^2 = \mathbf{s}_x^2 + \mathbf{s}_y^2 - 2\mathbf{s}_{xy}$$

Note the role of the covariance term in the variance calculation. With positive covariance, the variance of the differences will be reduced compared to using independent samples.

For a comparison of the two population means, for some value \mathbf{a} , test the null hypothesis:

$$\mathbf{H_0 : \mu_X - \mu_Y = a}$$

against a two-sided alternative.

The test statistic is:

$$\mathbf{t = \frac{\bar{d} - a}{s_d / \sqrt{n}}}$$

With the assumption of normal population distributions for the two sample means, the test statistic can be compared with a t-distribution with $(n-1)$ degrees of freedom.

An interesting application is testing for equal population means. That is, $\mathbf{a = 0}$ and the null hypothesis is:

$$\mathbf{H_0 : \mu_X - \mu_Y = 0}$$

For this test, the test statistic is:

$$\mathbf{t = \frac{\bar{d}}{s_d / \sqrt{n}}}$$

Choose a significance level α (the probability of a Type I error).

When testing against a two-sided alternative, a decision rule can be set by one of three equivalent methods.

- (1) Use the Appendix Table of the t-distribution to find a critical value t_c that satisfies:

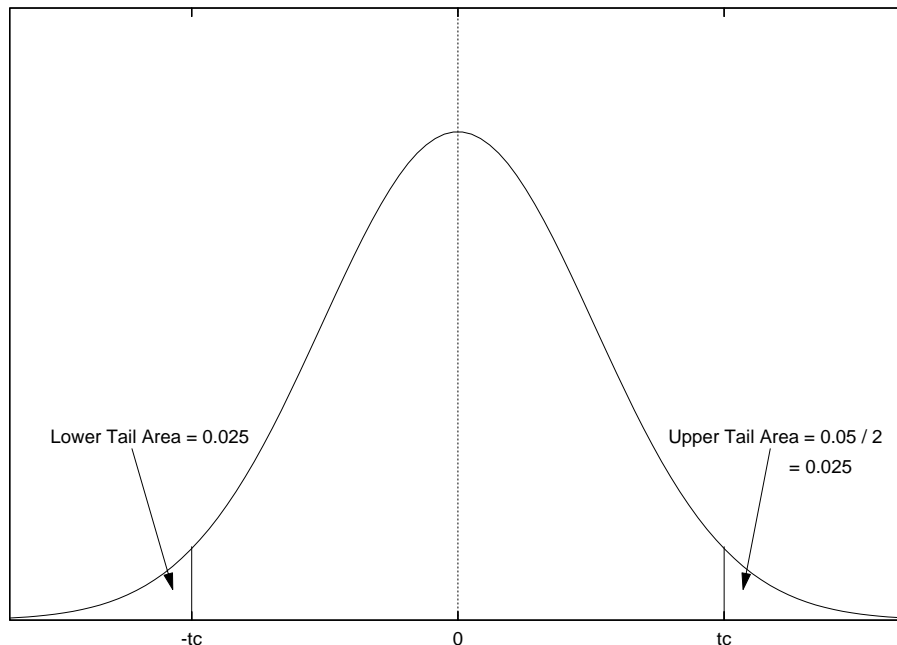
$$P(t_{(n-1)} > t_c) = \alpha/2$$

Reject the null hypothesis if:

$$|t| > t_c$$

For $\alpha = 0.05$ (a 5% significance level), the graph below shows the critical value and the rejection region.

PDF of $t_{(n-1)}$



Reject H_0 \leftarrow $|$ \rightarrow do not reject \leftarrow $|$ \rightarrow Reject H_0

(2) Calculate a p-value for the test as:

$$\text{p-value} = 2 \cdot \mathbf{P}(t_{(n-1)} > |t|)$$

The decision rule is reject the null hypothesis if:

$$\text{p-value} < \alpha$$

The p-value must be calculated with computer software.

(3) Calculate a $100(1 - \alpha)\%$ confidence interval estimate for the difference in population means $\mu_X - \mu_Y$.

The method was presented in Chapter 9.1.

The decision rule is to reject the null hypothesis of equal populations means if the value zero is not contained between the lower and upper limits of the confidence interval estimate.

Example: The stock market data set introduced in Chapter 9.1 contained observations for 20 successive business days described by:

x_1, x_2, \dots, x_n daily percentage returns for a company, and

y_1, y_2, \dots, y_n daily percentage returns for a market portfolio.

The differences are generated as:

$$d_i = x_i - y_i \quad \text{for } i = 1, 2, \dots, 20$$

The sample mean and standard deviation of the differences were calculated as:

$$\bar{d} = -0.173 \quad \text{and} \quad s_d = 1.391$$

To test the null hypothesis of equal population means for the two samples against a two-sided alternative the t-test statistic is:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{-0.173}{1.391 / \sqrt{20}} = -0.556$$

With Microsoft Excel, the p-value calculation $2 \cdot \mathbf{P}(t_{(n-1)} > |t|)$ is obtained by selecting Insert Function:

TDIST(0.556, 19, 2)

This returns a p-value of 0.58.

That is, for the t-distribution with 19 degrees of freedom, the area under the probability density function to the right of the value 0.556 is $0.58/2 = 0.29$, and, by symmetry, the area to the left of the value -0.556 is also 0.29.

The calculated p-value of 0.58 exceeds any reasonable significance level and therefore, with the given data set, there is no evidence to reject the null hypothesis of equal population means for the company returns and the overall market returns.

The same conclusion is obtained by comparing the test statistic with a t-distribution critical value. A significance level must be chosen.

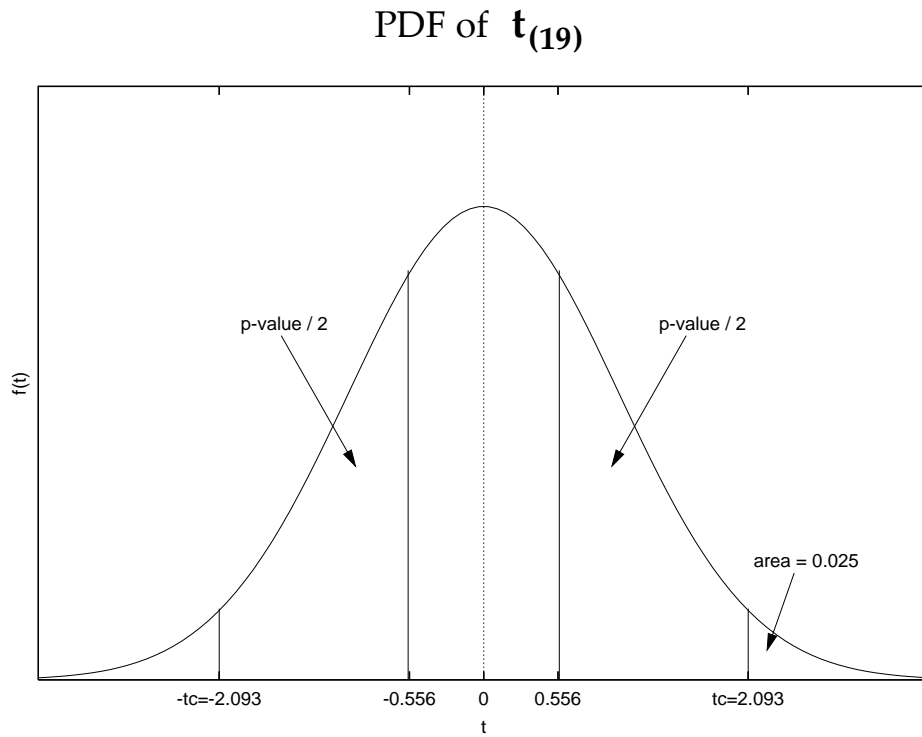
A sensible choice is a 5% significance level.

From the Appendix Table for the t-distribution, using 19 degrees of freedom and an upper tail area of $0.05/2 = 0.025$ the critical value is:

$$t_c = 2.093$$

It is clear that $|t| = 0.556$ is smaller than the critical value and so there is no evidence to reject the null hypothesis.

The graph below shows both the p-value calculation and the rejection region for the stock market returns example.



❖ Comparing Two Means from Independent Samples

The construction of a test statistic for the comparison of two means from independent samples depends on the assumptions made about the population variances. Different variance assumptions lead to different test statistics.

The lecture notes for Chapter 9.2 presented one set-up that will be presented here.

The key assumption is that the two populations have a common variance σ^2 that is estimated from the sample observations.

Example: The Chapter 9.2 example analyzed daily closing prices for the company Johnson & Johnson for the sample period January to June and the sample period July to December of the year 1999.

The data set contained $\mathbf{n}_x = 124$ observations for the first sample period and $\mathbf{n}_y = 128$ observations for the second sample period.

Denote the population means of the daily closing prices in the two sample periods by μ_X and μ_Y . Test the null hypothesis:

$$H_0: \mu_X = \mu_Y \quad \text{population means are equal}$$

against the one-sided alternative hypothesis:

$$H_1: \mu_X < \mu_Y \quad \text{higher mean in the second sample period}$$

That is, test:

$$H_0: \mu_X - \mu_Y = 0 \quad \text{against} \quad H_1: \mu_X - \mu_Y < 0$$

The sample means and variances for the closing prices in the two sample periods are denoted by \bar{x} , \bar{y} and s_x^2 , s_y^2 .

By using the statistical results given in the lecture notes for Chapter 9.2 a test statistic is calculated as:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s^2}{n_x} + \frac{s^2}{n_y}}}$$

where s^2 is the pooled sample variance constructed by combining all the observations in the two sample periods:

$$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)}$$

The test statistic can be compared with the t-distribution with $(n_x + n_y - 2)$ degrees of freedom.

For the comparison of the closing prices for the two sample periods the calculated test statistic is:

$$t = \frac{89.96 - 97.98}{\sqrt{\frac{26.91}{124} + \frac{26.91}{128}}} = -12.27$$

The degrees of freedom is $(124 + 128 - 2) = 250$.

For this one-tailed test, the p-value is calculated as:

$$P(t_{(250)} < -12.27)$$

By symmetry, this is the same as:

$$P(t_{(250)} > 12.27)$$

With Microsoft Excel the p-value is calculated with the function:

$$\text{TDIST}(12.27, 250, 1)$$

This gives the answer:

$$\text{p-value} = 1.05\text{E}-27 = \frac{1.05}{10^{27}} < 0.00005$$

The calculated p-value is lower than any reasonable significance level to give strong evidence to reject the null hypothesis. The information in the data suggests that, for 1999, the mean daily closing price for the company Johnson & Johnson was higher in the second half of the year compared to the first half of the year.

Chapter 11.4 Testing the Equality of Variances

Consider two random samples. Assume:

- independent samples, and
- normally distributed populations.

The samples have n_x and n_y observations and the population variances are σ_X^2 and σ_Y^2 .

Estimators of the population variances are the random variables s_X^2 and s_Y^2 .

Introduce the random variable:

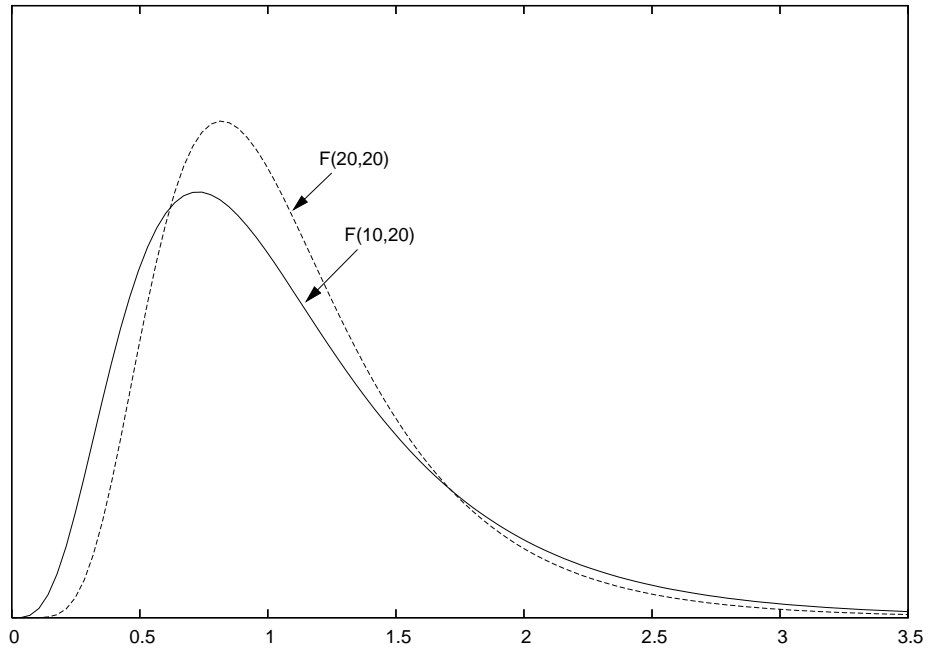
$$F = \frac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2}$$

F is the ratio of two independently distributed chi-square random variables. A result from statistical theory is that F has an **F-distribution** with numerator degrees of freedom $(n_x - 1)$ and denominator degrees of freedom $(n_y - 1)$.

That is, $F \sim F_{(n_x - 1, n_y - 1)}$
↑
'is distributed as'

Like the chi-square distribution, the F distribution is defined only for non-negative values and the skewed shape of the probability density function depends on the degrees of freedom.

PDF of the F distribution with
(10, 20) and (20, 20) degrees of freedom.



Appendix Table 9, pages 871–3 of the textbook, lists cutoff points or critical values that give an upper tail area of either 0.05 or 0.01.

The application of interest is to test the null hypothesis:

$$H_0: \sigma_X^2 = \sigma_Y^2 \quad \text{population variances are equal}$$

against the two-sided alternative:

$$H_1: \sigma_X^2 \neq \sigma_Y^2 \quad \text{population variances are not equal}$$

When the null hypothesis is true, $\sigma_X^2 = \sigma_Y^2$, and the random variable:

$$F = \frac{s_X^2}{s_Y^2} \quad \text{has an } F_{(n_x-1, n_y-1)} \text{ distribution.}$$

From the numeric data set calculate the sample variances s_x^2 and s_y^2 . Arrange the two samples so that $s_x^2 > s_y^2$.

A test statistic is calculated as the variance ratio:

$$\frac{s_x^2}{s_y^2}$$

The test statistic exceeds one since s_x^2 is bigger than s_y^2 .

For a chosen significance level α , the decision rule is to reject the null hypothesis of equal variances if:

$$\frac{s_x^2}{s_y^2} > F_c$$

where F_c is the critical value from the F-distribution that satisfies:

$$P(F_{(n_x-1, n_y-1)} > F_c) = \alpha/2$$

Appendix Table 9 lists critical values for upper tail probabilities of 0.05 or 0.01. Therefore, when testing against a two-sided alternative, the information in the printed table offers a choice of the significance level as either $\alpha = 0.10$ (a 10% level) or $\alpha = 0.02$ (a 2% level).

In applied work, the computer software is used for the calculation of descriptive statistics, the test statistic and an accompanying p-value. A p-value for the two-tailed test is calculated as:

$$\text{p-value} = 2 \cdot P\left(F_{(n_x-1, n_y-1)} > \frac{s_x^2}{s_y^2}\right)$$

For a chosen significance level α , the decision rule is to reject the null hypothesis if:

$$\text{p-value} < \alpha$$

A one-sided alternative can also be specified.

Test the null hypothesis:

$$\mathbf{H}_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{or} \quad \mathbf{H}_0 : \sigma_X^2 \leq \sigma_Y^2$$

against the alternative:

$$\mathbf{H}_1 : \sigma_X^2 > \sigma_Y^2$$

If a comparison of the calculated sample variances shows $\mathbf{s}_x^2 < \mathbf{s}_y^2$ then there is no evidence to reject the null hypothesis.

If $\mathbf{s}_x^2 > \mathbf{s}_y^2$ then calculate the test statistic:

$$\frac{\mathbf{s}_x^2}{\mathbf{s}_y^2}$$

For a chosen significance level α , the decision rule is to reject the null hypothesis of equal variances if:

$$\frac{\mathbf{s}_x^2}{\mathbf{s}_y^2} > \mathbf{F}_c$$

where \mathbf{F}_c is the critical value from the F-distribution that satisfies:

$$\mathbf{P}(\mathbf{F}_{(n_x-1, n_y-1)} > \mathbf{F}_c) = \alpha$$

Example: Exercise 11.28, page 393 of the textbook

Annual total sales of a company are reported for two sample periods:

Period 1: active price competition in the industry (4 years)

Period 2: price collusion in the industry (7 years)

It is hypothesized that total sales should vary more in an industry with active price competition compared to a market with price collusion.

Denote σ_X^2 and σ_Y^2 as the population variances in the two sample periods. Test the null hypothesis:

$$H_0: \sigma_X^2 = \sigma_Y^2 \quad \text{against the alternative:} \quad H_1: \sigma_X^2 > \sigma_Y^2$$

From the data set, the sample statistics are:

$$\text{Period 1: } n_x = 4 \quad s_x^2 = 114.09 \quad s_x = 10.68$$

$$\text{Period 2: } n_y = 7 \quad s_y^2 = 16.08 \quad s_y = 4.01$$

The test statistic is the variance ratio:

$$\frac{s_x^2}{s_y^2} = \frac{114.09}{16.08} = 7.095$$

This can be compared with the critical values reported in Appendix Table 9 for the F-distribution.

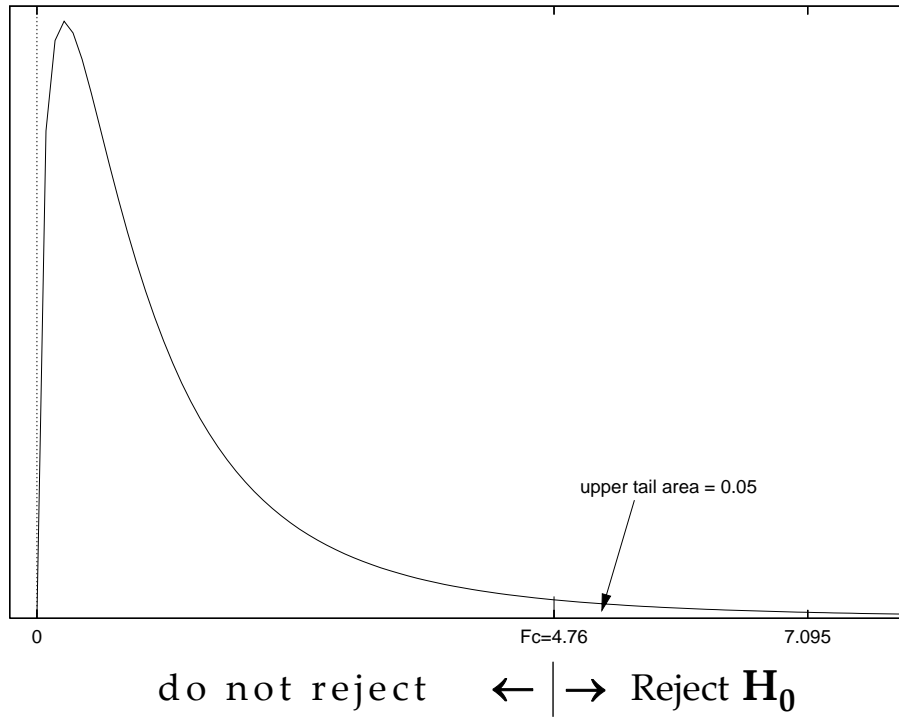
For a 5% significance level, with numerator degrees of freedom $(n_x - 1) = 3$ and denominator degrees of freedom $(n_y - 1) = 6$ the F-distribution critical value is:

$$F_c = 4.76$$

The calculated test statistic of 7.095 exceeds the critical value and therefore, there is evidence to reject the null hypothesis of equal variance in the two sample periods. The data suggests that the variance of sales is significantly higher in the period of active price competition.

The graph below illustrates that the calculated test statistic of 7.095 is in the rejection region for a 5% level one-tailed test.

PDF of $F_{(3,6)}$



Now approach a decision rule by calculating and interpreting a p-value for the test.

The p-value of the test is the smallest significance level at which a null hypothesis can be rejected.

In the above exercise, the null hypothesis was rejected at a 5% level and so the p-value must be smaller than 5%.

An exact p-value is obtained as the upper tail area of the F-distribution probability density function to the right of the calculated test statistic:

$$\text{p-value} = \mathbf{P(F_{(3,6)} > 7.095)}$$

With Microsoft Excel select Insert Function:

$$\begin{array}{c} \text{numerator degrees of freedom} \\ \downarrow \\ \text{FDIST}(7.095, 3, 6) \\ \uparrow \\ \text{denominator degrees of freedom} \end{array}$$

This returns the answer:

$$\text{p-value} = 0.021$$

The calculated p-value falls between 0.01 and 0.05.

The interpretation is that, although the null hypothesis can be rejected at a 5% level, it is not rejected at a 1% significance level.

This conclusion can be confirmed by checking the F-distribution tables. With $\alpha = 0.01$ the critical value from the $F_{(3,6)}$ distribution is:

$$F_c = 9.78$$

The calculated test statistic of 7.095 is below the 1% level critical value to suggest that the null hypothesis of equal population variances in the two sample periods is not rejected.

- Comment on methodology – in applied work the computer software used for the statistical analysis of the data can also be used for the calculation of p-values for test statistics. Therefore, hypothesis testing conclusions can rely on the interpretation of p-values. The statistical tables printed in textbook Appendices can have a useful role as a backup check of the test decision.

❖ More Examples of Testing for Equal Population Variances in Two Independent Samples

Consider daily stock market closing prices for a company observed in two sample periods. For Johnson & Johnson, a test of equal means in the first and second half of 1999, was developed in the lecture notes for Chapters 9.2 and 11.1. The test method assumed the same variance in the two sample periods. If this variance assumption does not describe the data then the test conclusions may be unreliable.

Is the assumption of equal variance in the two sample periods a reasonable assumption? To answer this question, test the null hypothesis of equal variances in the two samples against a two-sided alternative.

Summary statistics for the closing prices in the two sample periods were reported as:

$$\text{January to June} \quad \mathbf{n}_x = 124 \quad \mathbf{s}_x^2 = 32.54$$

$$\text{July to December} \quad \mathbf{n}_y = 128 \quad \mathbf{s}_y^2 = 21.45$$

The test statistic is the variance ratio:

$$\frac{s_x^2}{s_y^2} = \frac{32.54}{21.45} = 1.517$$

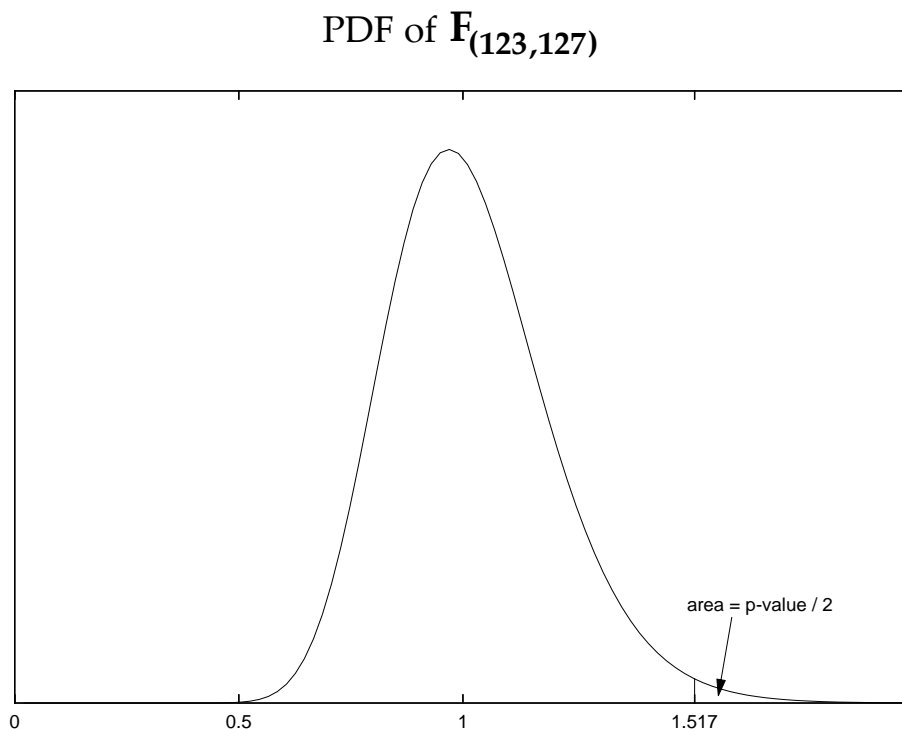
Note that the larger variance is placed in the numerator to give a test statistic that exceeds one.

The test statistic can be compared with an F-distribution with numerator degrees of freedom $(124 - 1) = 123$ and denominator degrees of freedom $(128 - 1) = 127$.

For a two-tailed test, the p-value is found as:

$$\text{p-value} = 2 \cdot \mathbf{P}(\mathbf{F}_{(123,127)} > 1.517)$$

The p-value calculation is illustrated in the graph below.



To find an exact p-value, with Microsoft Excel select Insert Function:

$$\text{FDIST}(1.517, 123, 127)$$

This gives the answer: $\text{pvalue} / 2 = 0.0102$

Therefore, the p-value for the two-tailed test is:

$$2 (0.0102) = 0.0204$$

At a 1% significance level the p-value exceeds $\alpha = 0.01$.

This gives evidence that, for the Johnson & Johnson daily closing prices, the null hypothesis of equal variance in the two sample periods is not rejected.

A Test of Normality

Textbook Reference: Chapter 16.2, pages 619–21.

The calculation of p-values for hypothesis testing typically is based on the assumption that the population distribution is normal. Therefore, a test of the normality assumption may be useful to inspect. A variety of tests of normality have been developed by various statisticians. One of these tests will be described here.

To start, the calculation of descriptive statistics is reviewed.

A data set has the numeric observations: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$.

Familiar descriptive statistics are the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

and the sample variance:

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

Now introduce two new statistics.

The sample **skewness** is defined as:

$$S = \frac{1}{n} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(\tilde{\sigma}^2)^{3/2}} \quad \text{where} \quad \tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Skewness gives a measure of how symmetric the observations are about the mean. For a normal distribution the skewness is **0**. A distribution skewed to the right has positive skewness and a distribution skewed to the left has negative skewness.

The sample **kurtosis** is defined as:

$$K = \frac{1}{n} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{(\tilde{\sigma}^2)^2}$$

Kurtosis gives a measure of the thickness in the tails of a probability density function. For a normal distribution the kurtosis is **3**.

Excess kurtosis is defined as:

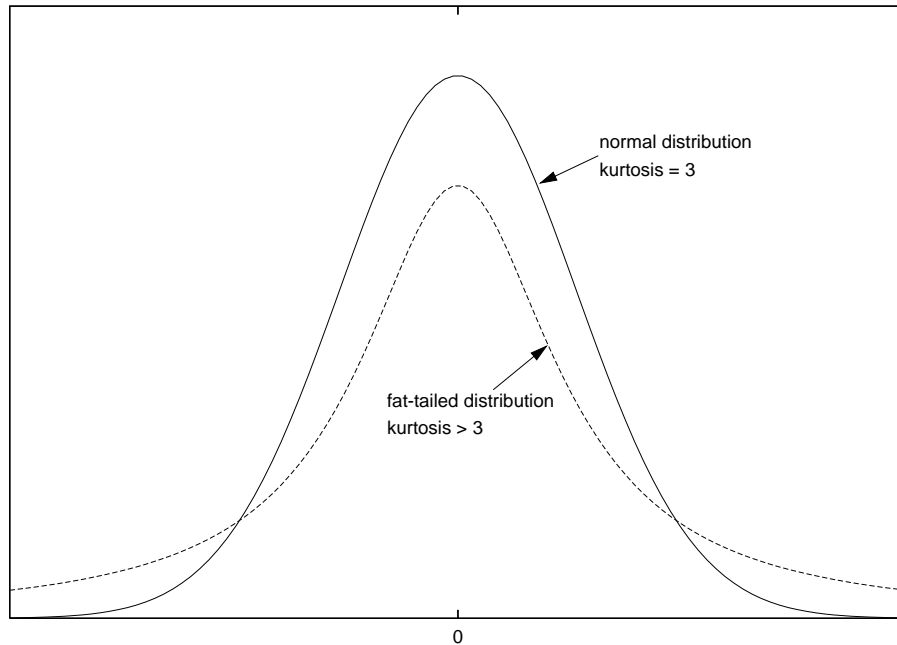
$$EK = K - 3$$

It follows that, for a normal distribution, the excess kurtosis is **0**.

A fat-tailed or thick-tailed distribution has a value for kurtosis that exceeds **3**. That is, excess kurtosis is positive.

This is called **leptokurtosis**.

The graph below compares the shape of the probability density function for the normal distribution and a fat-tailed distribution.



The above calculation formula for skewness and kurtosis are considered suitable for 'large samples'.

Formula that incorporate 'small sample' adjustments are available. The adjusted calculation formula for skewness is:

$$g_1 = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(s^2)^{3/2}}$$

The adjusted calculation formula for excess kurtosis is:

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

With Microsoft Excel the function SKEW reports skewness and the function KURT reports *excess* kurtosis using the formula g_1 and g_2 .

A test of normality is now proposed.

Consider testing the null hypothesis:

H₀: normal distribution,
skewness is zero and excess kurtosis is zero;

against the alternative hypothesis:

H₁: non-normal distribution.

A test statistic is:

$$\mathbf{n} \cdot \left[\frac{\mathbf{S}^2}{\mathbf{6}} + \frac{(\mathbf{EK})^2}{\mathbf{24}} \right]$$

It turns out that this test statistic can be compared with a χ^2 (chi-square) distribution with 2 degrees of freedom.

The null hypothesis of normality is rejected if the calculated test statistic exceeds a critical value from the $\chi^2_{(2)}$ distribution.

The critical values can be found from the Appendix Table for the chi-square distribution as:

Significance Level α	Critical Value
0.10	4.61
0.05	5.99
0.01	9.21

The presentation of this test of normality is valid for 'large samples'. For 'small samples' the decision rule can be viewed as approximate.

The test is called the Bowman-Shelton test.

For application to economic data the test is known as the Jarque-Bera test.

Example: A stock market data set has daily percentage returns observed for the year 1997 for two companies – Barrick Gold and Bank of New York.

The sample has observations for $n = 253$ trading days.

For each company, an exercise is to test for normality of the daily returns. Various statistics are given in the table below.

Both the ‘small sample’ and ‘large sample’ versions of the skewness and excess kurtosis statistics are presented to give emphasis to the methodology.

	Barrick Gold	Bank of NY
‘Small sample’ statistics		
Skewness \mathbf{g}_1	0.01	-0.14
Excess Kurtosis \mathbf{g}_2	1.38	0.41
‘Large sample’ statistics		
Skewness \mathbf{S}	0.01	-0.14
Excess Kurtosis \mathbf{EK}	1.33	0.38
Bowman-Shelton test of normality – calculated with the ‘large sample’ statistics		
test statistic	18.73	2.31
p-value	< 0.0005	0.315

For Barrick Gold, the normality test statistic of 18.73 exceeds the critical values for any reasonable significance level to lead to the conclusion that the daily returns do not follow a normal distribution.

Since the excess kurtosis statistic is greater than zero, the appearance is that the daily returns follow a distribution that features leptokurtosis. Researchers have suggested that the leptokurtosis arises from a pattern of volatility in financial markets where periods of high volatility are followed by periods of relative stability.

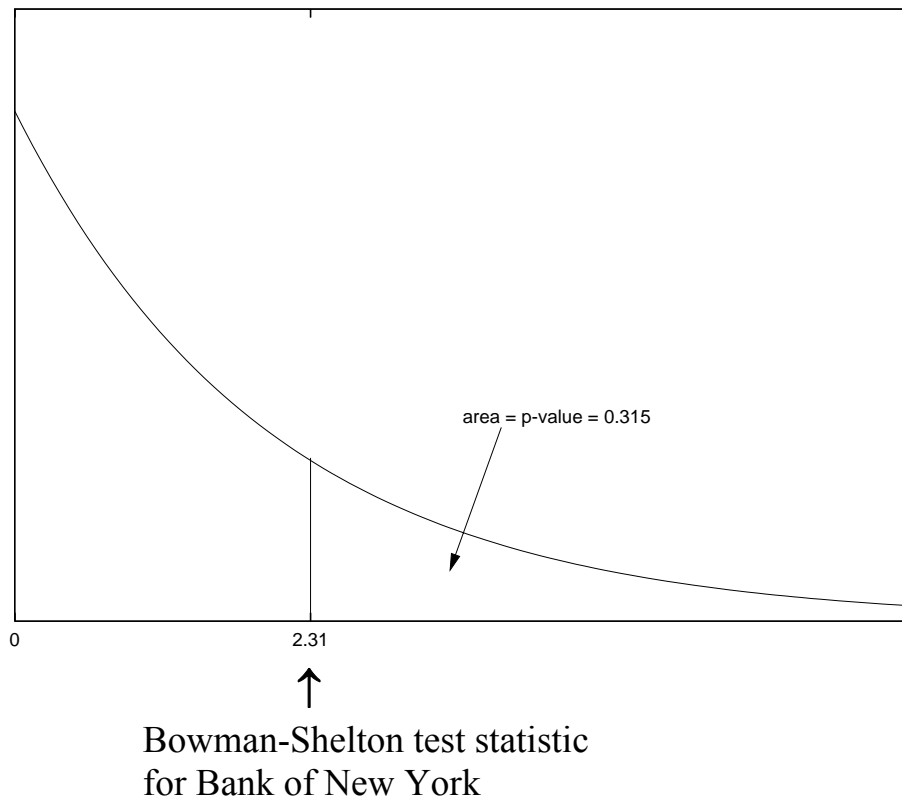
A p-value for the test statistic is calculated as a chi-square distribution probability and, with Microsoft Excel, is computed with the function:

CHIDIST(test_statistic, 2)
 ↑
 degrees of freedom

For the Bank of New York, the calculation of a p-value for the normality test statistic is illustrated in the graph below.

A statistical result is that the chi-square distribution with two degrees of freedom is an exponential distribution.

PDF for the χ^2 distribution
with 2 degrees of freedom



It is clear that the calculated p-value is greater than any standard significance level α to suggest that there is no evidence to reject the null hypothesis of a normal distribution for the daily returns of the Bank of New York.