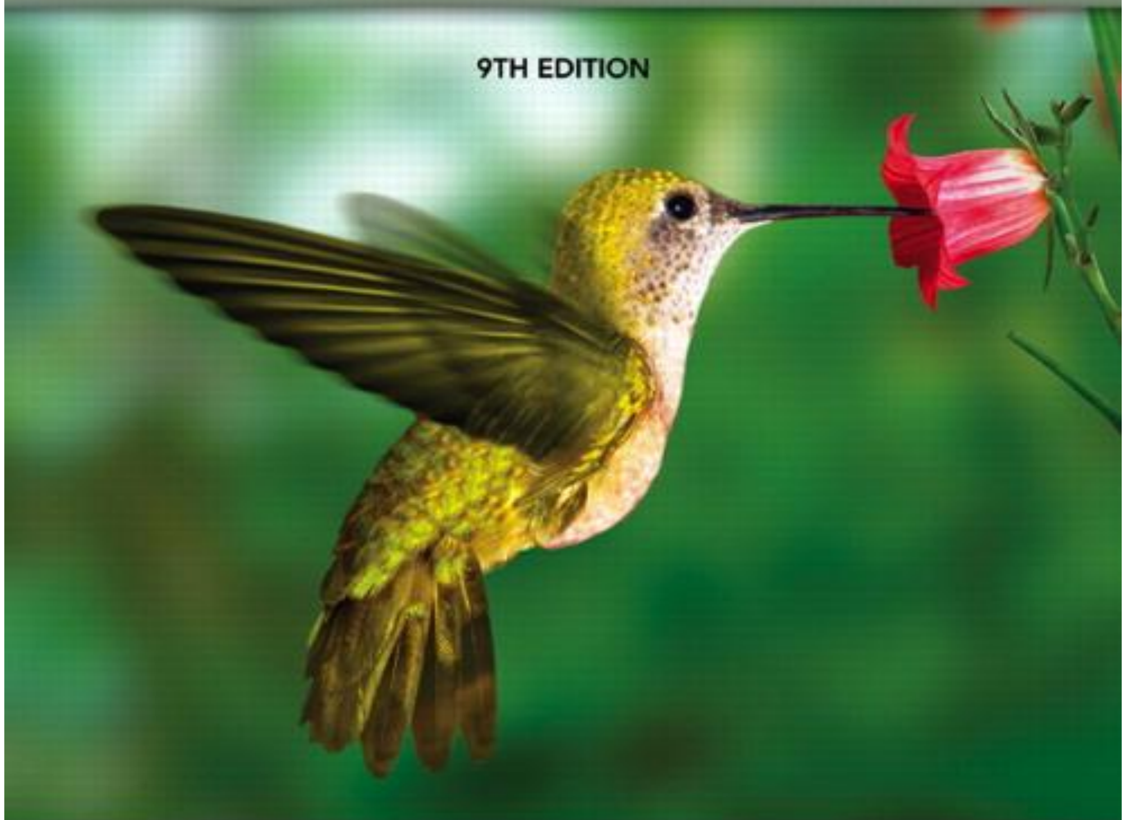
Introductory
STATISTICS



9TH EDITION

Neil
WEISS

Chapter 1

The Nature of Statistics

Section 1.1

Statistics Basics

Definition 1.1

Descriptive Statistics

Descriptive Statistics consists of methods for organizing and summarizing information.

Descriptive statistics includes the construction of graphs, charts, and tables and the calculation of various descriptive measures such as averages, measures of variation, and percentiles.

The 1948 Baseball Season. In 1948, the Washington Senators played 153 games, winning 56 and losing 97. They finished seventh in the American League and were led in hitting by Bud Stewart, whose batting average was .279.

Definition 1.2

Population and Sample

Population: The collection of all individuals or items under consideration in a statistical study.

Sample: That part of the population from which information is obtained.

Political polling provides an example of **inferential statistics**. Interviewing everyone of voting age in the United States on their voting preferences would be expensive and unrealistic. Statisticians who want to gauge the sentiment of the entire **population** of U.S. voters can afford to interview only a carefully chosen group of a few thousand voters. This group is called a **sample** of the **population**.

Definition 1.3

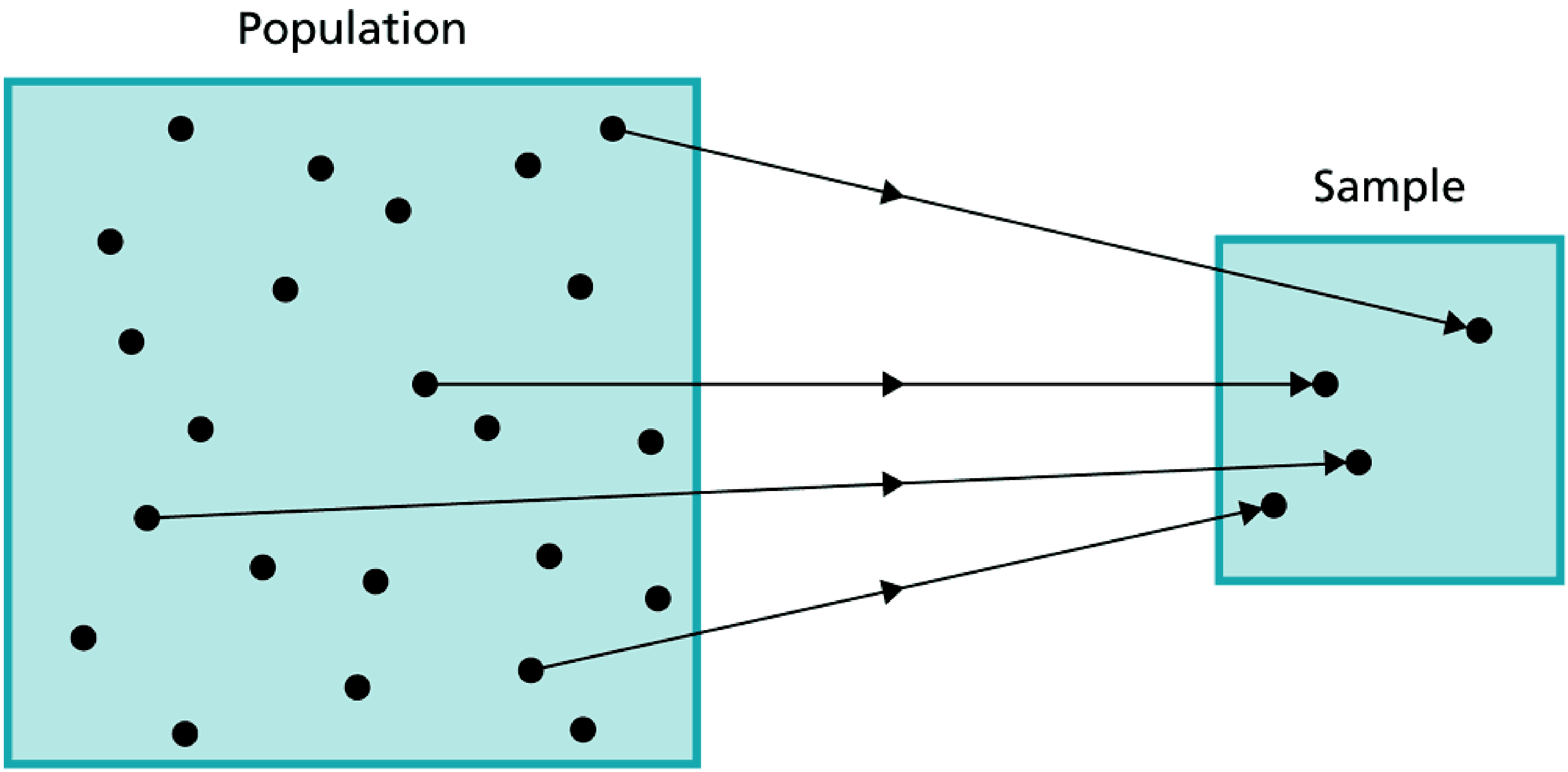
Inferential Statistics

Inferential statistics: consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population.

Statisticians analyze the information obtained from a **sample** of the voting **population** to make **inferences** (draw conclusions) about the preferences of the entire voting **population**. Inferential statistics provides methods for drawing such conclusions.

Figure 1.1

Relationship between population and sample



Section 1.2

Simple Random Sampling

Definition 1.4

Simple Random Sampling; Simple Random Sample

Simple random sampling: A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained.

Simple random sample: A sample obtained by simple random sampling.

There are two types of **simple random sampling**. One is simple random sampling **with replacement**, whereby a member of the population can be selected more than once; the other is simple random sampling **without replacement**, whereby a member of the population can be selected at most once.

Random-Number Tables

Obtaining a simple random sample by picking slips of paper out of a box is usually impractical, especially when the population is large. Fortunately, we can use several practical procedures to get simple random samples. One common method involves a **table of random numbers** – a table of randomly chosen digits, as illustrated in Table 1.5.

Table 1.5

Random numbers

Line number	00–09		10–19		20–29		30–39		40–49	
	Column number									
00	15544	80712	97742	21500	97081	42451	50623	56071	28882	28739
01	01011	21285	04729	39986	73150	31548	30168	76189	56996	19210
02	47435	53308	40718	29050	74858	64517	93573	51058	68501	42723
03	91312	75137	86274	59834	69844	19853	06917	17413	44474	86530
04	12775	08768	80791	16298	22934	09630	98862	39746	64623	32768
05	31466	43761	94872	92230	52367	13205	38634	55882	77518	36252
06	09300	43847	40881	51243	97810	18903	53914	31688	06220	40422
07	73582	13810	57784	72454	68997	72229	30340	08844	53924	89630
08	11092	81392	58189	22697	41063	09451	09789	00637	06450	85990
09	93322	98567	00116	35605	66790	52965	62877	21740	56476	49296
10	80134	12484	67089	08674	70753	90959	45842	59844	45214	36505
11	97888	31797	95037	84400	76041	96668	75920	68482	56855	97417
12	92612	27082	59459	69380	98654	20407	88151	56263	27126	63797
13	72744	45586	43279	44218	83638	05422	00995	70217	78925	39097
14	96256	70653	45285	26293	78305	80252	03625	40159	68760	84716
15	07851	47452	66742	83331	54701	06573	98169	37499	67756	68301
16	25594	41552	96475	56151	02089	33748	65289	89956	89559	33687
17	65358	15155	59374	80940	03411	94656	69440	47156	77115	99463
18	09402	31008	53424	21928	02198	61201	02457	87214	59750	51330
19	97424	90765	01634	37328	41243	33564	17884	94747	93650	77668

Random-Number Generators

Nowadays, statisticians prefer statistical software packages or graphing calculators, rather than random-number tables, to obtain simple random samples. The built-in programs for doing so are called **random-number generators**. When using random-number generators, be aware of whether they provide samples with replacement or samples without replacement.

Section 1.3

Other Sampling Designs

Procedure 1.1

Systematic Random Sampling

Step 1 Divide the population size by the sample size and round the result down to the nearest whole number, m.

Step 2 Use a random-number table or a similar device to obtain a number, k, between 1 and m.

Step 3 Select for the sample those members of the population that are numbered $k, k + m, k + 2m, \ldots$.

Procedure 1.2

Cluster Sampling

Step 1 Divide the population into groups (clusters).

Step 2 Obtain a simple random sample of the clusters.

Step 3 Use all the members of the clusters obtained in Step 2 as the sample.

Procedure 1.3

Stratified Random Sampling with Proportional Allocation

Step 1 Divide the population into subpopulations (strata).

Step 2 From each stratum, obtain a simple random sample of size proportional to the size of the stratum; that is, the sample size for a stratum equals the total sample size times the stratum size divided by the population size.

Step 3 Use all the members obtained in Step 2 as the sample.

Section 1.4

Experimental Designs

Definition 1.5

Experimental Units; Subjects

In a designed experiment, the individuals or items on which the experiment is performed are called **experimental units.** When the experimental units are humans, the term **subject** is often used in place of experimental unit.

Folic Acid and Birth Defects. For the study, the doctors enrolled 4753 women prior to conception, and divided them randomly into two groups. One group took daily multivitamins containing 0.8 mg of folic acid, whereas the other group received only trace elements. In the language of experimental design, each woman in the folic acid study is an experimental unit, or a subject.

Key Fact 1.1

Principles of Experimental Design

The following principles of experimental design enable a researcher to conclude that differences in the results of an experiment not reasonably attributable to chance are likely caused by the treatments.

- **Control:** Two or more treatments should be compared.
- **Randomization:** The experimental units should be randomly divided into groups to avoid unintentional selection bias in constituting the groups.
- **Replication:** A sufficient number of experimental units should be used to ensure that randomization creates groups that resemble each other closely and to increase the chances of detecting any differences among the treatments.

Folic Acid and Birth Defects

- ***Control***: The doctors compared the rate of major birth defects for the women who took folic acid to that for the women who took only trace elements.

- ***Randomization***: The women were divided randomly into two groups to avoid unintentional selection bias.

- ***Replication***: A large number of women were recruited for the study to make it likely that the two groups created by randomization would be similar and also to increase the chances of detecting any effect due to the folic acid.

Folic Acid and Birth Defects

One of the most common experimental situations involves a specified treatment and placebo, an inert or innocuous medical substance. Technically, both the specified treatment and placebo are treatments. The group receiving the specified treatment is called the **treatment group**, and the group receiving placebo is called the **control group**. In the folic acid study, the women who took folic acid constituted the **treatment group** and those who took only trace elements constituted the **control group**.

Definition 1.6

Response Variable, Factors, Levels, and Treatments

Response variable: The characteristic of the experimental outcome that is to be measured or observed.

Factor: A variable whose effect on the response variable is of interest in the experiment.

Levels: The possible values of a factor.

Treatment: Each experimental condition. For one-factor experiments, the treatments are the levels of the single Factor. For multifactor experiments, each treatment is a Combination of levels of the factors.

Example 1.12 **Experimental Design**

Weight Gain of Golden Torch Cacti

The Golden Torch Cactus (Trichocereus spachianus), a cactus native to Argentina, has excellent landscape potential. William Feldman and Frank Crosswhite, two researchers at the Boyce Thompson Southwestern Arboretum, investigated the optimal method for producing these cacti. The researchers examined, among other things, the effects of a hydrophilic polymer and irrigation regime on weight gain. Hydrophilic polymers are used as soil additives to keep moisture in the root zone. For this study, the researchers chose Broadleaf P-4 polyacrylamide, abbreviated P4. The hydrophilic polymer was either used or not used, and five irrigation regimes were employed: none, light, medium, heavy, and very heavy.

Example 1.12 Experimental Design
Weight Gain of Golden Torch Cacti

Identify the

a. experimental units.
b. response variable.
c. factors.
d. levels of each factor.
e. treatments.

Example 1.12 **Experimental Design**

Weight Gain of Golden Torch Cacti

Solution

a. The experimental units are the cacti used in the study.
b. The response variable is weight gain.
c. The factors are hydrophilic polymer and irrigation regime.
d. Hydrophilic polymer has two levels: with and without. Irrigation regime has five levels: none, light, medium, heavy, and very heavy.
e. Each treatment is a combination of a level of hydrophilic polymer and a level of irrigation regime. Table 1.8 depicts the 10 treatments for this experiment. In the table, we abbreviated “very heavy” as “Xheavy.”

Table 1.8

Schematic for the 10 treatments in the cactus study

		Irrigation regime				
		None	Light	Medium	Heavy	Xheavy
Polymer	No P4	No water No P4 (Treatment 1)	Light water No P4 (Treatment 2)	Medium water No P4 (Treatment 3)	Heavy water No P4 (Treatment 4)	Xheavy water No P4 (Treatment 5)
	With P4	No water With P4 (Treatment 6)	Light water With P4 (Treatment 7)	Medium water With P4 (Treatment 8)	Heavy water With P4 (Treatment 9)	Xheavy water With P4 (Treatment 10)

Definition 1.7

Completely Randomized Design

In a **completely randomized design,** all the experimental units are assigned randomly among all the treatments.

Once we have chosen the treatments, we must decide how the experimental units are to be assigned to the treatments (or vice versa). The women in the folic acid study were randomly divided into two groups; one group received folic acid and the other only trace elements. In the cactus study, 40 cacti were divided randomly into 10 groups of four cacti each and then each group was assigned a different treatment from among the 10 depicted in Table 1.8. Both of these experiments used a **completely randomized design**.

Definition 1.8

Randomized Block Design

In a **randomized block design,** the experimental units are assigned randomly among all the treatments separately within each block.

Although the completely randomized design is commonly used and simple, it is not always the best design. Several alternatives to that design exist. For instance, in a **randomized block design**, experimental units that are similar in ways that are expected to affect the response variable are grouped in **blocks**. Then the random assignment of experimental units to the treatments is made block by block.

Example 1.13 Statistical Designs

Golf Ball Driving Distances

Suppose we want to compare the driving distances for five different brands of golf ball. For 40 golfers, discuss a method of comparison based on

a. a completely randomized design.
b. a randomized block design.

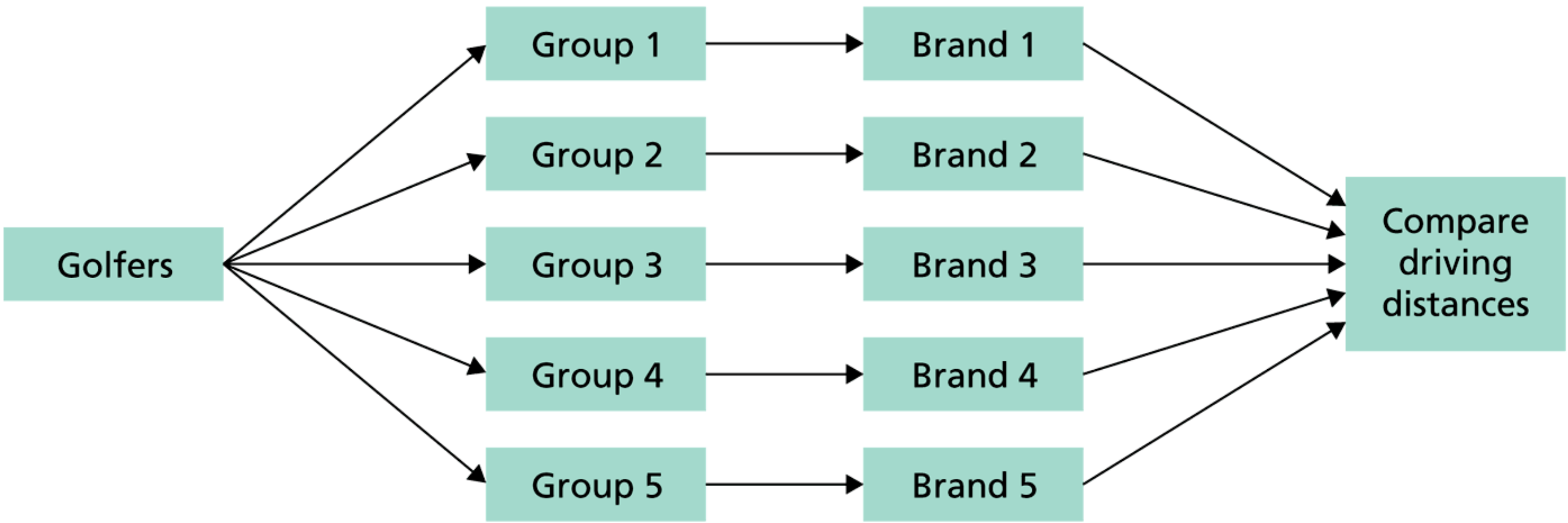
Solution

Here the experimental units are the golfers, the response variable is driving distance, the factor is brand of golf ball, and the levels (and treatments) are the five brands.

a. For a completely randomized design, we would randomly divide the 40 golfers into five groups of 8 golfers each and then randomly assign each group to drive a different brand of ball,as illustrated in Fig.1.5.

Figure 1.5

Completely randomized design for golf ball experiment



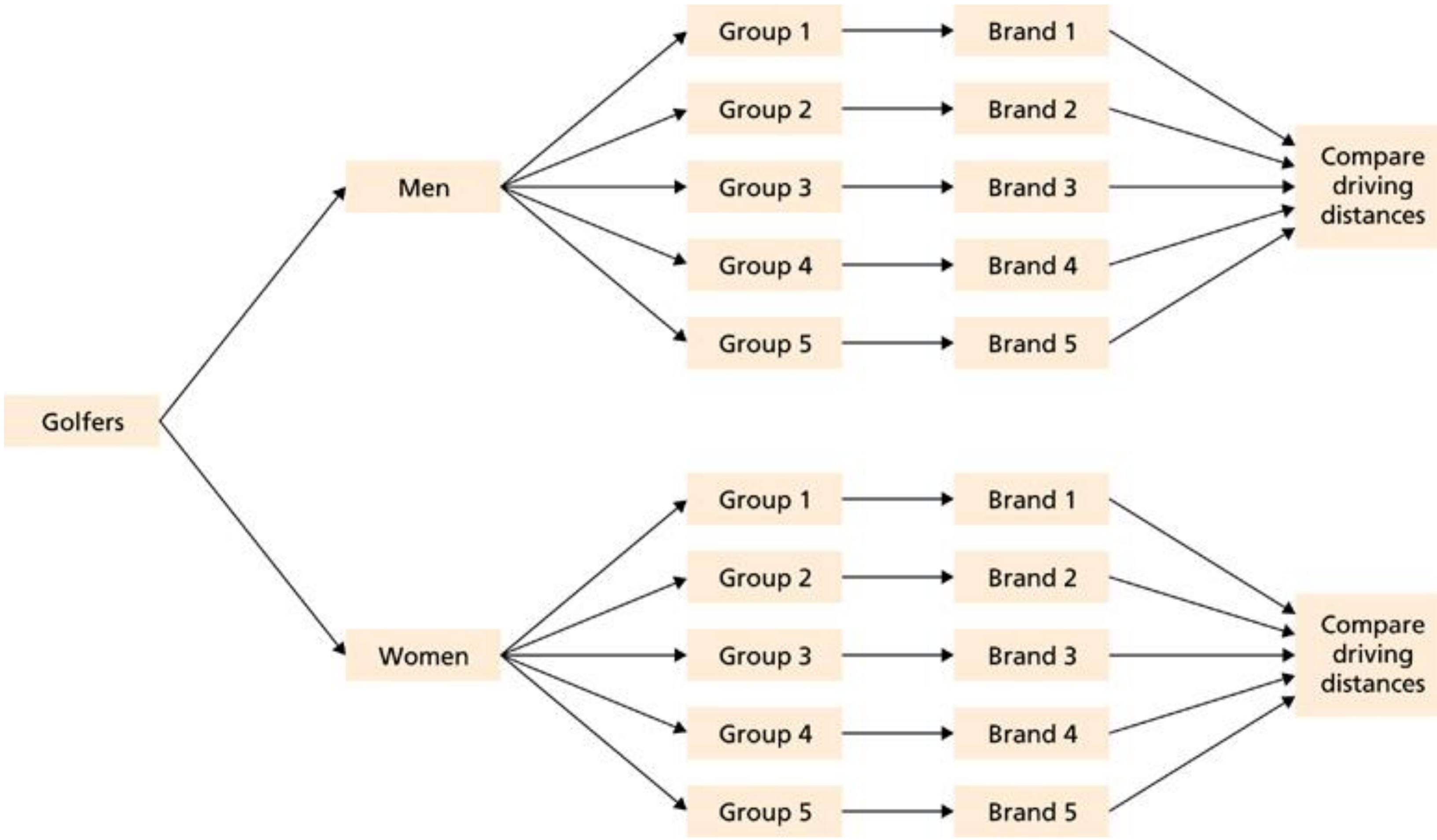
Example 1.13 Statistical Designs

Golf Ball Driving Distances

b. Because driving distance is affected by gender, using a randomized block design that blocks by gender is probably a better approach. We could do so by using 20 men golfers and 20 women golfers. We would randomly divide the 20 men into five groups of 4 men each and then randomly assign each group to drive a different brand of ball, as shown in Fig.1.6. Likewise, we would randomly divide the 20 women into five groups of 4 women each and then randomly assign each group to drive a different brand of ball, as also shown in Fig.1.6.

Figure 1.6

Randomized block design for golf ball experiment



By blocking, we can isolate and remove the variation in driving distances between men and women and thereby make it easier to detect any differences in driving distances among the five brands of golf ball. Additionally, blocking permits us to analyze separately the differences in driving distances among the five brands for men and women.

As illustrated in Example 1.13, blocking can isolate and remove systematic differences among blocks, thereby making any differences among treatments easier to detect. Blocking also makes possible the separate analysis of treatment effects on each block.