Probability & Statistics

Notes

Saad Bin Khalid 20K-0161 9 June 2022

Chapter 4 – Probability Concepts

Exercise 4.2 – Events

DEFINITION 4.2

Sample Space and Event

Sample space: The collection of all possible outcomes for an experiment.

Event: A collection of outcomes for the experiment, that is, any subset of the sample space. An event **occurs** if and only if the outcome of the experiment is a member of the event.

DEFINITION 4.3

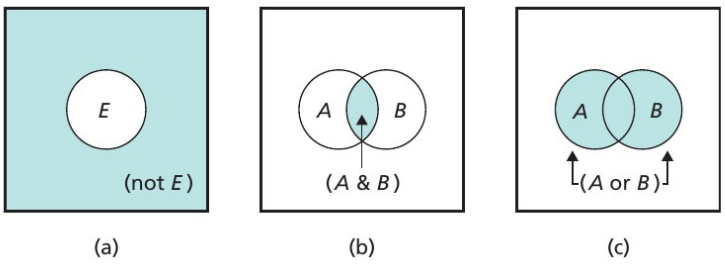
Relationships Among Events

(not *E*): The event "*E* does not occur"

(*A* & *B*): The event "both *A* and *B* occur"

(*A* or *B*): The event "either *A* or *B* or both occur"

FIGURE 4.9
Venn diagrams for (a) event (not *E*), (b) event (*A* & *B*), and (c) event (*A* or *B*)



DEFINITION 4.4

Mutually Exclusive Events

Two or more events are **mutually exclusive events** if no two of them have outcomes in common.

Does It Mean?

are mutually
no two of them can

FIGURE 4.14
(a) Two mutually exclusive events; (b) two non–mutually exclusive events

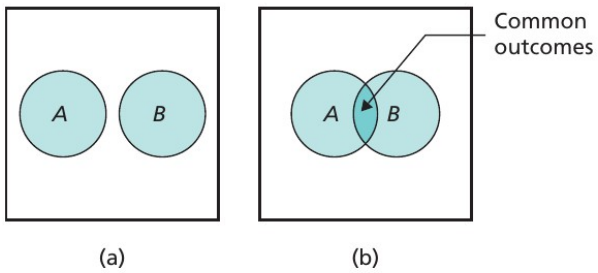
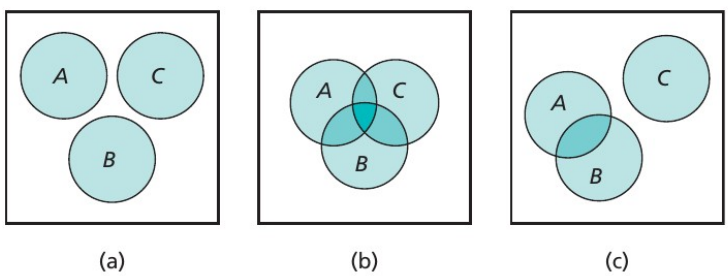


FIGURE 4.15
(a) Three mutually exclusive events; (b) three non–mutually exclusive events; (c) three non–mutually exclusive events



Exercise 4.3 – Some Rules of Probability

FORMULA 4.1 **The Special Addition Rule**

If event A and event B are mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B).$$

More generally, if events A, B, C, ... are mutually exclusive, then

$$P(A \text{ or } B \text{ or } C \text{ or } \cdots) = P(A) + P(B) + P(C) + \cdots.$$

FORMULA 4.3 **The General Addition Rule**

If A and B are any two events, then

$$P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B).$$

Exercise 4.4 – Contingency Tables; Joint and Marginal Probabilities

Data from two variables of a population are called *bivariate data*, and a frequency distribution for bivariate data is called *a contingency table or two-way table.*

		Rank			
		Full professor R_1	Associate professor R_2	Assistant professor R_3	Instructor
Age (yr)	Under 30 A_1	2	3	57	
	30–39 A_2	52	170	163	
	40–49 A_3	156	125	61	
	50–59 A_4	145	68	36	
	60 & over A_5	75	15	3	
	Total	430	381	320	

	R_1	R_2	R_3	R_4
A_1	$(A_1 \,\&\, R_1)$	$(A_1 \,\&\, R_2)$	$(A_1 \,\&\, R_3)$	$(A_1 \,\&\, R_4)$
A_2	$(A_2 \,\&\, R_1)$	$(A_2 \,\&\, R_2)$	$(A_2 \,\&\, R_3)$	$(A_2 \,\&\, R_4)$
A_3	$(A_3 \,\&\, R_1)$	$(A_3 \,\&\, R_2)$	$(A_3 \,\&\, R_3)$	$(A_3 \,\&\, R_4)$
A_4	$(A_4 \,\&\, R_1)$	$(A_4 \,\&\, R_2)$	$(A_4 \,\&\, R_3)$	$(A_4 \,\&\, R_4)$
A_5	$(A_5 \,\&\, R_1)$	$(A_5 \,\&\, R_2)$	$(A_5 \,\&\, R_3)$	$(A_5 \,\&\, R_4)$

These nine probabilities are often called *marginal probabilities* because they correspond to events represented in the margin of the contingency table. We can also find probabilities for joint events, so-called joint probabilities. For instance, the probability that the selected faculty member is an associate professor under 30 [event (A 1 & R 2)].

Exercise 4.5 – Conditional Probability

DEFINITION 4.6

Conditional Probability

The probability that event B occurs given that event A occurs is called a **conditional probability.** It is denoted $\boldsymbol{P(B \mid A)}$, which is read "the probability of B given A." We call A the **given event.**

If A and B are any two events with $P(A) > 0$, then

$$P(B \mid A) = \frac{P(A \,\&\, B)}{P(A)}.$$

Exercise 4.6 – The Multiplication Rule; Independence*

FORMULA 4.5

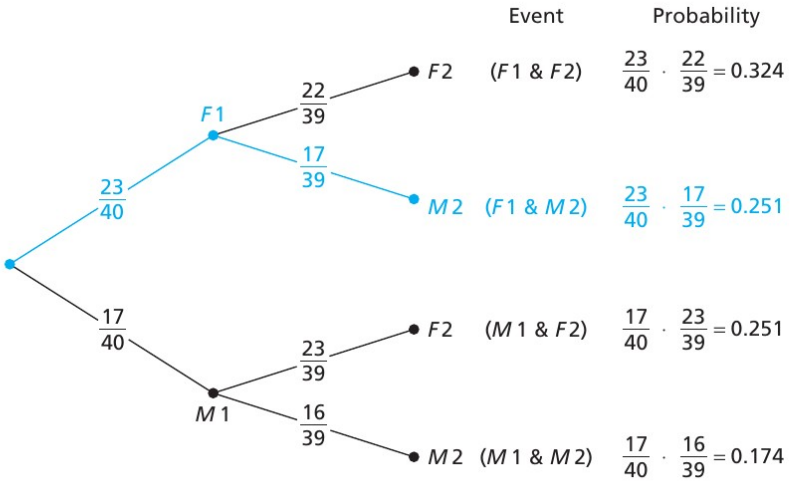
The General Multiplication Rule

If A and B are any two events, then

$$P(A \,\&\, B) = P(A) \cdot P(B \mid A).$$

FIGURE 4.25
Tree diagram for student-selection problem

Gender	Frequency
Male	17
Female	23
	40



DEFINITION 4.7

Independent Events

Event B is said to be **independent** of event A if $P(B \mid A) = P(B)$.

FORMULA 4.6

The Special Multiplication Rule (for Two Independent Events)

If A and B are independent events, then

$$P(A \,\&\, B) = P(A) \cdot P(B),$$

and conversely, if $P(A \,\&\, B) = P(A) \cdot P(B)$, then A and B are independent events.

The terms mutually exclusive and independent refer to different concepts. Mutually exclusive events are those that cannot occur simultaneously; independent events are those for which the occurrence of some does not affect the probabilities of the others occurring.
In fact, if two or more events are mutually exclusive, the occurrence of one precludes the occurrence of the others. Two or more (nonimpossible) events cannot be both mutually exclusive and independent.

Exercise 4.7 – Bayes's Rule

Exhaustive Events

Events $A_1, A_2, \ldots, A_k$ are said to be exhaustive events if one or more of them must occur.

In general, if events are both exhaustive and mutually exclusive, exactly one of them must occur. An event and its complement are always mutually exclusive and exhaustive.

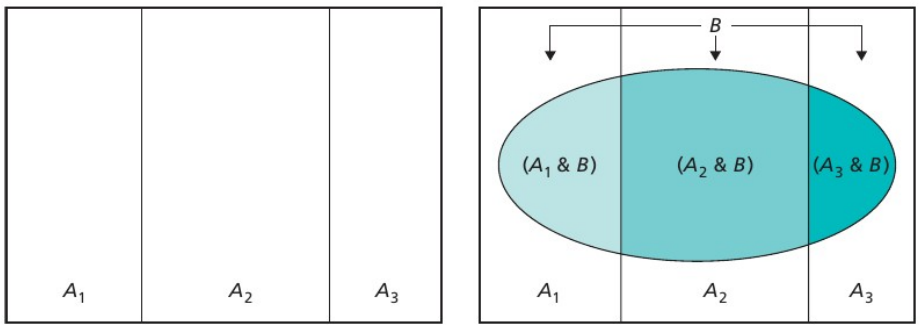
FORMULA 4.8

The Rule of Total Probability

Suppose that events $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive; that is, exactly one of the events must occur. Then for any event B,

$$P(B) = \sum_{j=1}^{k} P(A_j) \cdot P(B \mid A_j).$$

FIGURE 4.26
(a) Three mutually exclusive and exhaustive events; (b) an event B and three mutually exclusive and exhaustive events



(a) (b)

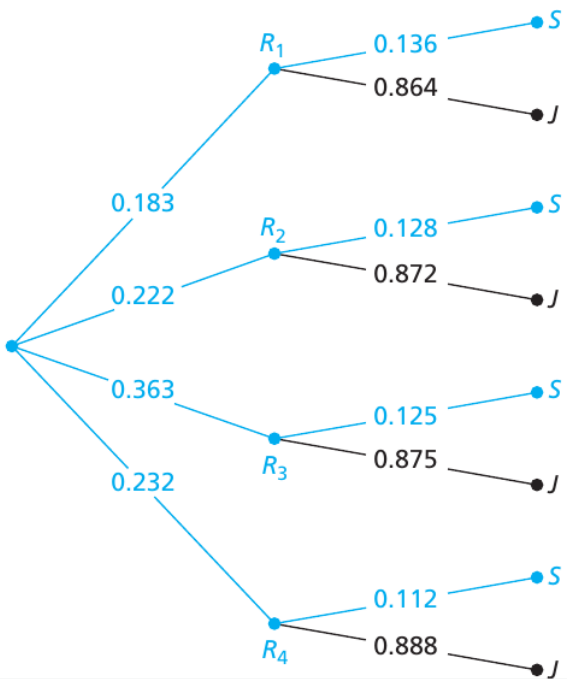
TABLE 4.11
Percentage distribution for region of residence and percentage of seniors in each region

Region	Percentage of U.S. population	Percentage seniors
Northeast	18.3	13.6
Midwest	22.2	12.8
South	36.3	12.5
West	23.2	11.2
	100.0	

TABLE 4.12
Probabilities derived from Table 4.11

$P(R_1) = 0.183 \quad P(S \mid R_1) = 0.136$
$P(R_2) = 0.222 \quad P(S \mid R_2) = 0.128$
$P(R_3) = 0.363 \quad P(S \mid R_3) = 0.125$
$P(R_4) = 0.232 \quad P(S \mid R_4) = 0.112$

FIGURE 4.27
Tree diagram for calculating $P(S)$, using the rule of total probability



$$P(S) = \sum_{j=1}^{4} P(R_j) \cdot P(S \mid R_j)$$
$$= 0.183 \cdot 0.136 + 0.222 \cdot 0.128 + 0.363 \cdot 0.125 + 0.232 \cdot 0.112$$
$$= 0.125.$$

FORMULA 4.9

Bayes's Rule

Suppose that events $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive. Then for any event B,

$$P(A_i \mid B) = \frac{P(A_i) \cdot P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j) \cdot P(B \mid A_j)},$$

where A_i can be any one of events $A_1, A_2, \ldots, A_k$.

Chapter 5 – Discrete Random Variables

Exercise 5.1 – Discrete Random Variables and Probability Distributions

DEFINITION 5.1 **Random Variable**

A **random variable** is a quantitative variable whose value depends on chance.

DEFINITION 5.2 **Discrete Random Variable**

A **discrete random variable** is a random variable whose possible values can be listed.

Recall that we use lowercase letters such as x, y, and z to denote variables. To represent random variables, however, we usually use uppercase letters. For example, let X denote the number of siblings of a randomly selected student. Then we can represent the event that the selected student has two siblings by $\{X = 2\}$, read "X equals two," and the probability of that event as $P(X = 2)$, read "the probability that X equals two."

KEY FACT 5.1 **Sum of the Probabilities of a Discrete Random Variable**

For any discrete random variable X, we have $\Sigma P(X = x) = 1$.†

Exercise 5.2 – The Mean & Standard Deviation of a Discrete Random Variable

DEFINITION 5.4 **Mean of a Discrete Random Variable**

The **mean of a discrete random variable** X is denoted μ_X or, when no confusion will arise, simply μ. It is defined by

$$\mu = \Sigma x P(X = x).$$

The terms **expected value** and **expectation** are commonly used in place of the term *mean*.†

DEFINITION 5.5 **Standard Deviation of a Discrete Random Variable**

The **standard deviation of a discrete random variable** X is denoted σ_X or, when no confusion will arise, simply σ. It is defined as

$$\sigma = \sqrt{\Sigma (x - \mu)^2 P(X = x)}.$$

The standard deviation of a discrete random variable can also be obtained from the computing formula

$$\sigma = \sqrt{\Sigma x^2 P(X = x) - \mu^2}.$$

x	$P(X = x)$	x^2	$x^2 P(X = x)$
0	0.029	0	0.000
1	0.049	1	0.049
2	0.078	4	0.312
3	0.155	9	1.395
4	0.212	16	3.392
5	0.262	25	6.550
6	0.215	36	7.740
			19.438

Exercise 5.3 – The Binomial Distribution

DEFINITION 5.7

Binomial Coefficients

If n is a positive integer and x is a nonnegative integer less than or equal to n, then the **binomial coefficient** $\binom{n}{x}$ is defined as

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}.^{\dagger}$$

DEFINITION 5.8

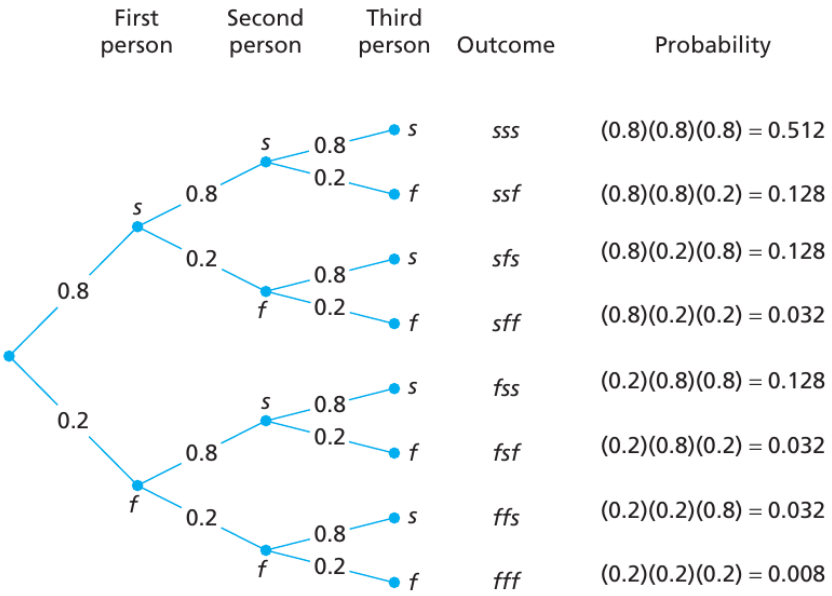
Bernoulli Trials

Repeated trials of an experiment are called **Bernoulli trials** if the following three conditions are satisfied:

1. The experiment (each trial) has two possible outcomes, denoted generically ***s***, for **success,** and ***f***, for **failure.**
2. The trials are independent.
3. The probability of a success, called the **success probability** and denoted ***p***, remains the same from trial to trial.

In United States, a person of age 20 years has about an 80% chance of being alive at age 65 years. Suppose three people of age 20 years are selected at random. Find:

- Probability of each outcome.
- Probability that exactly two of the three people will be alive at age 65.



Exactly 2 of the 3 people are alive consists of the ssf, sfs, and fss. By the special addition rule:

$$\begin{aligned} P(\text{Exactly two will be alive}) &= P(ssf) + P(sfs) + P(fss) \\ &= \underbrace{0.128 + 0.128 + 0.128}_{\text{3 times}} = 3 \cdot 0.128 = 0.384. \end{aligned}$$

KEY FACT 5.4

Number of Outcomes Containing a Specified Number of Successes

In n Bernoulli trials, the number of outcomes that contain exactly x successes equals the binomial coefficient $\binom{n}{x}$.

FORMULA 5.1

Binomial Probability Formula

Let X denote the total number of successes in n Bernoulli trials with success probability p. Then the probability distribution of the random variable X is given by

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

The random variable X is called a **binomial random variable** and is said to have the **binomial distribution** with parameters n and p.

PROCEDURE 5.1 To Find a Binomial Probability Formula

Assumptions

1. n trials are to be performed.
2. Two outcomes, success or failure, are possible for each trial.
3. The trials are independent.
4. The success probability, p, remains the same from trial to trial.

Step 1 **Identify a success.**

Step 2 **Determine p, the success probability.**

Step 3 **Determine n, the number of trials.**

Step 4 **The binomial probability formula for the number of successes, X, is**

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

FORMULA 5.2

Mean and Standard Deviation of a Binomial Random Variable

The mean and standard deviation of a binomial random variable with parameters n and p are

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{np(1 - p)},$$

respectively.

Exercise 5.5 – From 2nd Book

Definition 3.1: A **random variable** is a function that associates a real number with each element in the sample space.

Definition 3.8: The function $f(x, y)$ is a **joint probability distribution** or **probability mass function** of the discrete random variables X and Y if

1. $f(x, y) \geq 0$ for all (x, y),
2. $\sum_x \sum_y f(x, y) = 1$,
3. $P(X = x, Y = y) = f(x, y)$.

For any region A in the xy plane, $P[(X, Y) \in A] = \sum\sum_A f(x, y)$.

In dealing with continuous variables, f (x) is usually called the probability density function, or simply the density function, of X.

Definition 3.6: The function $f(x)$ is a **probability density function** (pdf) for the continuous random variable X, defined over the set of real numbers, if

1. $f(x) \geq 0$, for all $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\, dx$.

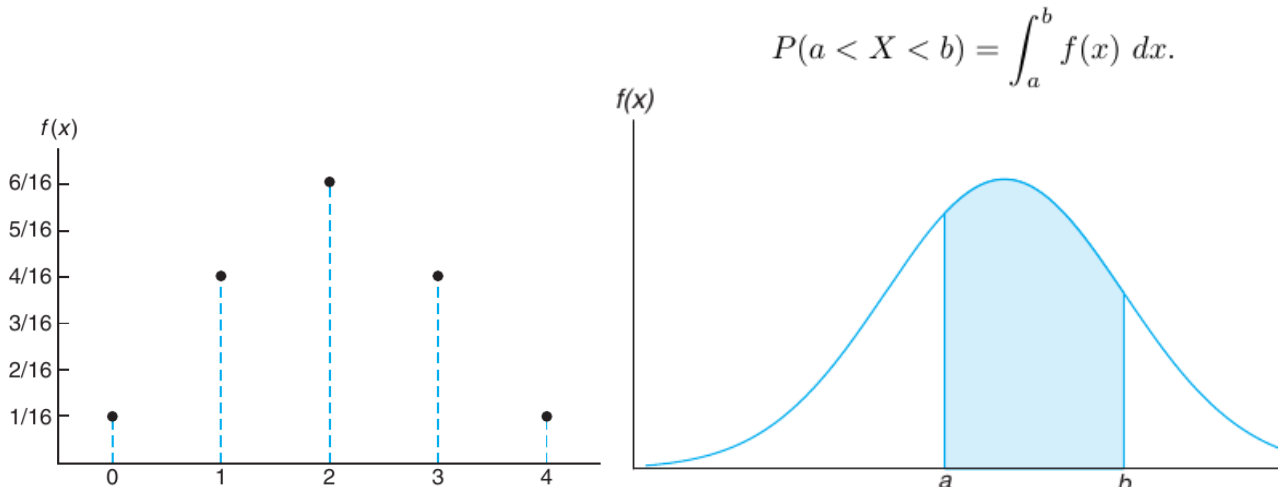


Figure 3.1: Probability mass function plot.

Example 3.10: Find the cumulative distribution function of the random variable X in Example 3.9. Using $F(x)$, verify that $f(2) = 3/8$.

***Solution*:** Direct calculations of the probability distribution of Example 3.9 give $f(0)= 1/16$, $f(1) = 1/4$, $f(2)= 3/8$, $f(3)= 1/4$, and $f(4)= 1/16$. Therefore,

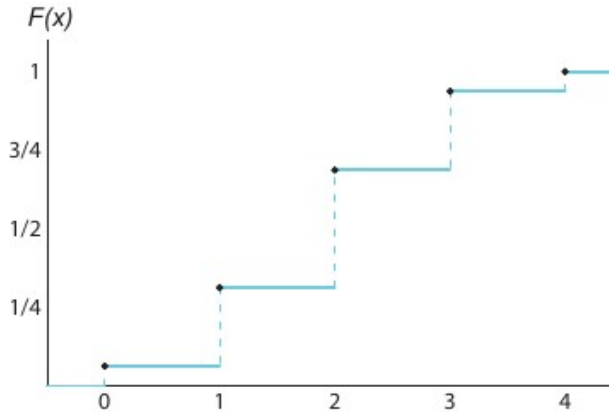
$$\begin{aligned}
F(0) &= f(0) = \frac{1}{16},\\
F(1) &= f(0) + f(1) = \frac{5}{16},\\
F(2) &= f(0) + f(1) + f(2) = \frac{11}{16},\\
F(3) &= f(0) + f(1) + f(2) + f(3) = \frac{15}{16},\\
F(4) &= f(0) + f(1) + f(2) + f(3) + f(4) = 1.
\end{aligned}$$

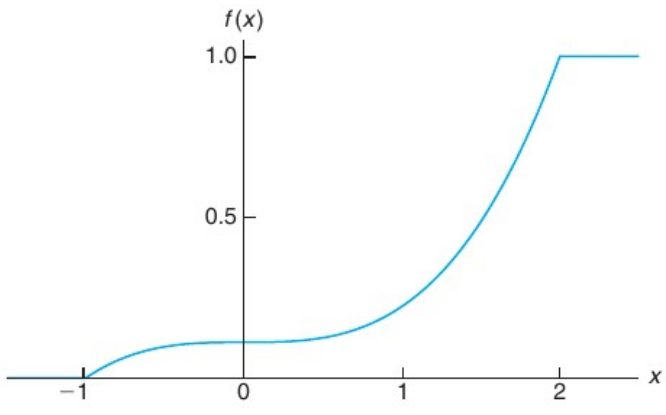
Hence,

$$F(x) = \begin{cases} 0, & \text{for } x < 0,\\ \frac{1}{16}, & \text{for } 0 \le x < 1,\\ \frac{5}{16}, & \text{for } 1 \le x < 2,\\ \frac{11}{16}, & \text{for } 2 \le x < 3,\\ \frac{15}{16}, & \text{for } 3 \le x < 4,\\ 1 & \text{for } x \ge 4. \end{cases}$$

Now

$$f(2) = F(2) - F(1) = \frac{11}{16} - \frac{5}{16} = \frac{3}{8}.$$





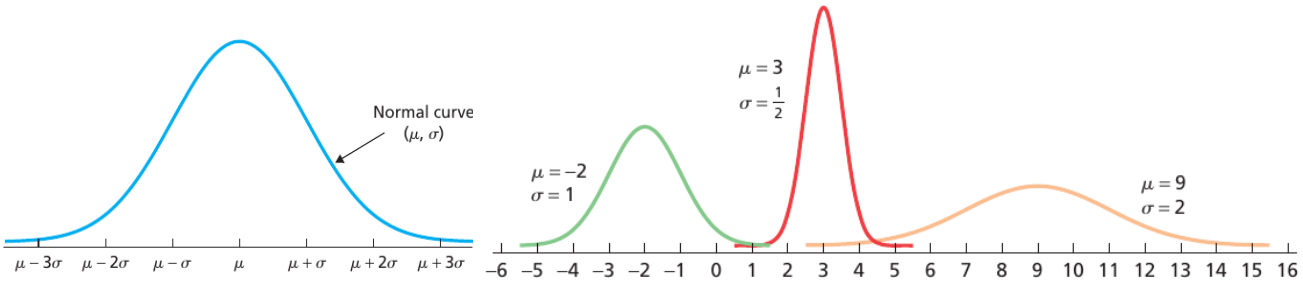
Chapter 6 – The Normal Distribution

Exercise 6.1 – Introduction

DEFINITION 6.1 **Normally Distributed Variable**

A variable is said to be a **normally distributed variable** or to have a **normal distribution** if its distribution has the shape of a normal curve.

A normal distribution is completely determined by the *mean* and *standard deviation.* These are the parameters of the normal curve. It is symmetric about and centered at the mean, and its spread depends on the standard deviation of the variable.



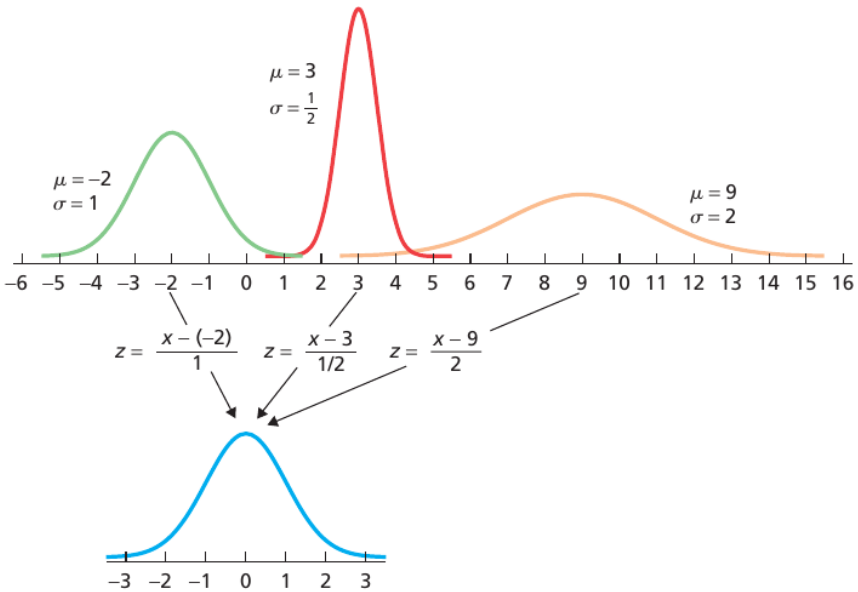
DEFINITION 6.2 **Standard Normal Distribution; Standard Normal Curve**

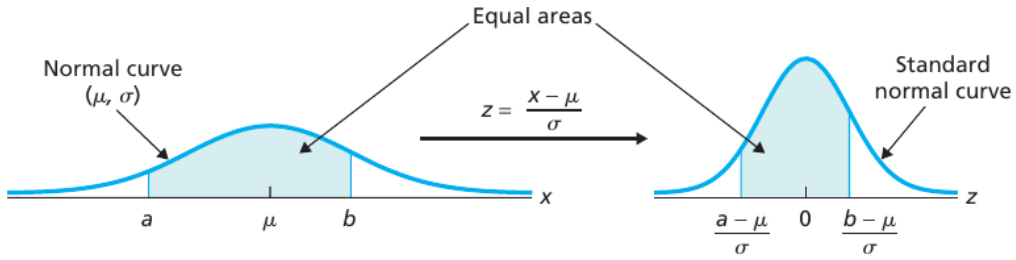
A normally distributed variable having mean 0 and standard deviation 1 is said to have the **standard normal distribution.** Its associated normal curve is called the **standard normal curve,** which is shown in Fig. 6.5.

The standardized version of a normally distributed variable x,

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

has the standard normal distribution.





Exercise 6.2 – Areas under the Standard Normal Curve

KEY FACT 6.5

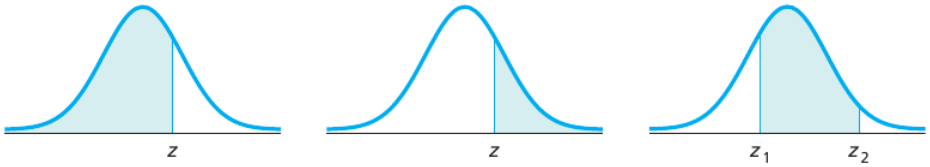
Basic Properties of the Standard Normal Curve

Property 1: The total area under the standard normal curve is 1.

Property 2: The standard normal curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.

Property 3: The standard normal curve is symmetric about 0; that is, the part of the curve to the left of the dashed line in Fig. 6.8 is the mirror image of the part of the curve to the right of it.

Property 4: Almost all the area under the standard normal curve lies between −3 and 3.



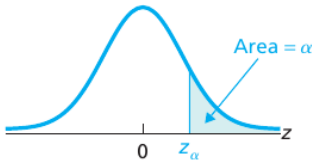
(a) Shaded area: Area to left of z

(b) Shaded area: 1 – (Area to left of z)

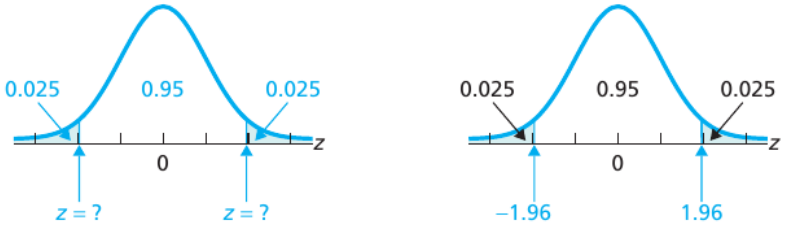
(c) Shaded area: (Area to left of z_2) – (Area to left of z_1)

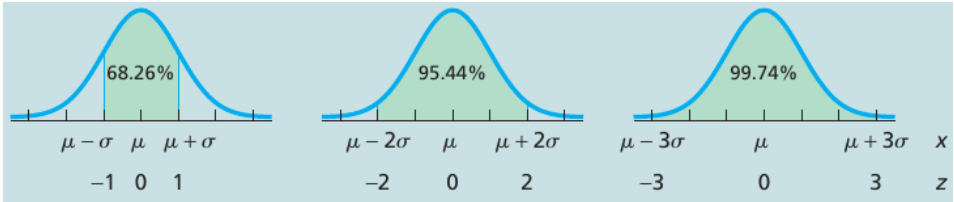
DEFINITION 6.3

The z_α Notation



The symbol $\boldsymbol{z_\alpha}$ is used to denote the z-score that has an area of α (alpha) to its right under the standard normal curve, as illustrated in Fig. 6.14. Read "z_α" as "z sub α" or more simply as "$z\,\alpha$."





Chapter 7 – The Sample Distribution of the Sample Mean

Exercise 7.1 – Sampling Error; the Need for Sampling Distributions

DEFINITION 7.1

Sampling Error

Sampling error is the error resulting from using a sample to estimate a population characteristic.

DEFINITION 7.2

Sampling Distribution of the Sample Mean

For a variable x and a given sample size, the distribution of the variable $\bar{x}$ is called the **sampling distribution of the sample mean.**

Does It Mean?

mpling distribution

All possible observations of the statistic for samples of a given size is called the sampling distribution of the statistic.

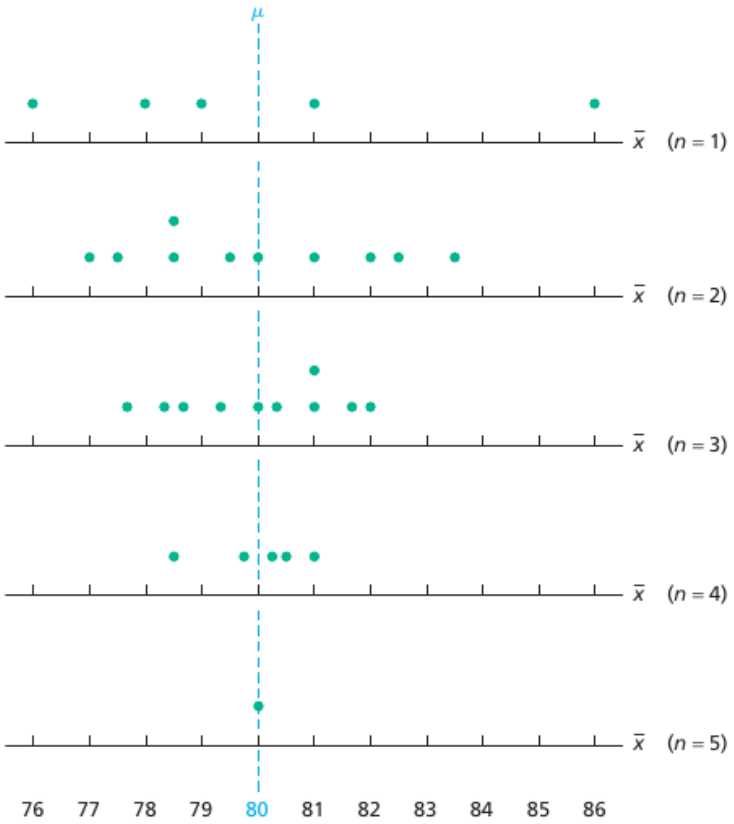


Figure 7.3 vividly illustrates that the possible sample means cluster more closely around the population mean as the sample size increases. This result suggests that sampling error tends to be smaller for large samples than for small samples.

KEY FACT 7.1

Sample Size and Sampling Error

The larger the sample size, the smaller the sampling error tends to be in estimating a population mean, μ, by a sample mean, $\bar{x}$.

es It Mean?

le sample

Exercise 7.2 – The Mean and Standard Deviation of Sample Mean

FORMULA 7.1

Mean of the Sample Mean

For samples of size n, the mean of the variable $\bar{x}$ equals the mean of the variable under consideration. In symbols,

$$\mu_{\bar{x}} = \mu.$$

FORMULA 7.2

Standard Deviation of the Sample Mean

For samples of size n, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration divided by the square root of the sample size. In symbols,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

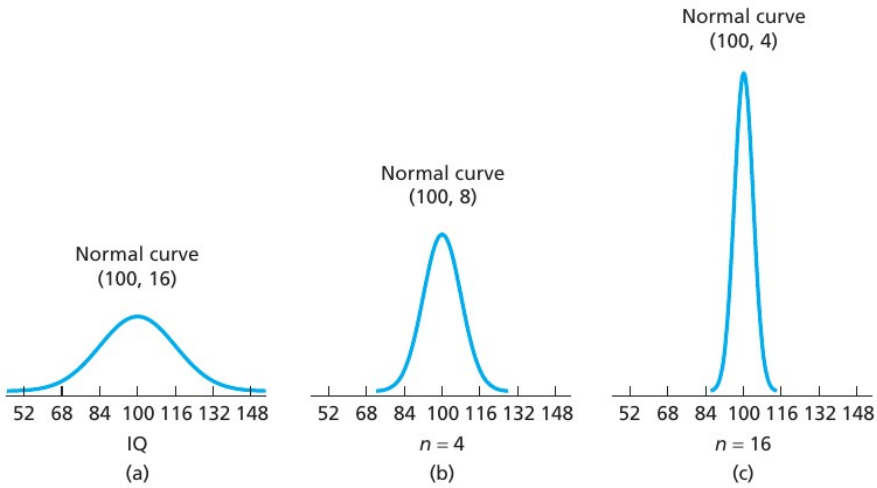
Exercise 7.3 – The Sampling Distribution of the Sample Mean

KEY FACT 7.2

Sampling Distribution of the Sample Mean for a Normally Distributed Variable

Suppose that a variable x of a population is normally distributed with mean μ and standard deviation σ. Then, for samples of size n, the variable $\bar{x}$ is also normally distributed and has mean μ and standard deviation $\sigma/\sqrt{n}$.

FIGURE 7.4
(a) Normal distribution for IQs; (b) sampling distribution of the sample mean for $n = 4$; (c) sampling distribution of the sample mean for $n = 16$



KEY FACT 7.3

The Central Limit Theorem (CLT)

For a relatively large sample size, the variable $\bar{x}$ is approximately normally distributed, regardless of the distribution of the variable under consideration. The approximation becomes better with increasing sample size.

The sample size must be large for a normal distribution to provide an adequate approximation to $\bar{x}$. Usually, a sample size of 30 or more ($n \geq 30$) is large enough.

KEY FACT 7.4

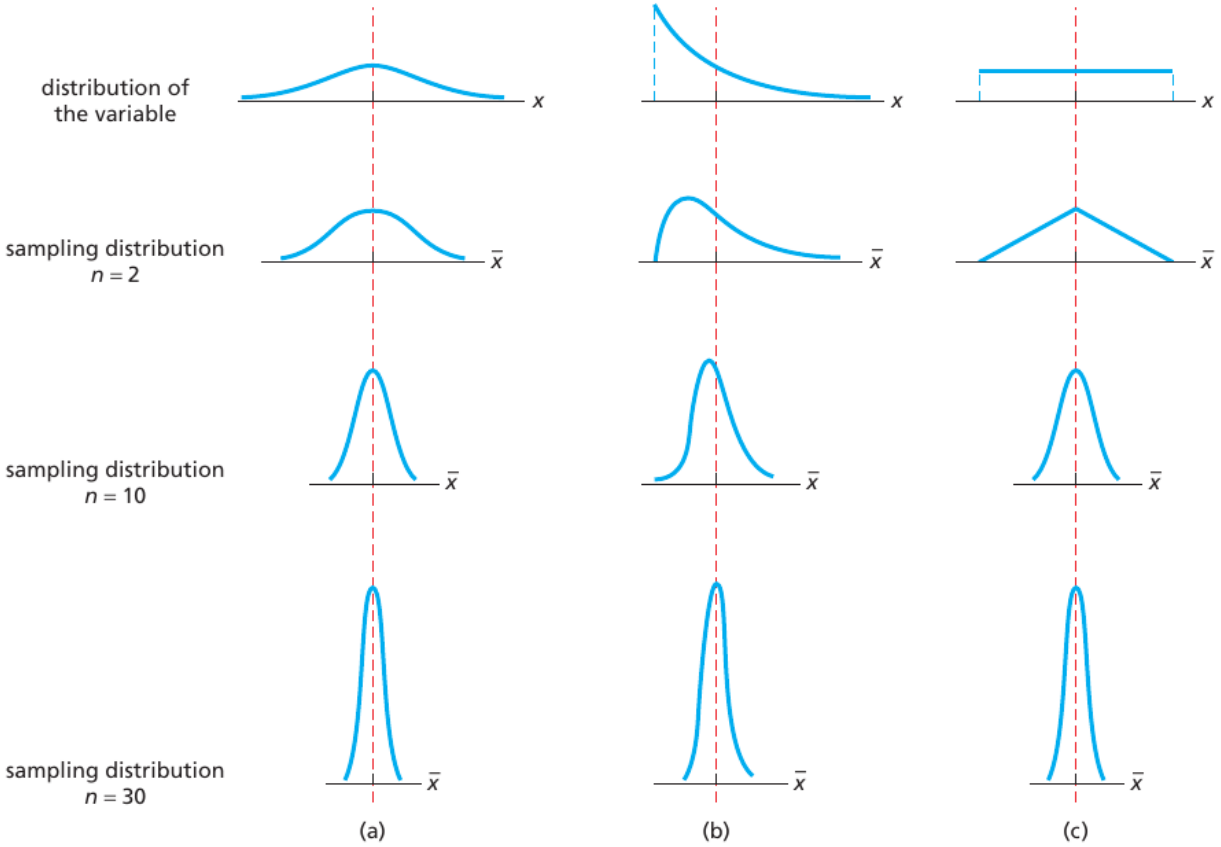
Does It Mean?

the variable under
n is normally
r the sample size is
ne possible sample
at least approxi-
mal distribution
and standard
$\sqrt{n}$

Sampling Distribution of the Sample Mean

Suppose that a variable x of a population has mean μ and standard deviation σ. Then, for samples of size n,

- the mean of $\bar{x}$ equals the population mean, or $\mu_{\bar{x}} = \mu$;
- the standard deviation of $\bar{x}$ equals the population standard deviation divided by the square root of the sample size, or $\sigma_{\bar{x}} = \sigma/\sqrt{n}$;
- if x is normally distributed, so is $\bar{x}$, regardless of sample size; and
- if the sample size is large, $\bar{x}$ is approximately normally distributed, regardless of the distribution of x.



Chapter 8 – Confidence Intervals for One Population Mean

Exercise 8.1 – Estimating a Population Mean

DEFINITION 8.2

What Does It Mean?

A confidence-interval estimate for a parameter provides a range of numbers along with a percentage confidence that the parameter lies in that range.

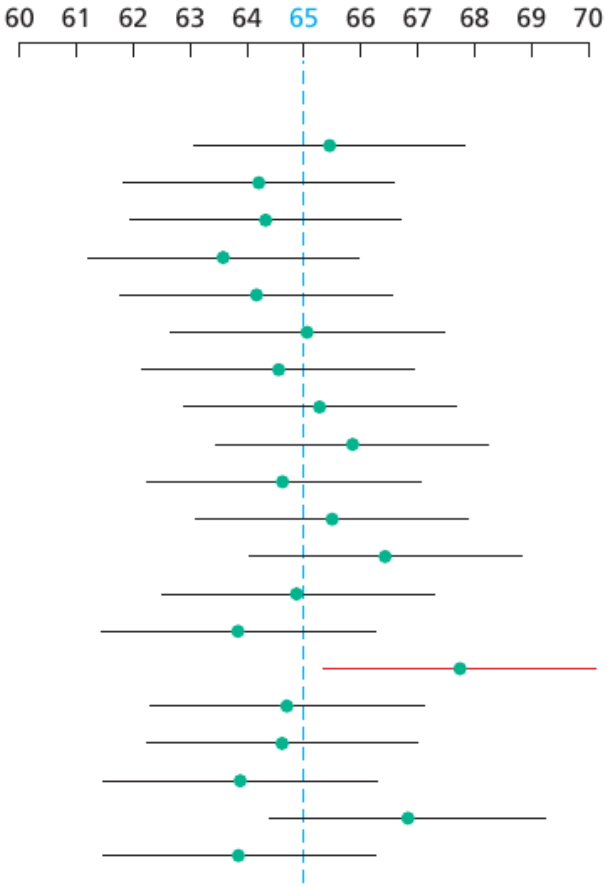
Confidence-Interval Estimate

Confidence interval (CI): An interval of numbers obtained from a point estimate of a parameter.

Confidence level: The confidence we have that the parameter lies in the confidence interval (i.e., that the confidence interval contains the parameter).

Confidence-interval estimate: The confidence level and confidence interval.

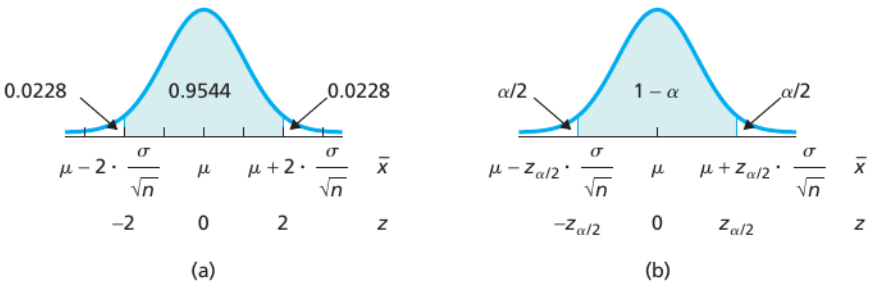
Sample	$\bar{x}$	95.44% CI	μ in CI?
1	65.45	63.06 to 67.85	yes
2	64.21	61.81 to 66.61	yes
3	64.33	61.93 to 66.73	yes
4	63.59	61.19 to 65.99	yes
5	64.17	61.77 to 66.57	yes
6	65.07	62.67 to 67.47	yes
7	64.56	62.16 to 66.96	yes
8	65.28	62.88 to 67.68	yes
9	65.87	63.48 to 68.27	yes
10	64.61	62.22 to 67.01	yes
11	65.51	63.11 to 67.91	yes
12	66.45	64.05 to 68.85	yes
13	64.88	62.48 to 67.28	yes
14	63.85	61.45 to 66.25	yes
15	67.73	65.33 to 70.13	no
16	64.70	62.30 to 67.10	yes
17	64.60	62.20 to 67.00	yes
18	63.88	61.48 to 66.28	yes
19	66.82	64.42 to 69.22	yes
20	63.84	61.45 to 66.24	yes



Exercise 8.2 – Confidence Intervals for One Population Mean When σ Is Known

FIGURE 8.3

(a) 95.44% of all samples have means within 2 standard deviations of μ; (b) 100(1 − α)% of all samples have means within $z_{\alpha/2}$ standard deviations of μ



PROCEDURE 8.1 One-Mean *z*-Interval Procedure

Purpose To find a confidence interval for a population mean, μ

Assumptions

1. Simple random sample
2. Normal population or large sample
3. σ known

Step 1 **For a confidence level of $1 - \alpha$, use Table II to find $z_{\alpha/2}$.**

Step 2 **The confidence interval for μ is from**

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{to} \quad \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is found in Step 1, n is the sample size, and $\bar{x}$ is computed from the sample data.

Step 3 **Interpret the confidence interval.**

Note: The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

Exercise 8.3 – Confidence Intervals for Population Mean When σ Is Unknown

KEY FACT 8.5

›es It Mean?

mally distributed
tudentized
sample mean
ibution with
edom 1 less

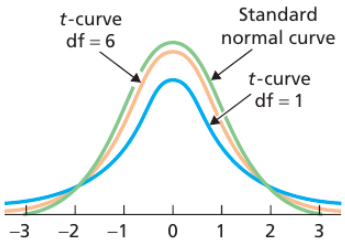
Studentized Version of the Sample Mean

Suppose that a variable x of a population is normally distributed with mean μ. Then, for samples of size n, the variable

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

has the t-distribution with $n - 1$ degrees of freedom.

KEY FACT 8.6



Basic Properties of *t*-Curves

Property 1: The total area under a t-curve equals 1.

Property 2: A t-curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.

Property 3: A t-curve is symmetric about 0.

Property 4: As the number of degrees of freedom becomes larger, t-curves look increasingly like the standard normal curve.

PROCEDURE 8.2 One-Mean *t*-Interval Procedure

Purpose To find a confidence interval for a population mean, μ

Assumptions

1. Simple random sample
2. Normal population or large sample
3. σ unknown

Step 1 **For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n - 1$, where n is the sample size.**

Step 2 **The confidence interval for μ is from**

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},$$

where $t_{\alpha/2}$ is found in Step 1 and $\bar{x}$ and s are computed from the sample data.

Step 3 **Interpret the confidence interval.**

Note: The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

Chapter 9 – Hypothesis Tests for One Population Mean

Exercise 9.1 – The Nature of Hypothesis Testing

DEFINITION 9.1

Null and Alternative Hypotheses; Hypothesis Test

Null hypothesis: A hypothesis to be tested. We use the symbol H_0 to represent the null hypothesis.

Alternative hypothesis: A hypothesis to be considered as an alternative to the null hypothesis. We use the symbol H_a to represent the alternative hypothesis.

Hypothesis test: The problem in a hypothesis test is to decide whether the null hypothesis should be rejected in favor of the alternative hypothesis.

Does It Mean?

ally, the word *null* in esis stood for "no or "the difference is the years, however, esis has come to y a hypothesis to

Basic Logic of Hypothesis Testing

Take a random sample from the population. If the sample data are consistent with the null hypothesis, do not reject the null hypothesis; if the sample data are inconsistent with the null hypothesis and supportive of the alternative hypothesis, reject the null hypothesis in favor of the alternative hypothesis.

DEFINITION 9.2

Type I and Type II Errors

Type I error: Rejecting the null hypothesis when it is in fact true.

Type II error: Not rejecting the null hypothesis when it is in fact false.

Decision:	H_0 is: True	H_0 is: False
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

DEFINITION 9.3

Significance Level

The probability of making a Type I error, that is, of rejecting a true null hypothesis, is called the **significance level, α,** of a hypothesis test.

KEY FACT 9.1

Relation between Type I and Type II Error Probabilities

For a fixed sample size, the smaller we specify the significance level, α, the larger will be the probability, β, of not rejecting a false null hypothesis.

KEY FACT 9.2

Possible Conclusions for a Hypothesis Test

Suppose that a hypothesis test is conducted at a small significance level.

- If the null hypothesis is rejected, we conclude that the data provide sufficient evidence to support the alternative hypothesis.
- If the null hypothesis is not rejected, we conclude that the data do not provide sufficient evidence to support the alternative hypothesis.

Exercise 9.2 – Critical Value Approach to Hypothesis Testing

DEFINITION 9.4

Rejection Region, Nonrejection Region, and Critical Values

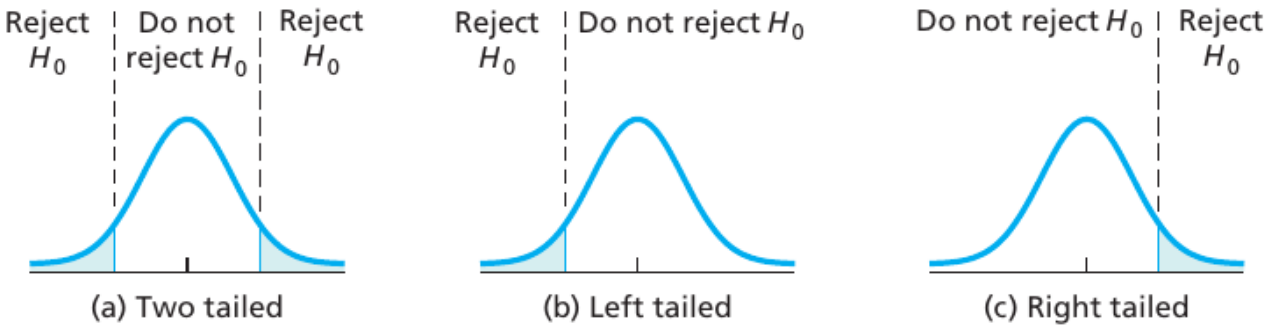
Rejection region: The set of values for the test statistic that leads to rejection of the null hypothesis.

Nonrejection region: The set of values for the test statistic that leads to nonrejection of the null hypothesis.

Critical value(s): The value or values of the test statistic that separate the rejection and nonrejection regions. A critical value is considered part of the rejection region.

Does It Mean?

alue of the test
in the rejection
ct the null
otherwise, do
e null hypothesis.



CRITICAL-VALUE APPROACH TO HYPOTHESIS TESTING

Step 1	State the null and alternative hypotheses.
Step 2	Decide on the significance level, α.
Step 3	Compute the value of the test statistic.
Step 4	Determine the critical value(s).
Step 5	If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.
Step 6	Interpret the result of the hypothesis test.

Exercise 9.3 – P-Value Approach to Hypothesis Testing

DEFINITION 9.5

P-Value

The ***P*-value** of a hypothesis test is the probability of getting sample data at least as inconsistent with the null hypothesis (and supportive of the alternative hypothesis) as the sample data actually obtained.† We use the letter ***P*** to denote the *P*-value.

Does It Mean?

-values provide
ainst the null
larger *P*-values

KEY FACT 9.4

Decision Criterion for a Hypothesis Test Using the *P*-Value

If the *P*-value is less than or equal to the specified significance level, reject the null hypothesis; otherwise, do not reject the null hypothesis. In other words, if $P \le \alpha$, reject H_0; otherwise, do not reject H_0.

KEY FACT 9.5

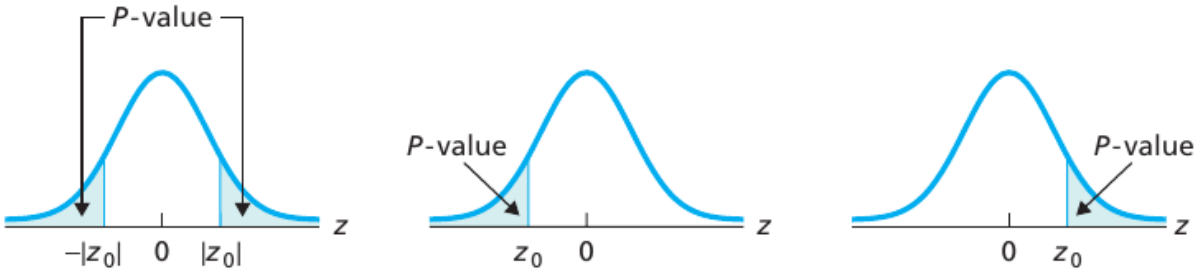
P-Value as the Observed Significance Level

The *P*-value of a hypothesis test equals the smallest significance level at which the null hypothesis can be rejected, that is, the smallest significance level for which the observed semple data results in rejection of H_0.

KEY FACT 9.6

Determining a *P*-Value

To determine the *P*-value of a hypothesis test, we assume that the null hypothesis is true and compute the probability of observing a value of the test statistic as extreme as or more extreme than that observed. By *extreme* we mean "far from what we would expect to observe if the null hypothesis is true."



(a) Two tailed (b) Left tailed (c) Right tailed

P-VALUE APPROACH TO HYPOTHESIS TESTING

Step 1	State the null and alternative hypotheses.
Step 2	Decide on the significance level, α.
Step 3	Compute the value of the test statistic.
Step 4	Determine the P-value, P.
Step 5	If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.
Step 6	Interpret the result of the hypothesis test.

Exercise 9.4 – Hypothesis Intervals for One Population Mean When σ Is Known

PROCEDURE 9.1 One-Mean *z*-Test

Purpose To perform a hypothesis test for a population mean, μ

Assumptions

1. Simple random sample
2. Normal population or large sample
3. σ known

Step 1 **The null hypothesis is H_0: $\mu = \mu_0$, and the alternative hypothesis is**

H_a: $\mu \neq \mu_0$	or	H_a: $\mu < \mu_0$	or	H_a: $\mu > \mu_0$
(Two tailed)		(Left tailed)		(Right tailed)

Step 2 **Decide on the significance level, α.**

Step 3 **Compute the value of the test statistic**

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

and denote that value z_0.

Exercise 9.5 – Hypothesis Intervals for Population Mean When σ Is Unknown

PROCEDURE 9.2 One-Mean *t*-Test

Purpose To perform a hypothesis test for a population mean, μ

Assumptions

1. Simple random sample
2. Normal population or large sample
3. σ unknown

Step 1 **The null hypothesis is H_0: $\mu = \mu_0$, and the alternative hypothesis is**

H_a: $\mu \neq \mu_0$	or	H_a: $\mu < \mu_0$	or	H_a: $\mu > \mu_0$
(Two tailed)		(Left tailed)		(Right tailed)

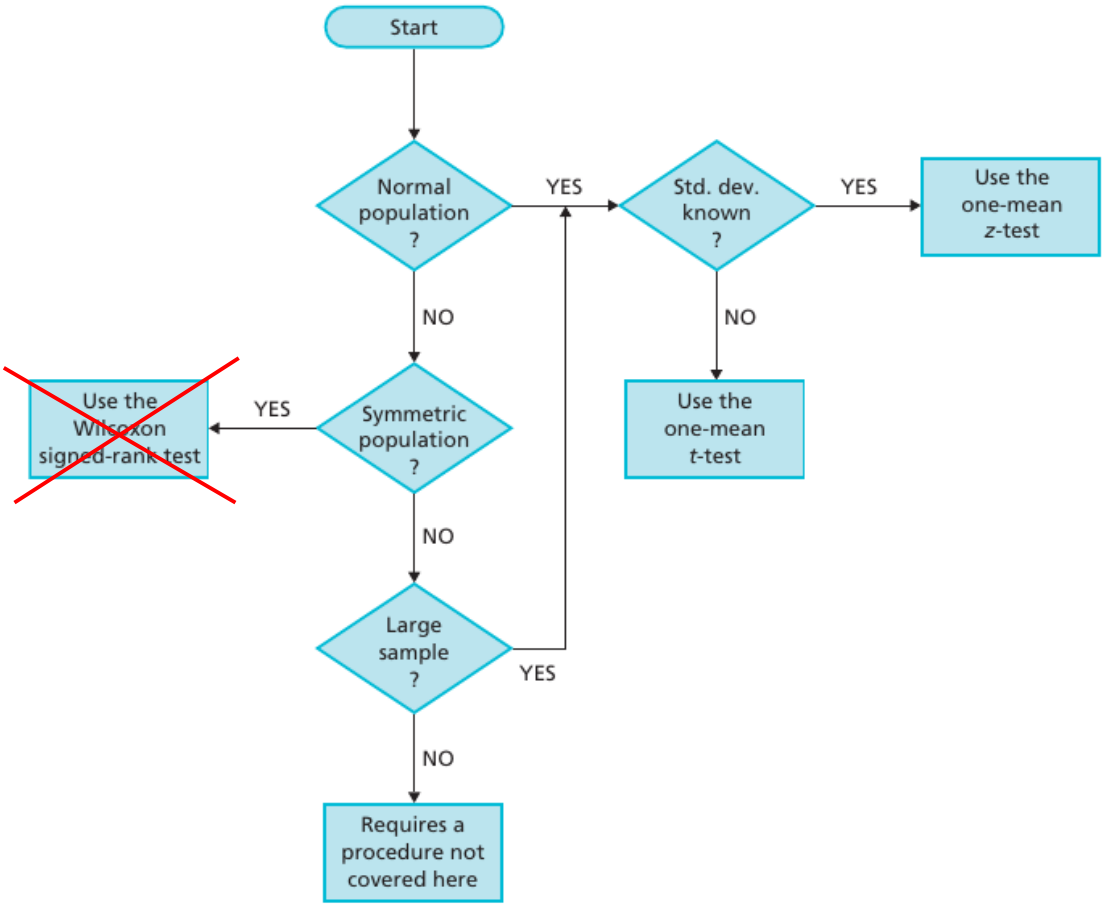
Step 2 **Decide on the significance level, α.**

Step 3 **Compute the value of the test statistic**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and denote that value t_0.

Exercise 9.8 – Which Procedure should be used?



Chapter 10 – Inferences for Two Population Means

Exercise 10.1 – The Sampling Distribution of the Difference between Two Sample Means for Independent Samples

KEY FACT 10.1

The Sampling Distribution of the Difference between Two Sample Means for Independent Samples

Suppose that x is a normally distributed variable on each of two populations. Then, for independent samples of sizes n_1 and n_2 from the two populations,

- $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$,
- $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$, and
- $\bar{x}_1 - \bar{x}_2$ is normally distributed.

Under the conditions of Key Fact 10.1, the standardized version of $\bar{x}_1 - \bar{x}_2$,

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}},$$

Exercise 10.2 – Inferences for Two Population Means, Using Independent Samples: Standard Deviations Assumed Equal

PROCEDURE 10.1 Pooled t-Test

Purpose To perform a hypothesis test to compare two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1 **The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is**

H_a: $\mu_1 \neq \mu_2$ (Two tailed) or H_a: $\mu_1 < \mu_2$ (Left tailed) or H_a: $\mu_1 > \mu_2$ (Right tailed)

Step 2 **Decide on the significance level, α.**

Step 3 **Compute the value of the test statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{(1/n_1) + (1/n_2)}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

Denote the value of the test statistic t_0.

PROCEDURE 10.2 Pooled t-Interval Procedure

Purpose To find a confidence interval for the difference between two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1 **For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $df = n_1 + n_2 - 2$.**

Step 2 **The endpoints of the confidence interval for $\mu_1 - \mu_2$ are**

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot s_p\sqrt{(1/n_1) + (1/n_2)}.$$

Step 3 **Interpret the confidence interval.**

Note: The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

Exercise 10.3 – Inferences for Two Population Means, Using Independent Samples: Standard Deviations Assumed Unequal

KEY FACT 10.3

Distribution of the Nonpooled *t*-Statistic

Suppose that x is a normally distributed variable on each of two populations. Then, for independent samples of sizes n_1 and n_2 from the two populations, the variable

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

has approximately a t-distribution. The degrees of freedom used is obtained from the sample data. It is denoted Δ and given by

$$\Delta = \frac{\left[\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right]^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}},$$

rounded down to the nearest integer.

PROCEDURE 10.3 Nonpooled *t*-Test

Purpose To perform a hypothesis test to compare two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples

Step 1 **The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is**

H_a: $\mu_1 \neq \mu_2$ (Two tailed)	or	H_a: $\mu_1 < \mu_2$ (Left tailed)	or	H_a: $\mu_1 > \mu_2$ (Right tailed)

Step 2 **Decide on the significance level, α.**

Step 3 **Compute the value of the test statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}.$$

Denote the value of the test statistic t_0.

PROCEDURE 10.4 Nonpooled t-Interval Procedure

Purpose To find a confidence interval for the difference between two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples

Step 1 **For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = \Delta$, where**

$$\Delta = \frac{\left[\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right]^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

rounded down to the nearest integer.

Step 2 **The endpoints of the confidence interval for $\mu_1 - \mu_2$ are**

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot \sqrt{\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)}.$$

Step 3 **Interpret the confidence interval.**

Exercise 10.5 – Inferences for Two Population Means, Using Paired (Dependent) Samples

KEY FACT 10.6 **Distribution of the Paired t-Statistic**

Suppose that x is a variable on each of two populations whose members can be paired. Further suppose that the paired-difference variable d is normally distributed. Then, for paired samples of size n, the variable

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)}{s_d/\sqrt{n}}$$

has the t-distribution with $df = n - 1$.

PROCEDURE 10.6 Paired t-Test

Purpose To perform a hypothesis test to compare two population means, μ_1 and μ_2

Assumptions

1. Simple random paired sample
2. Normal differences or large sample

Step 1 **The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is**

H_a: $\mu_1 \neq \mu_2$ (Two tailed) or H_a: $\mu_1 < \mu_2$ (Left tailed) or H_a: $\mu_1 > \mu_2$ (Right tailed)

Step 2 **Decide on the significance level, α.**

Step 3 **Compute the value of the test statistic**

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

and denote that value t_0.

$$\bar{d} = \frac{\Sigma d_i}{n} \qquad s_d = \sqrt{\frac{\Sigma d_i^2 - (\Sigma d_i)^2/n}{n-1}}$$

d is the sample mean of the paired differences.
s_d is the variance of paired differences

PROCEDURE 10.7 Paired t-Interval Procedure

Purpose To find a confidence interval for the difference between two population means, μ_1 and μ_2

Assumptions

1. Simple random paired sample
2. Normal differences or large sample

Step 1 **For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $df = n - 1$.**

Step 2 **The endpoints of the confidence interval for $\mu_1 - \mu_2$ are**

$$\bar{d} \pm t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}.$$

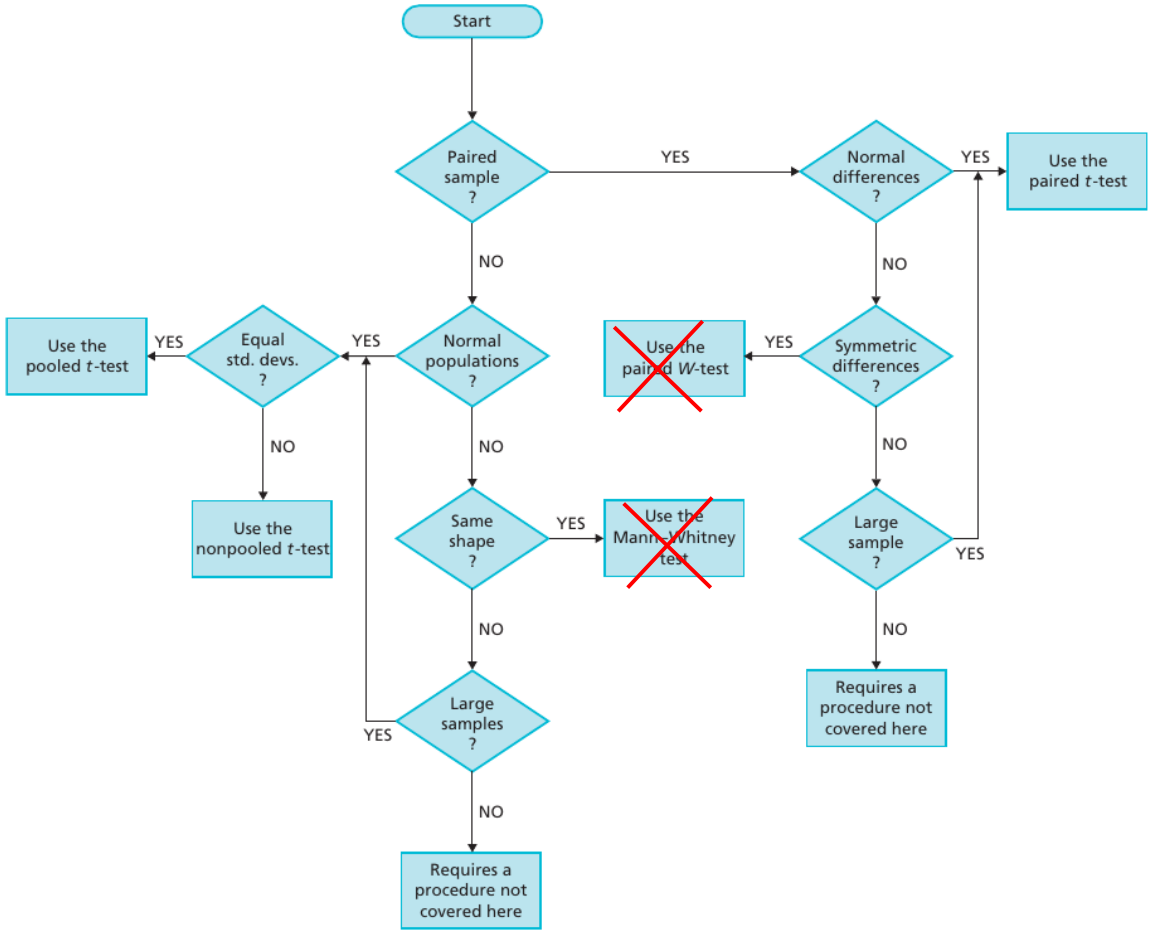
Step 3 **Interpret the confidence interval.**

Note: The confidence interval is exact for normal differences and is approximately correct for large samples and nonnormal differences.

Exercise 10.7 – Which Procedure should be used?

Type	Assumptions	Test statistic	Procedure to use
Pooled t-test	1. Simple random samples 2. Independent samples 3. Normal populations or large samples 4. Equal population standard deviations	$t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{(1/n_1) + (1/n_2)}}$ † $(\text{df} = n_1 + n_2 - 2)$	10.1 (page 441)
Nonpooled t-test	1. Simple random samples 2. Independent samples 3. Normal populations or large samples	$t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$ ‡	10.3 (page 453)
Mann–Whitney test	1. Simple random samples 2. Independent samples 3. Same-shape populations	M = sum of the ranks for sample data from Population 1	10.5 (page 468)
Paired t-test	1. Simple random paired sample 2. Normal differences or large sample	$t = \dfrac{\bar{d}}{s_d/\sqrt{n}}$ $(\text{df} = n - 1)$	10.6 (page 481)
Paired W-test	1. Simple random paired sample 2. Symmetric differences	W = sum of positive ranks	10.8 (page 492)

† $s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ ‡ $\text{df} = \dfrac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$



Chapter 11 – Inferences for Population Standard Deviations

Exercise 11.1 – Inferences for One Population Standard Deviation

A chi-square distribution has the shape of a special type of right-skewed curve, called a chi-square (χ 2) curve. Actually, there are infinitely many chi-square distributions, and we identify them by its number of degrees of freedom, just as we did for t-distributions.

KEY FACT 11.1 **Basic Properties of χ^2-Curves**

Property 1: The total area under a χ^2-curve equals 1.

Property 2: A χ^2-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.

Property 3: A χ^2-curve is right skewed.

Property 4: As the number of degrees of freedom becomes larger, χ^2-curves look increasingly like normal curves.

PROCEDURE 11.1 One-Standard-Deviation χ^2-Test

Purpose To perform a hypothesis test for a population standard deviation, σ

Assumptions

1. Simple random sample
2. Normal population

Step 1 **The null hypothesis is H_0: $\sigma = \sigma_0$, and the alternative hypothesis is**

$$H_a: \sigma \neq \sigma_0 \text{ (Two tailed)} \quad \text{or} \quad H_a: \sigma < \sigma_0 \text{ (Left tailed)} \quad \text{or} \quad H_a: \sigma > \sigma_0 \text{ (Right tailed)}$$

Step 2 **Decide on the significance level, α.**

Step 3 **Compute the value of the test statistic**

$$\chi^2 = \frac{n-1}{\sigma_0^2} s^2$$

and denote that value χ_0^2.

PROCEDURE 11.2 One-Standard-Deviation χ^2-Interval Procedure

Purpose To find a confidence interval for a population standard deviation, σ

Assumptions

1. Simple random sample
2. Normal population

Step 1 **For a confidence level of $1 - \alpha$, use Table VII to find $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ with df $= n - 1$.**

Step 2 **The confidence interval for σ is from**

$$\sqrt{\frac{n-1}{\chi^2_{\alpha/2}}} \cdot s \quad \text{to} \quad \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2}}} \cdot s,$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are found in Step 1, n is the sample size, and s is computed from the sample data obtained.

Step 3 **Interpret the confidence interval.**

Chapter 14 – Descriptive Methods in Regression and Correlation

Exercise 14.2 – The Regression Equation

DEFINITION 14.2

Regression Line and Regression Equation

Regression line: The line that best fits a set of data points according to the least-squares criterion.

Regression equation: The equation of the regression line.

DEFINITION 14.3

Notation Used in Regression and Correlation

For a set of n data points, the defining and computing formulas for S_{xx}, S_{xy}, and S_{yy} are as follows.

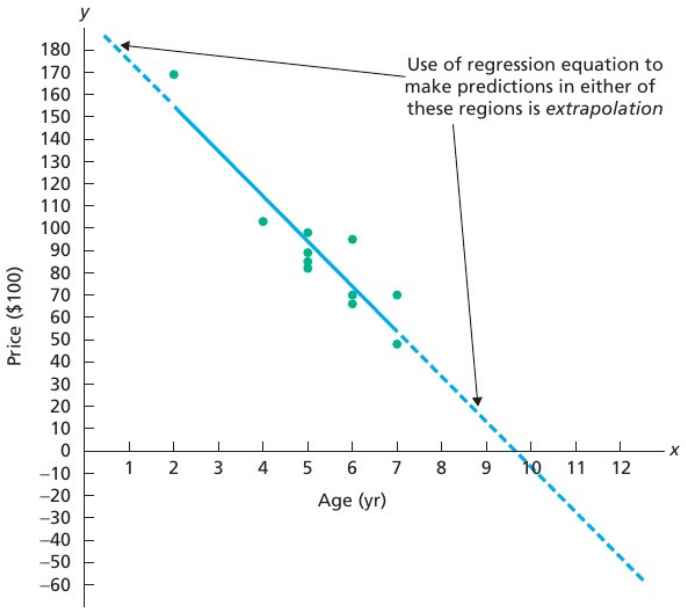
Quantity	Defining formula	Computing formula
S_{xx}	$\Sigma(x_i - \bar{x})^2$	$\Sigma x_i^2 - (\Sigma x_i)^2/n$
S_{xy}	$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n$
S_{yy}	$\Sigma(y_i - \bar{y})^2$	$\Sigma y_i^2 - (\Sigma y_i)^2/n$

FORMULA 14.1

Regression Equation

The regression equation for a set of n data points is $\hat{y} = b_0 + b_1 x$, where

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \frac{1}{n}(\Sigma y_i - b_1 \Sigma x_i) = \bar{y} - b_1 \bar{x}.$$



Exercise 14.3 – The Coefficient of Determination

DEFINITION 14.6

Coefficient of Determination

The **coefficient of determination, r^2,** is the proportion of variation in the observed values of the response variable explained by the regression. Thus,

$$r^2 = \frac{SSR}{SST}.$$

It Does It Mean?

coefficient of
ation is a descriptive
of the utility of the
n equation for making

FORMULA 14.2

Computing Formulas for the Sums of Squares

The computing formulas for the three sums of squares are

$$SST = \Sigma y_i^2 - (\Sigma y_i)^2/n, \quad SSR = \frac{[\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n]^2}{\Sigma x_i^2 - (\Sigma x_i)^2/n},$$

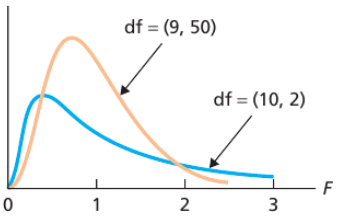
and $SSE = SST - SSR$.

Chapter 16 – Analysis of Variance (ANOVA)

Exercise 16.1 – The F-distribution

An F-distribution has the shape of a special type of right-skewed curve, called an F-curve. Actually, there are infinitely many F-distributions, and we identify them by its number of degrees of freedom, just as we did for t-distributions and chi-square distributions.

Two different *F*-curves



KEY FACT 11.3

Basic Properties of *F*-Curves

Property 1: The total area under an *F*-curve equals 1.

Property 2: An *F*-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.

Property 3: An *F*-curve is right skewed.

Exercise 16.2 – One way ANOVA – The Logic

ANOVA provides methods for comparing several population means, that is, the means of a single variable for several populations.

KEY FACT 16.2 **Assumptions (Conditions) for One-Way ANOVA**

1. **Simple random samples:** The samples taken from the populations under consideration are simple random samples.
2. **Independent samples:** The samples taken from the populations under consideration are independent of one another.
3. **Normal populations:** For each population, the variable under consideration is normally distributed.
4. **Equal standard deviations:** The standard deviations of the variable under consideration are the same for all the populations.

As before, when dealing with several populations, we use subscripts on parameters and statistics. Thus, for Population j, we use μ_j, $\bar{x}_j$, s_j, T_j and n_j to denote the population mean, sample mean, sample standard deviation, sum of all samples, and sample size, respectively.

DEFINITION 16.1 **Mean Squares and *F*-Statistic in One-Way ANOVA**

Treatment mean square, *MSTR*: The variation among the sample means: $MSTR = SSTR/(k-1)$, where $SSTR$ is the treatment sum of squares and k is the number of populations under consideration.

Error mean square, *MSE*: The variation within the samples: $MSE = SSE/(n-k)$, where SSE is the error sum of squares and n is the total number of observations.

***F*-statistic, *F*:** The ratio of the variation among the sample means to the variation within the samples: $F = MSTR/MSE$.

$$SST = \sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2$$

$$SSTR = \sum_{j=1}^{k} \left(\frac{T_j^2}{n_j}\right) - \frac{(\sum x_i)^2}{n}$$

$$SSE = SST - SSTR$$

Source	df	*SS*	*MS* = *SS*/df	*F*-statistic
Treatment	$k-1$	$SSTR$	$MSTR = \frac{SSTR}{k-1}$	$F = \frac{MSTR}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	SST		