POPULATIONS AND SAMPLES (DA_2022)

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

DA_2021 Lecture n. 3, Rome 9th march 2021

DIPARTIMENTO DI FISICA

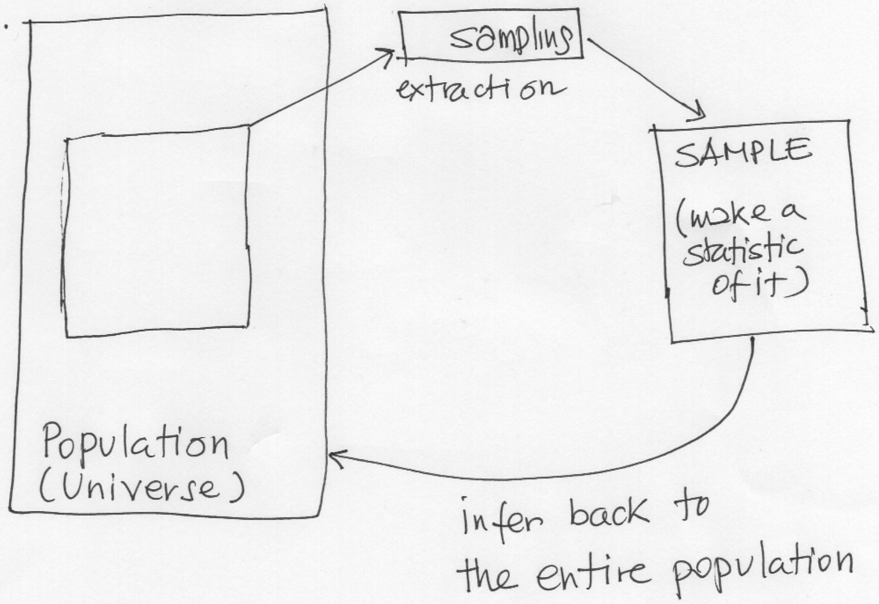
SAPIENZA
UNIVERSITÀ DI ROMA

KEYWORDS OF THIS LECTURE

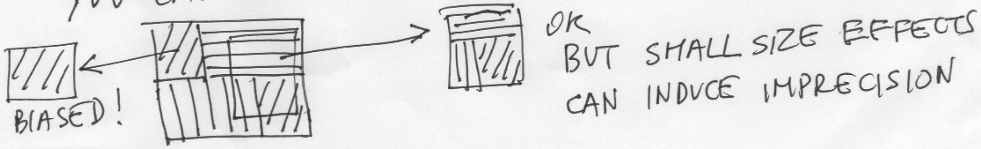
- Population
- Sample
- Parameter
- Statistic
- Estimation
- Random sampling
- Histograms -> probability distributions

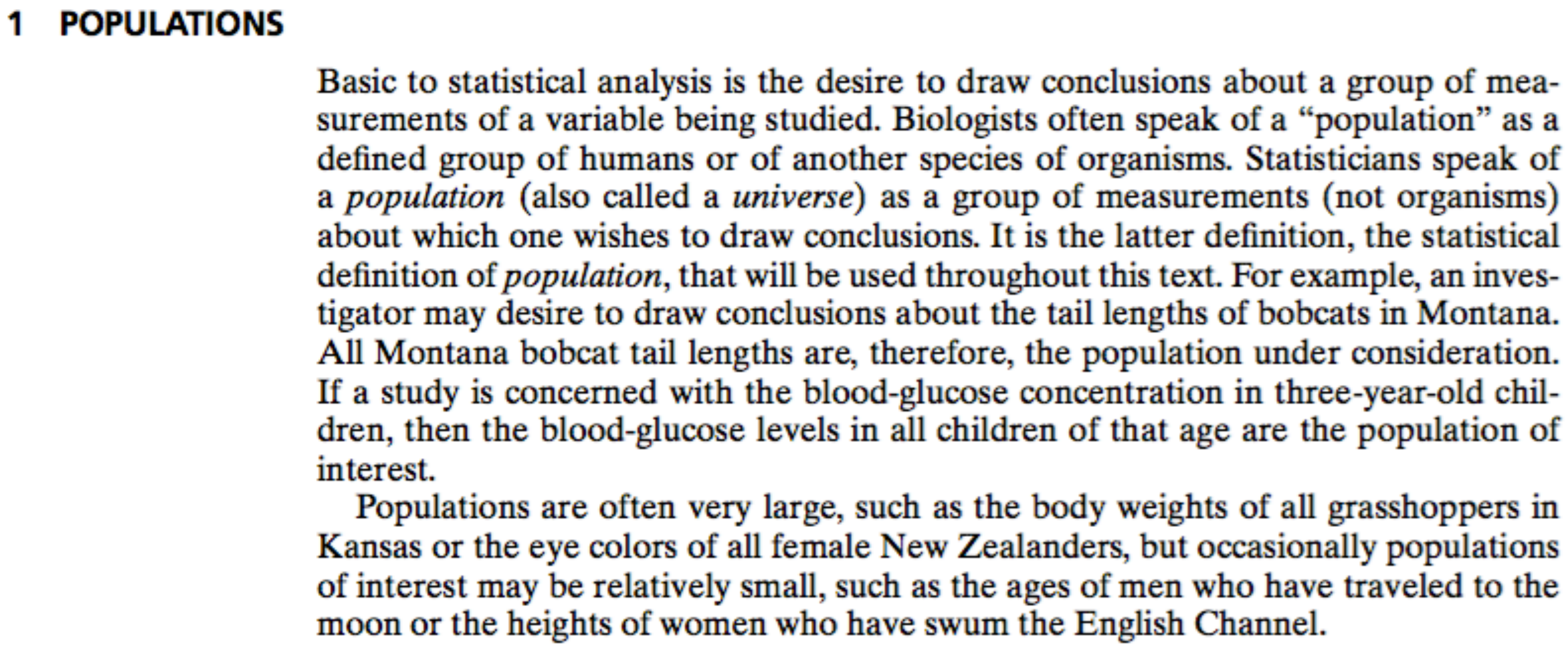
See: R 6.1,6.2, 6.3, 6.4, and also WS Chap.1

WHAT IS STATISTICS ALL ABOUT



OBS. IF THE POPULATION is WELL MIXED, HOMOGENEOUS (Physicists say that the population is 'ERGODIC'), WHATEVER SAMPLE IS GOOD. IF FOR DIFFERENT REASONS THE POPULATION IS NOT UNIFORM YOU CAN GET A BIASED SAMPLE



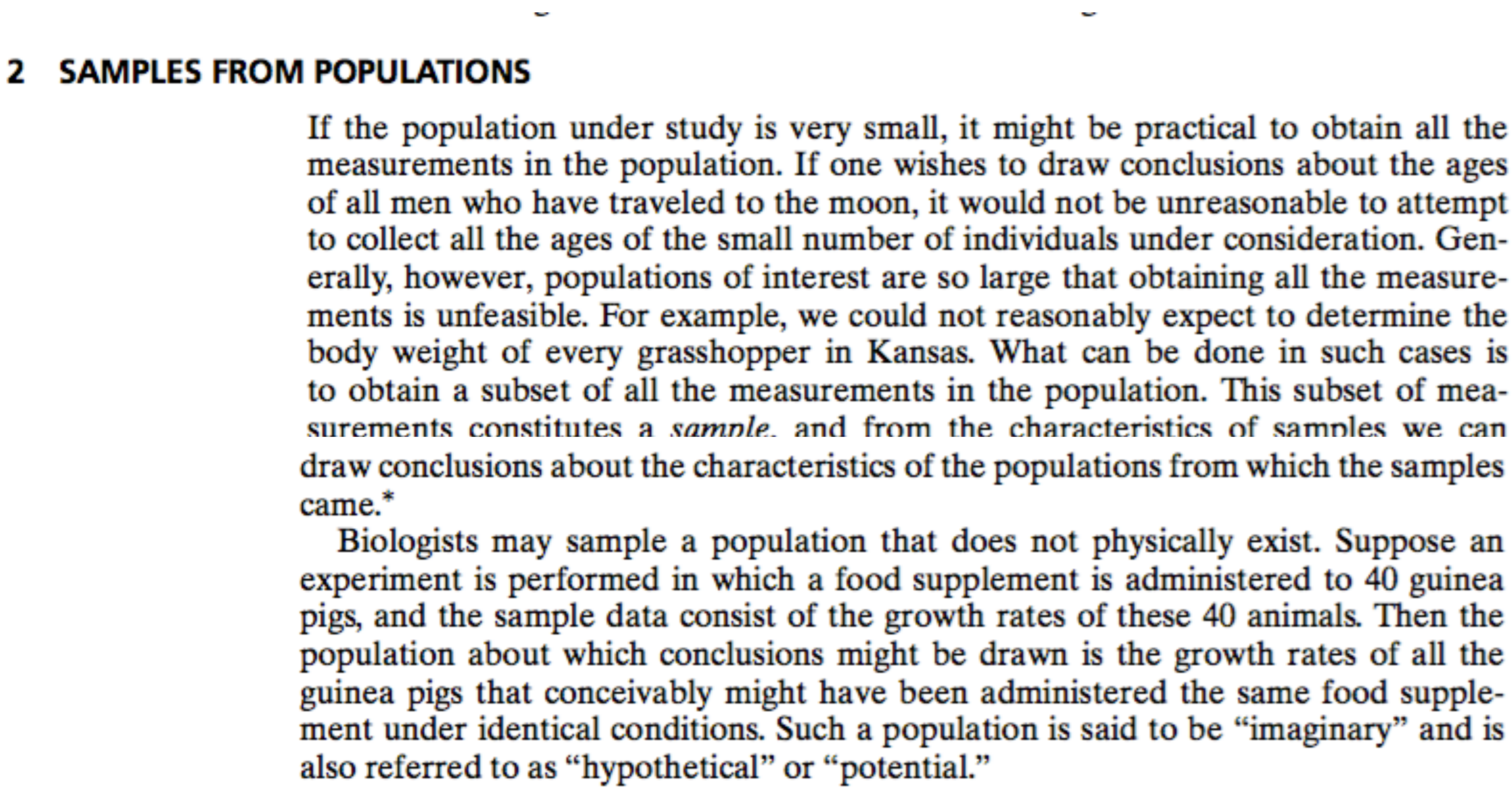


1 POPULATIONS

Basic to statistical analysis is the desire to draw conclusions about a group of measurements of a variable being studied. Biologists often speak of a "population" as a defined group of humans or of another species of organisms. Statisticians speak of a *population* (also called a *universe*) as a group of measurements (not organisms) about which one wishes to draw conclusions. It is the latter definition, the statistical definition of *population*, that will be used throughout this text. For example, an investigator may desire to draw conclusions about the tail lengths of bobcats in Montana. All Montana bobcat tail lengths are, therefore, the population under consideration. If a study is concerned with the blood-glucose concentration in three-year-old children, then the blood-glucose levels in all children of that age are the population of interest.

Populations are often very large, such as the body weights of all grasshoppers in Kansas or the eye colors of all female New Zealanders, but occasionally populations of interest may be relatively small, such as the ages of men who have traveled to the moon or the heights of women who have swum the English Channel.

From:[**JHZ**] Jerrold H Zar - Biostatistical Analysis_ Pearson New International Edition-Pearson (2014).

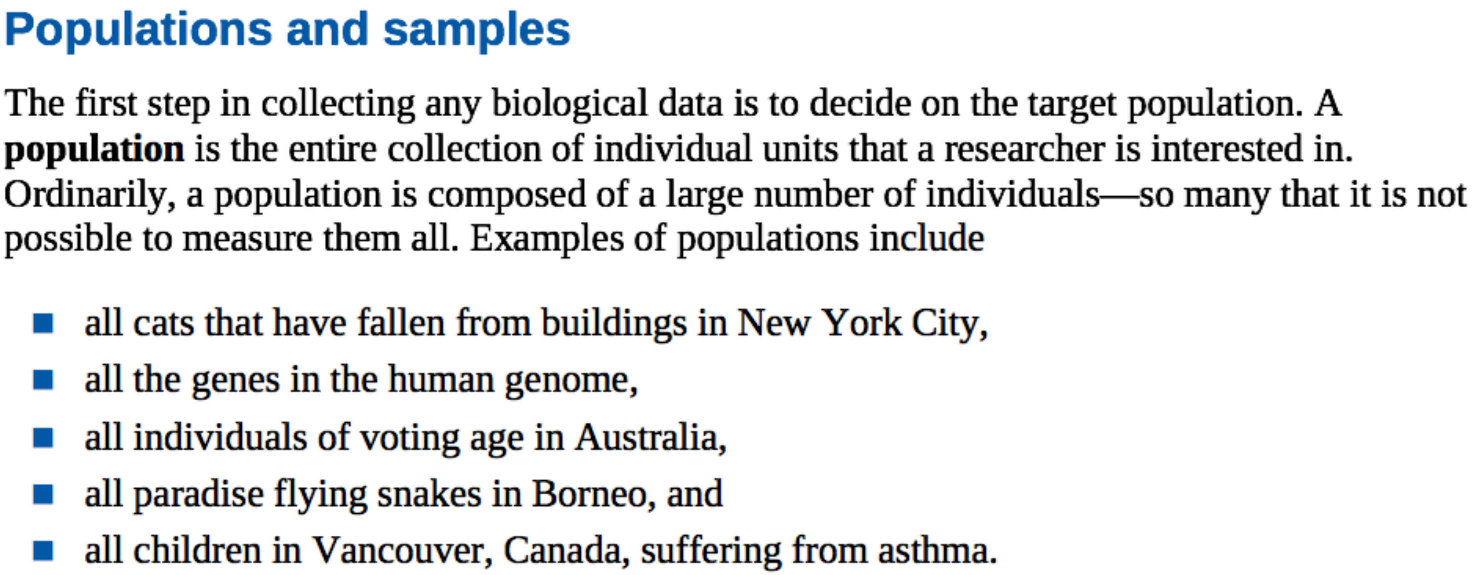


2 SAMPLES FROM POPULATIONS

If the population under study is very small, it might be practical to obtain all the measurements in the population. If one wishes to draw conclusions about the ages of all men who have traveled to the moon, it would not be unreasonable to attempt to collect all the ages of the small number of individuals under consideration. Generally, however, populations of interest are so large that obtaining all the measurements is unfeasible. For example, we could not reasonably expect to determine the body weight of every grasshopper in Kansas. What can be done in such cases is to obtain a subset of all the measurements in the population. This subset of measurements constitutes a *sample*, and from the characteristics of samples we can draw conclusions about the characteristics of the populations from which the samples came.*

Biologists may sample a population that does not physically exist. Suppose an experiment is performed in which a food supplement is administered to 40 guinea pigs, and the sample data consist of the growth rates of these 40 animals. Then the population about which conclusions might be drawn is the growth rates of all the guinea pigs that conceivably might have been administered the same food supplement under identical conditions. Such a population is said to be “imaginary” and is also referred to as “hypothetical” or “potential.”

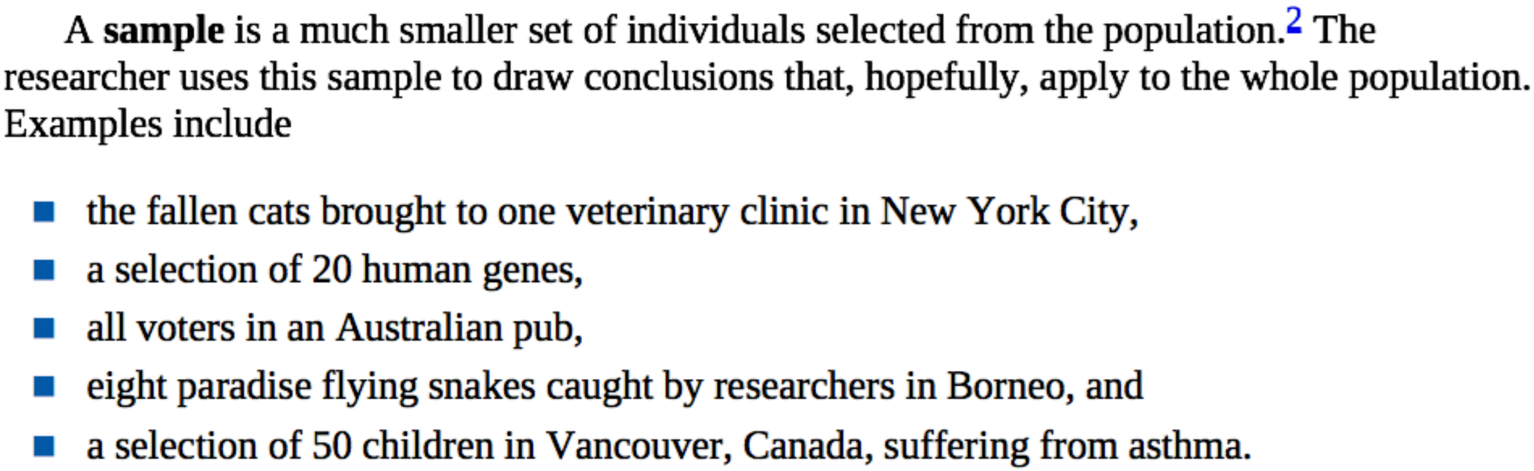
From: Jerrold H Zar - Biostatistical Analysis_ Pearson New International Edition-Pearson (2014).



Populations and samples

The first step in collecting any biological data is to decide on the target population. A **population** is the entire collection of individual units that a researcher is interested in. Ordinarily, a population is composed of a large number of individuals—so many that it is not possible to measure them all. Examples of populations include

- all cats that have fallen from buildings in New York City,
- all the genes in the human genome,
- all individuals of voting age in Australia,
- all paradise flying snakes in Borneo, and
- all children in Vancouver, Canada, suffering from asthma.

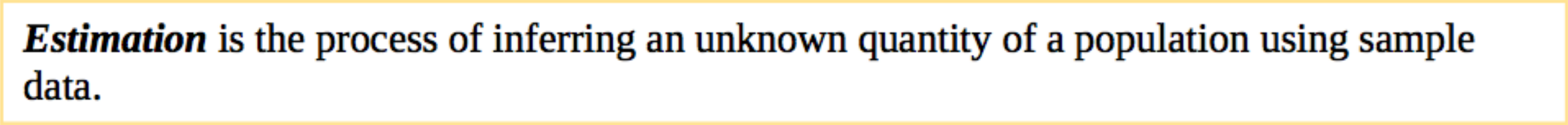


A **sample** is a much smaller set of individuals selected from the population.[2] The researcher uses this sample to draw conclusions that, hopefully, apply to the whole population. Examples include

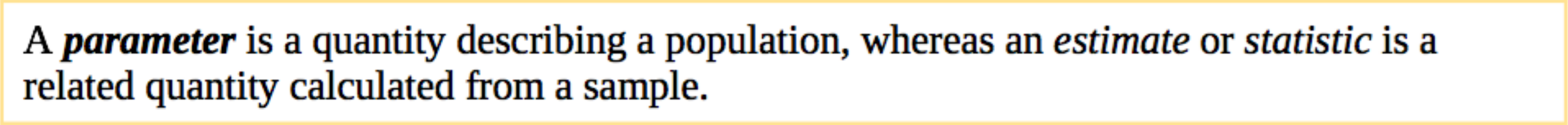
- the fallen cats brought to one veterinary clinic in New York City,
- a selection of 20 human genes,
- all voters in an Australian pub,
- eight paradise flying snakes caught by researchers in Borneo, and
- a selection of 50 children in Vancouver, Canada, suffering from asthma.

From:[**WS**] M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015).

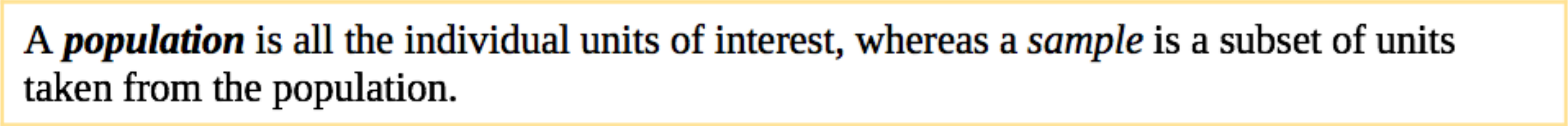
KEY CONCEPTS



Estimation is the process of inferring an unknown quantity of a population using sample data.

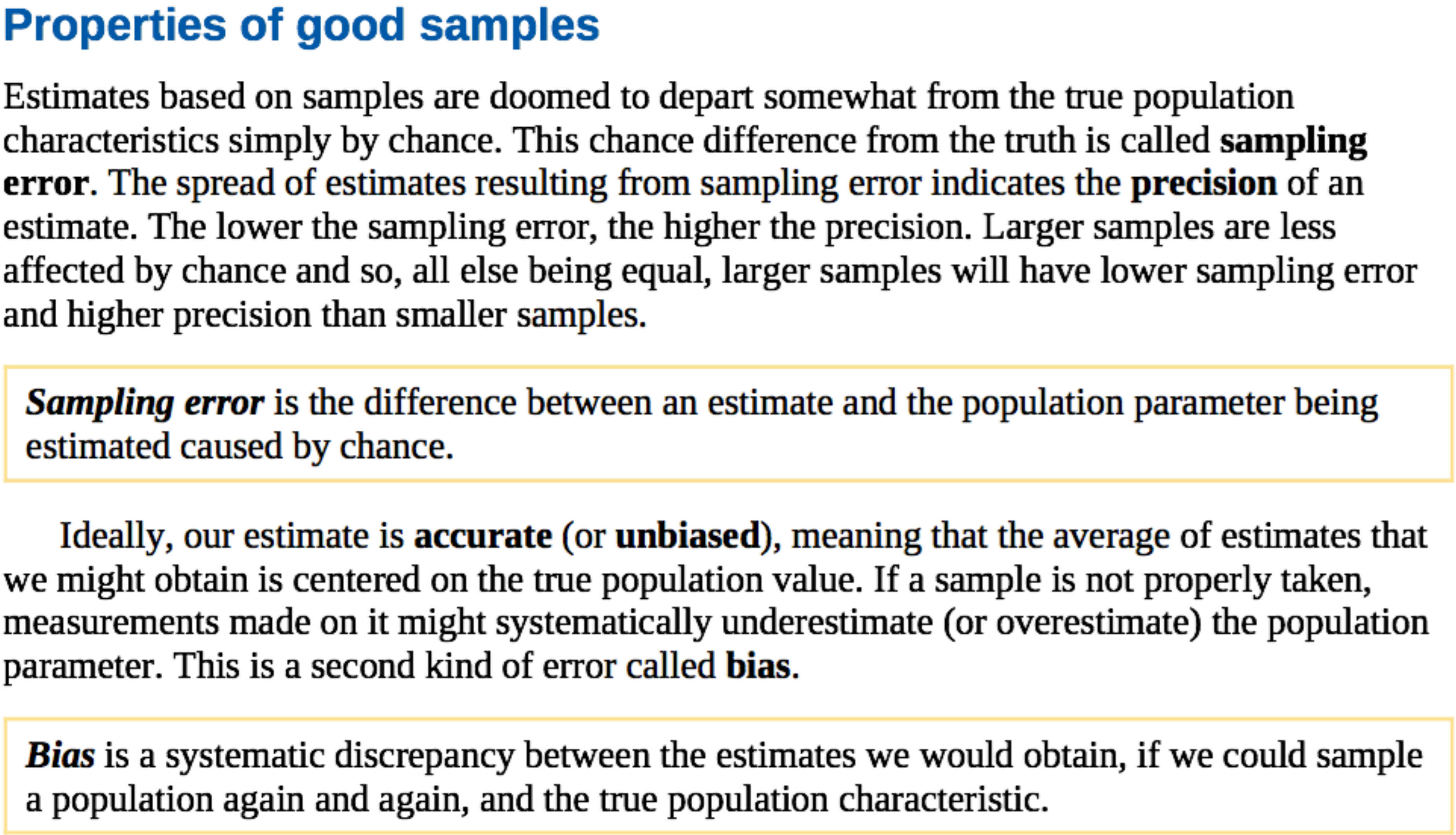


A ***parameter*** is a quantity describing a population, whereas an *estimate* or *statistic* is a related quantity calculated from a sample.



A ***population*** is all the individual units of interest, whereas a *sample* is a subset of units taken from the population.

From:[**WS**] M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015).



Properties of good samples

Estimates based on samples are doomed to depart somewhat from the true population characteristics simply by chance. This chance difference from the truth is called **sampling error**. The spread of estimates resulting from sampling error indicates the **precision** of an estimate. The lower the sampling error, the higher the precision. Larger samples are less affected by chance and so, all else being equal, larger samples will have lower sampling error and higher precision than smaller samples.

> ***Sampling error*** is the difference between an estimate and the population parameter being estimated caused by chance.

Ideally, our estimate is **accurate** (or **unbiased**), meaning that the average of estimates that we might obtain is centered on the true population value. If a sample is not properly taken, measurements made on it might systematically underestimate (or overestimate) the population parameter. This is a second kind of error called **bias**.

> ***Bias*** is a systematic discrepancy between the estimates we would obtain, if we could sample a population again and again, and the true population characteristic.

From:[**WS**] M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015).

Unbiased

Biased

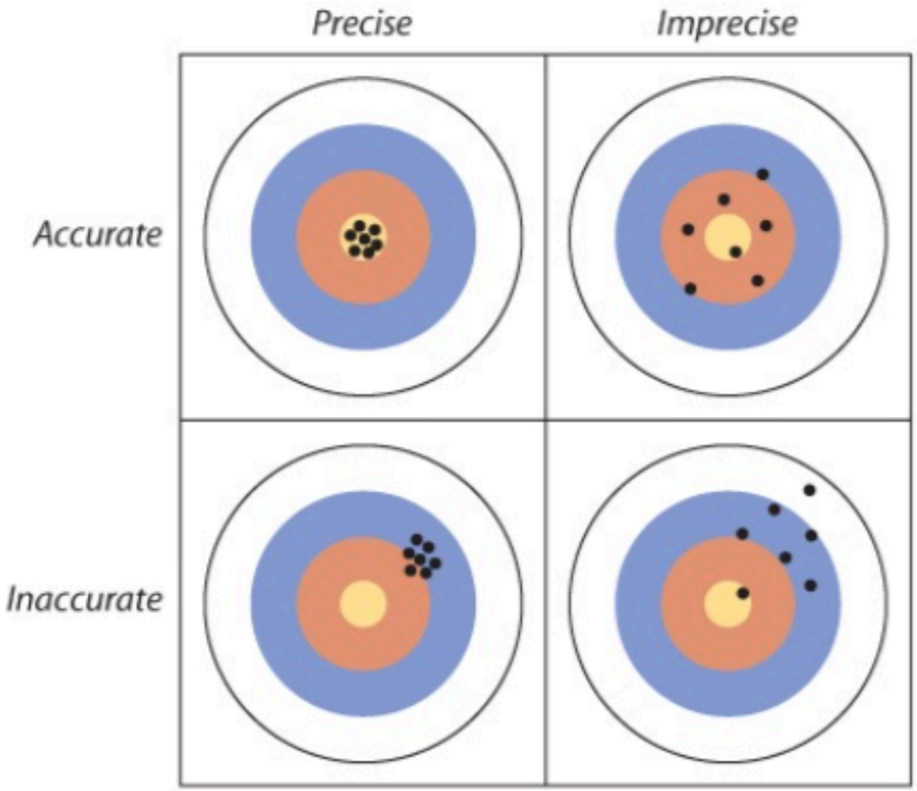
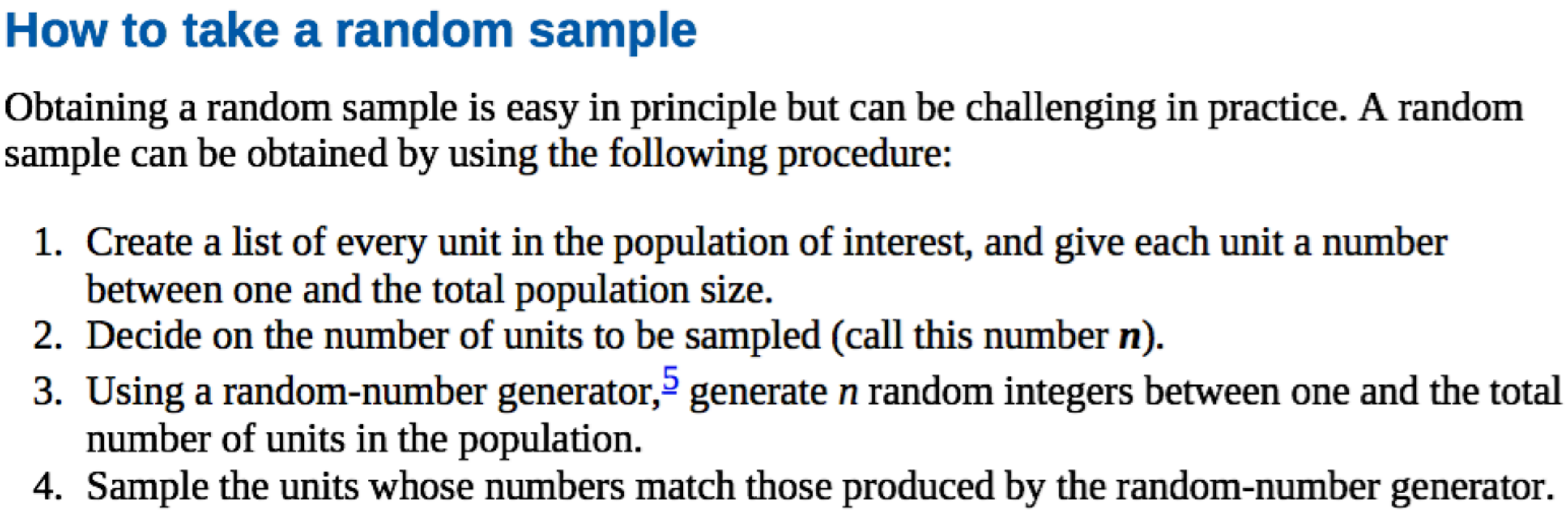


Figure 1.2-2
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015
W. H. Freeman and Company

FIGURE 1.2-2 Analogy between estimation and target shooting. An accurate estimate is centered around the bull's-eye, whereas a precise estimate has low spread.

From:[**WS**] M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015).

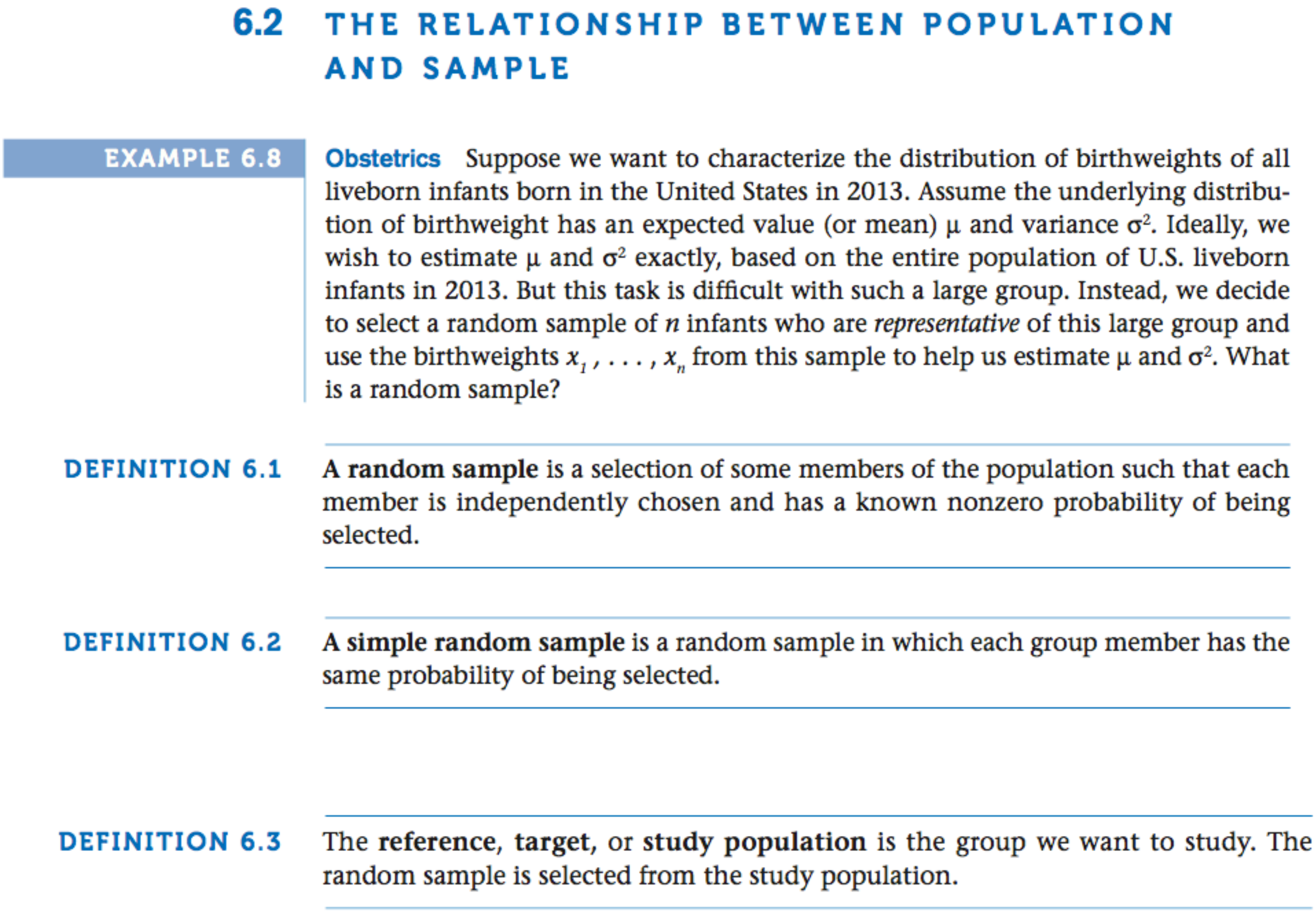


How to take a random sample

Obtaining a random sample is easy in principle but can be challenging in practice. A random sample can be obtained by using the following procedure:

1. Create a list of every unit in the population of interest, and give each unit a number between one and the total population size.
2. Decide on the number of units to be sampled (call this number ***n***).
3. Using a random-number generator,[5] generate *n* random integers between one and the total number of units in the population.
4. Sample the units whose numbers match those produced by the random-number generator.

From:[**WS**] M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015).



6.2 THE RELATIONSHIP BETWEEN POPULATION AND SAMPLE

EXAMPLE 6.8

Obstetrics Suppose we want to characterize the distribution of birthweights of all liveborn infants born in the United States in 2013. Assume the underlying distribution of birthweight has an expected value (or mean) μ and variance σ^2. Ideally, we wish to estimate μ and σ^2 exactly, based on the entire population of U.S. liveborn infants in 2013. But this task is difficult with such a large group. Instead, we decide to select a random sample of n infants who are *representative* of this large group and use the birthweights $x_1, \ldots, x_n$ from this sample to help us estimate μ and σ^2. What is a random sample?

DEFINITION 6.1

A random sample is a selection of some members of the population such that each member is independently chosen and has a known nonzero probability of being selected.

DEFINITION 6.2

A simple random sample is a random sample in which each group member has the same probability of being selected.

DEFINITION 6.3

The **reference**, **target**, or **study population** is the group we want to study. The random sample is selected from the study population.

From:Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

6.3 RANDOM-NUMBER TABLES

In this section, practical methods for selecting random samples are discussed.

DEFINITION 6.4 A **random number** (or **random digit**) is a random variable X that takes on the values 0, 1, 2, . . . , 9 with equal probability. Thus,

$$Pr(X=0)=Pr(X=1)=\cdots=Pr(X=9)=\frac{1}{10}$$

DEFINITION 6.5 Computer-generated **random numbers** are collections of digits that satisfy the following two properties:

(1) Each digit 0, 1, 2, . . . , 9 is equally likely to occur.

(2) The value of any particular digit is independent of the value of any other digit selected.

Table 4 in the Appendix lists 1000 random digits generated by a computer algorithm.

TABLE 4 Table of 1000 random digits

01	32924	22324	18125	09077	26	96772	16443	39877	04653
02	54632	90374	94143	49295	27	52167	21038	14338	01395
03	88720	43035	97081	83373	28	69644	37198	00028	98195
04	21727	11904	41513	31653	29	71011	62004	81712	87536
05	80985	70799	57975	69282	30	31217	75877	85366	55500
06	40412	58826	94868	52632	31	64990	98735	02999	35521
07	43918	56807	75218	46077	32	48417	23569	59307	46550
08	26513	47480	77410	47741	33	07900	65059	48592	44087
09	18164	35784	44255	30124	34	74526	32601	24482	16981
10	39446	01375	75264	51173	35	51056	04402	58353	37332
11	16638	04680	98617	90298	36	39005	93458	63143	21817
12	16872	94749	44012	48884	37	67883	76343	78155	67733
13	65419	87092	78596	91512	38	06014	60999	87226	36071
14	05207	36702	56804	10498	39	93147	88766	04148	42471
15	78807	79243	13729	81222	40	01099	95731	47622	13294
16	69341	79028	64253	80447	41	89252	01201	58138	13809
17	41871	17566	61200	15994	42	41766	57239	50251	64675
18	25758	04625	43226	32986	43	92736	77800	81996	45646
19	06604	94486	40174	10742	44	45118	36600	68977	68831
20	82259	56512	48945	18183	45	73457	01579	00378	70197
21	07895	37090	50627	71320	46	49465	85251	42914	17277
22	59836	71148	42320	67816	47	15745	37285	23768	39302
23	57133	76610	89104	30481	48	28760	81331	78265	60690
24	76964	57126	87174	61025	49	82193	32787	70451	91141
25	27694	17145	32439	68245	50	89664	50242	12382	39379

EXAMPLE 6.14

Hypertension How can the random digits in Appendix Table 4 be used to select 20 random participants in the hypertension treatment program in Example 6.12?

Solution: A roster of the 1000 participants must be compiled, and each participant must then be assigned a number from 000 to 999. Perhaps an alphabetical list of the participants already exists, which would make this task easy. Twenty groups of three digits would then be selected, starting at any position in the random-number table. For example, starting at the first row of Table 4 would yield the numbers listed in Table 6.1.

TABLE 6.1 Twenty random participants chosen from 1000 participants in the hypertension treatment program

First 3 rows of random-number table				Actual random numbers chosen				
32924	22324	18125	09077	329	242	232	418	125
54632	90374	94143	49295	090	775	463	290	374
88720	43035	97081	83373	941	434	929	588	720
				430	359	708	183	373

Frequency distributions and probability distributions

Different individuals in a sample will have different measurements. We can see this variability with a frequency distribution. The **frequency** of a specific measurement in a sample is the number of observations having a particular value of the measurement.[7] The **frequency distribution** shows how often each value of the variable occurs in the sample.

> The ***frequency distribution*** describes the number of times each value of a variable occurs in a sample.

Figure 1.4-1 shows the frequency distribution for the measured beak depths of a sample of 100 finches from a Galápagos island population.[8]

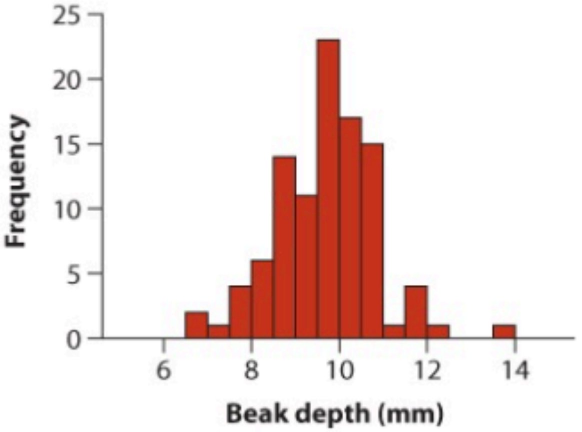
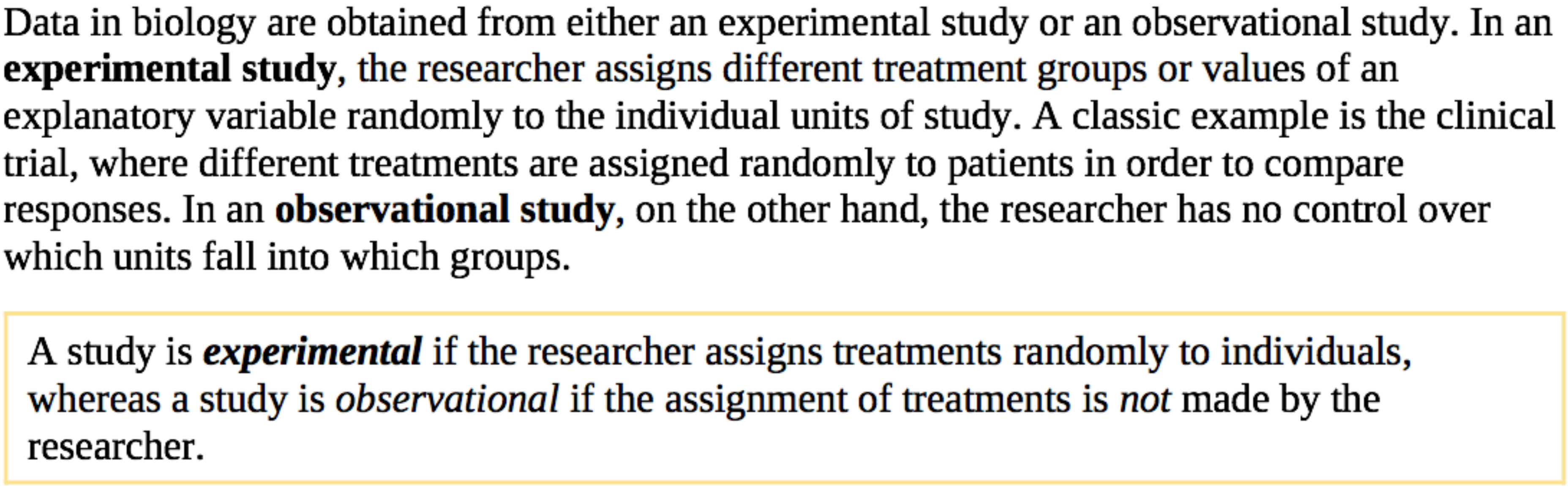


Figure 1.4-1
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

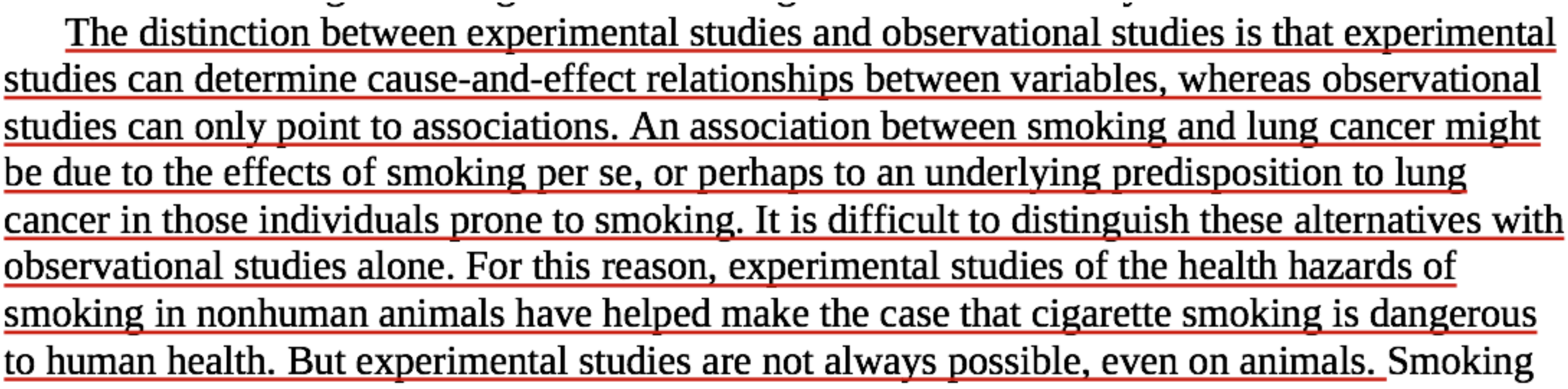
FIGURE 1.4-1 The frequency distribution of beak depths in a sample of 100 finches from a Galápagos island population (Boag and Grant 1984). The vertical axis indicates the frequency, the number of observations in each 0.5-mm interval of beak depth.

Types of studies



Data in biology are obtained from either an experimental study or an observational study. In an **experimental study**, the researcher assigns different treatment groups or values of an explanatory variable randomly to the individual units of study. A classic example is the clinical trial, where different treatments are assigned randomly to patients in order to compare responses. In an **observational study**, on the other hand, the researcher has no control over which units fall into which groups.

A study is ***experimental*** if the researcher assigns treatments randomly to individuals, whereas a study is *observational* if the assignment of treatments is *not* made by the researcher.



The distinction between experimental studies and observational studies is that experimental studies can determine cause-and-effect relationships between variables, whereas observational studies can only point to associations. An association between smoking and lung cancer might be due to the effects of smoking per se, or perhaps to an underlying predisposition to lung cancer in those individuals prone to smoking. It is difficult to distinguish these alternatives with observational studies alone. For this reason, experimental studies of the health hazards of smoking in nonhuman animals have helped make the case that cigarette smoking is dangerous to human health. But experimental studies are not always possible, even on animals. Smoking

The content of this presentation, in a nutshell can be further synthesized by the following keywords in association/opposition:

CASE/ POPULATION/SAMPLE
OBSERVATION / EXPERIMENTATION
RANDOM SAMPLING

These are, so to speak pompously, at the foundation of statistical methods

Summary

- Statistics is the study of methods for measuring aspects of populations from samples and for quantifying the uncertainty of the measurements.
- Much of statistics is about estimation, which infers an unknown quantity of a population using sample data.
- Statistics also allows hypothesis testing, a method to determine how well hypotheses about a population parameter fit the sample data.
- Sampling error is the chance difference between an estimate describing a sample and the corresponding parameter of the whole population. Bias is a systematic discrepancy between an estimate and the population quantity.
- The goals of sampling are to increase the accuracy and precision of estimates and to ensure that it is possible to quantify precision.
- In a random sample, every individual in a population has the same chance of being selected, and the selection of individuals is independent.
- A sample of convenience is a collection of individuals easily available to a researcher, but it is not usually a random sample.
- Volunteer bias is a systematic discrepancy in a quantity between the pool of volunteers and the population.
- Variables are measurements that differ among individuals.
- Variables are either categorical or numerical. A categorical variable describes which category an individual belongs to, whereas a numerical variable is expressed as a number.
- The frequency distribution describes the number of times each value of a variable occurs in a sample. A probability distribution describes the number of times each value occurs in a population. Probability distributions in populations can often be approximated by a normal distribution.
- In studies of association between two variables, one variable is typically used to predict the value of another variable and is designated as the explanatory variable. The other variable is designated as the response variable.
- In experimental studies, the researcher is able to assign subjects randomly to different treatments or groups. In observational studies, the assignment of individuals to treatments is not controlled by the researcher.