Chapter 1
Introduction

1.1 Basic ideas

Statistical methods deal with properties of groups or aggregates. In many applications the entity of primary interest is an actual, physical group (population) of objects. These objects may be animate (*e.g.*, people or animals) or inanimate (*e.g.*, farm field plots, trees, or days). We will refer to the individual objects that comprise the group of interest as **units**. In certain contexts we may refer to the unit as a population unit, a sampling unit, an experimental unit, or a treatment unit.

In order to obtain information about a group of units we first need to obtain information about each of the units in the group. A **variable** is a measurable characteristic of an individual unit. Since our goal is to learn something about the group, we are most interested in the **distribution of the variable**, *i.e.*, the way in which the possible values of the variable are distributed among the units in the group.

When the units are actual, physical objects we define the **population** as the collection of all of the units that we are interested in. In most applications it is unnecessary or undesirable to examine the entire population. Thus we define a **sample** as a subset or part of the population for which we have or will obtain data. The collection of observed values of one or more variables corresponding to the individual units in the sample constitute the **data**. Once the data are obtained we can use the distributions of the variables among the units in the sample to characterize the sample itself and to make inferences or generalizations about the entire population, *i.e.*, inferences about the distributions of these variables among the units in the population.

When discussing the distribution of a variable we need to consider the structure possessed by the possible values of the variable. This leads to the following classification of variables into four basic types.

A **qualitative** variable (categorical variable) classifies a unit into one of several possible categories. The possible values of a qualitative variable are names for these categories. We can distinguish between two types of qualitative variables. A qualitative variable is said to be **nominal** if there is no inherent ordering among its possible values. The sex of a person (female or male) and the color of a person's eyes (blue, brown, *etc.*) are examples of nominal qualitative variables. If there is an inherent ordering of the possible values of a qualitative variable, then it is said to be **ordinal**. The classification of a student (freshman, sophomore, junior, or senior), the ranking of a unit with respect to several size classes (small, medium, or large), and the degree to which a person agrees with a statement (recorded as strongly disagree, disagree, neutral, agree, or strongly agree) are examples of ordinal qualitative variables.

A **quantitative** variable (numerical variable) assigns a meaningful numerical value to a unit. Because the possible values of a quantitative variable are meaningful numerical quantities, they can be viewed as points on a number line. Therefore, it makes sense to talk about where the values of a quantitative variable are located on the number line, whether one value is larger than another, and how far apart two values are. If the possible values of a quantitative variable correspond to isolated points on the number line, then there is a discrete jump between adjacent possible values and the variable is said to be a **discrete** quantitative variable. The most common example of a discrete quantitative variable is a count such as the number of babies in a litter of animals or the number of plants in a field plot. If there is a continuous transition from one value of the variable to the next, then the variable is said to be a **continuous** quantitative variable. For a continuous quantitative variable there is always another possible value between any two possible values, no matter how close together the values are. In practice all quantitative variables are discrete in the sense that the observed values are rounded to a reasonable number of decimal places. Thus the distinction between a continuous quantitative variable and a discrete quantitative variable is often more conceptual than real. If a value of the variable represents a measurement of the size of a unit, such as height, weight, or length, or the amount of some quantity, then it is reasonable to think of the possible values of the variable as forming a continuum of values on the number line and to view the variable as continuous.

The values of ordinal variables are often recorded using numerical codes (ranks) such as 1:strongly disagree, 2:disagree, 3:neutral, 4:agree, or 5:strongly agree. This sort of coding of an ordinal variable does not make it quantitative. For example, the fact that these rankings are equally spaced points on the number line does not necessarily mean that the difference between 1:strongly disagree and 2:disagree is the same as the difference between 4:agree and 5:strongly agree. Therefore, the common practice of treating such ranking variables as quantitative must be used with caution and the fact that the values of the variable are simply ranks must be taken into account when interpreting an analysis of such a ranking variable.

We can also classify variables with respect to the roles they play in a statistical analysis. That is, we can distinguish between response variables and explanatory variables. A **response variable** is a variable that measures the response of a unit to natural or experimental stimuli. A response variable provides us with a measurement or observation that characterizes a unit with respect to a characteristic of primary interest. An **explanatory variable** is a variable that can be used to explain, in whole or in part, how a unit responds to natural or experimental stimuli. This terminology is clearest in the context of an experimental study. Consider an experiment where a unit is subjected to a treatment (some combination of conditions) and the response of the unit to the treatment is recorded. A variable that describes the treatment conditions is called an explanatory variable, since it may be used to explain the outcome of the experiment. A variable that measures the outcome of the experiment is called a response variable, since it measures the response of the unit to the treatment. An explanatory variable may also be used to subdivide a group so that the distributions of a response variable can be compared among subgroups.

In some applications, such as experimental studies, the population is best viewed as a hypothetical population of values of one or more variables. For example, suppose that we are interested in the effects of an alternative diet on weight gain in some population of experimental animals. We might conduct an experiment by randomly assigning animals to two groups; feeding one group a standard diet and the other group the alternative diet; and then recording the weight gained by each animal over some fixed period of time. In this example we can envision two hypothetical populations of weight gains: The population of weight gains we would have observed if all of the animals were given the standard diet; and, the population of weight gains we would have observed if all of the animals were given the alternative diet.

Statistics is often defined as a collection of methods for collecting, describing, and drawing conclusions from data. Methods for collecting data fall under the heading of sampling and experimentation; we will discuss these topics in Chapter 4. Descriptive statistical methods are used to describe the distributions of the values of variables among the units in a sample, *i.e.*, to gain insight about the sample. We will discuss univariate (one variable) descriptive statistical methods in Chapters 2 and 3 and bivariate (two variables) descriptive methods in Chapter 9. Inferential statistical methods are used to make inferences or generalizations, based on the data from the sample, about the distributions of the values of variables among the units in the population, *i.e.*, to gain insight about the population based on information obtained from the sample. Inferential methods are probabilistic in the sense that they are based on probability models for the distributions of variables. The majority of this book deals with inferential statistics; probability models are introduced in Chapter 4a.

We will use the following simple example to clarify the concepts and definitions from above. The data presented in Table 1 were collected on the first day of classes during the Spring 1999 semester. These data provide information about the 67 students who were present on the first day of classes for two sections of the statistics course Stat 214 at the University of Louisiana at Lafayette. Aside from being grouped by section, the data are

Table 1. Statistics 214 class data, spring 1999.

line	section	classification	sex	age	height	weight	siblings	BMI
1	1	senior	male	21	69	170	1	25.10
2	1	junior	male	25	71	165	3	23.01
3	1	junior	female	25	62	160	2	29.26
4	1	freshman	male	18	72	162	1	21.97
5	1	junior	female	22	63	170	1	30.11
6	1	freshman	female	18	64	110	2	18.88
7	1	freshman	female	18	60	103	1	20.11
8	1	freshman	female	18	68	135	3	20.52
9	1	sophomore	female	19	62	105	5	19.20
10	1	freshman	male	18	74	190	2	24.39
11	1	sophomore	female	20	70	150	1	21.52
12	1	senior	female	21	61	116	1	21.92
13	1	freshman	female	18	65	150	3	24.96
14	1	freshman	female	19	64	140	4	24.03
15	1	freshman	male	18	68	130	2	19.76
16	1	freshman	female	18	63	110	2	19.48
17	1	sophomore	female	21	62	125	1	22.86
18	1	freshman	female	18	63	115	2	20.37
19	1	freshman	female	19	64	135	3	23.17
20	1	freshman	female	18	69	155	1	22.89
21	1	sophomore	female	20	65	110	2	18.30
22	1	sophomore	female	19	68	140	1	21.28
23	1	freshman	female	47	66	110	1	17.75
24	1	sophomore	female	20	70	145	2	20.80
25	1	freshman	female	20	61	140	5	26.45
26	1	freshman	female	18	63	180	0	31.88
27	1	junior	male	22	70	175	2	25.11
28	1	freshman	female	18	63	120	1	21.25
29	1	senior	female	22	68	170	2	25.85
30	1	freshman	female	18	66	125	3	20.17
31	1	junior	male	22	75	205	2	25.62
32	1	freshman	female	18	67	110	1	17.23
33	1	senior	male	22	68	135	1	20.52
34	1	senior	female	22	64	185	2	31.75
35	1	freshman	female	41	61	96	1	18.14
36	1	junior	female	22	59	95	5	19.19

This table is continued on the next page.

Table 1. Statistics 214 class data (continuation).

line	section	classification	sex	age	height	weight	siblings	BMI
37	2	junior	female	20	66	110	1	17.75
38	2	junior	male	20	72	180	1	24.41
39	2	junior	female	21	66	120	1	19.37
40	2	sophomore	female	21	61	105	3	19.84
41	2	freshman	female	18	68	134	7	20.37
42	2	freshman	female	28	66	130	4	20.98
43	2	sophomore	female	26	64	135	4	23.17
44	2	sophomore	female	19	64	117	1	20.08
45	2	freshman	female	20	66	140	4	22.59
46	2	junior	female	20	64	130	1	22.31
47	2	senior	female	48	66	140	3	22.59
48	2	junior	female	22	67	115	2	18.01
49	2	sophomore	female	19	66	170	2	27.44
50	2	freshman	male	18	66	190	3	30.66
51	2	sophomore	female	21	67	135	4	21.14
52	2	freshman	female	20	68	140	2	21.28
53	2	sophomore	female	19	62	115	2	21.03
54	2	sophomore	female	20	60	110	2	21.48
55	2	freshman	male	18	72	185	3	25.09
56	2	senior	male	23	72	190	2	25.77
57	2	senior	male	24	69	170	4	25.10
58	2	junior	male	21	72	140	3	18.98
59	2	junior	female	20	65	112	2	18.64
60	2	junior	female	21	62	130	1	23.77
61	2	freshman	female	18	64	120	1	20.60
62	2	sophomore	female	25	66	145	2	23.40
63	2	junior	male	19	65	156	6	25.96
64	2	freshman	female	18	67	125	0	19.58
65	2	junior	female	44	66	165	4	26.63
66	2	sophomore	male	19	71	155	3	21.62
67	2	sophomore	female	19	62	133	2	24.32

presented in no particular order. These data correspond to a convenience sample of students which may or may not be representative of some larger population of students. Values are provided for eight variables: the section the student was registered in (1 or 2); the classification of the student (freshman, sophomore, junior, or senior); the sex of the student (female or male); the age of the student (in years); the height of the student (in inches); the weight of the student (in pounds); the number of siblings the student had $(0, 1, 2, \ldots)$; and the body mass index (BMI) of the student. The derived or constructed variable BMI (in kg/m^2) is the weight of the student (in kilograms) divided by the square of the student's height (in meters).

The sex of a student (with possible values of female and male) and the section the student was registered in (with possible values 1 and 2) are nominal qualitative variables. The classification of a student (with possible values of freshman, sophomore, junior, and senior) is an ordinal qualitative variable. The other variables are quantitative. The number of siblings that the student had (with possible values of $0, 1, 2, \ldots$) is inherently discrete. The other quantitative variables, age (in years), height (in inches), weight (in pounds), and BMI (in kg/m^2) can be viewed as continuous variables.

The section that the student was registered in was included as a potentially interesting explanatory variable which could be used to divide these students into two subgroups so that the distributions of the other variables for these subgroups could be compared. For an initial analysis of these data we would probably view all of the other variables as response variables. That is, a first analysis might consist of examination of the distributions of these response variables for the entire group or comparisons of these distributions by section. After looking at the overall distributions of the variables we might also want to group the students by sex (treat the sex of a student as an explanatory variable) and compare the distributions of height, weight, and BMI for the two sexes.

1.2 Some examples

This section contains a collection of examples which will be used in exercises and as examples in the sequel.

Example. DiMaggio and Mantle. Joe DiMaggio and Mickey Mantle were two well known baseball players. DiMaggio played center field for the New York Yankees for 13 years and was succeeded by Mantle who played center field for 18 years. There has been some argument about which of these two players was better at hitting home runs. The data given in Table 2 are the numbers of home runs hit by the player during each of the seasons he played. For each player these numbers of home runs are listed in order by the seasons he played.

Table 2. Home run data.

Joe DiMaggio:	29 46 32 30 31 30 21 25 20 39 14 32 12
Mickey Mantle:	13 23 21 27 37 52 34 42 31 40 54 30 15 35 19 23 22 18

Example. Weed seeds. C. W. Leggatt counted the number of seeds of the weed *potentilla* found in 98 quarter–ounce batches of the grass *Phleum praetense*. This example is taken from Snedecor and Cochran, *Statistical Methods*, Iowa State, (1980), 198; the original source is C. W. Leggatt, *Comptes rendus de l'association international d'essais de* *semences*, **5** (1935), 27. The 98 observed numbers of weed seeds, which varied from 0 to 7, are summarized in Table 3.

Table 3. Weed seed frequency distribution.

number of seeds	frequency
0	37
1	32
2	16
3	9
4	2
5	0
6	1
7	1
total	98

Example. Vole reproduction. An investigation was conducted to study reproduction in laboratory colonies of voles. This example is taken from Devore and Peck, *Statistics*, (1997), 33; the original reference is the article "Reproduction in laboratory colonies of voles", *Oikos*, (1983), 184. The data summarized in Table 4 are the numbers of babies in 170 litters born to voles in a particular laboratory.

Table 4. Vole baby frequency distribution.

number of babies	frequency
1	1
2	2
3	13
4	19
5	35
6	38
7	33
8	18
9	8
10	2
11	1
total	170

Example. Wooly–bear caterpillar cocoons. A study was conducted to investigate the relationship between air temperature and the temperature inside a wooly–bear caterpillar cocoon. It seems quite reasonable to expect the temperature inside a cocoon to be higher than the air temperature (outside the cocoon). The data given in Table 5 are pairs of air and cocoon temperatures made on 12 days at a location in the high arctic region. Each cocoon temperature is actually the average of two cocoon temperatures. This example comes from Kevan, P.C., T.S. Jensen, and J.D. Shorthouse, "Body temperatures and behaviorial thermoregulation of high arctic wooly–bear caterpillars and pupae (*Gynaephora rossii,* Lymantridae: Lepidoptera) and the importance of sunshine", *Arctic and Alpine Research*, **14**, (1982).

Table 5. Wooly–bear temperature data.

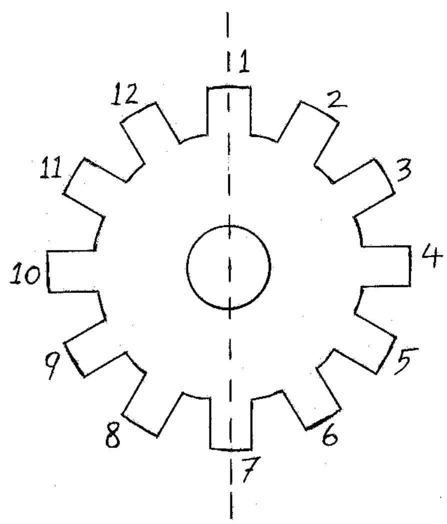
Day	Cocoon temp	Air temp	Day	Cocoon temp	Air temp
1	15.1	10.4	7	3.6	1.7
2	14.6	9.2	8	5.3	2.0
3	6.8	2.2	9	7.0	3.0
4	6.8	2.6	10	7.1	3.5
5	8.0	4.1	11	9.6	4.5
6	8.7	3.7	12	9.5	4.4

Example. Homophone confusion and Alzheimer's disease. A study was conducted to investigate the relationship between Alzheimer's disease and homophone spelling confusion. A homophone pair is a pair of words with the same pronunciation having different meanings and spellings. Twenty patients with Alzheimer's disease were asked to spell 24 homophone pairs (given in random order) and the number of homophone confusions, *e.g.* spelling doe given the context bake bread dough, was recorded for each patient. One year later, the same patients were again asked to spell the same 24 homophone pairs and the number of homophone confusions was again recorded. The data given in Table 6 are the numbers of homophone confusions at the two times of measurement for the 20 Alzheimer's patients. This example comes from Neils, J., D.P. Roeltgen, and F. Constantinidou, "Decline in homophone spelling associated with loss of semantic influence on spelling in Alzheimer's disease", *Brain and Language*, **49**, (1995).

Table 6. Alzheimer's homophone confusion data.

Patient	Time 1	Time 2	Patient	Time 1	Time 2
1	5	5	11	7	10
2	1	3	12	0	3
3	0	0	13	3	9
4	1	1	14	5	8
5	0	1	15	7	12
6	2	1	16	10	16
7	5	6	17	5	5
8	1	2	18	6	3
9	0	9	19	9	6
10	5	8	20	11	8

Example. Gear tooth strength. The data used in this example were published by B. Gunter, "Subversive data analysis, Part II: More graphics, including my favorite example", *Quality Progress*, Nov., 1988, 77–78. This description is adapted from Wild and Seber, *Chance Encounters*, Wiley, (2000), 118. These data concern gear blanks purchased by the Ford Motor Company. Ford engineers found that the teeth on these gears were breaking at too low a stress. The data given below are the impact strengths (in lb–ft) required to break a gear tooth. Each gear had 12 equally spaced teeth. The position



numbers for these teeth begin with 1 at 12 o'clock and proceed in a clockwise direction. The tooth positions are important since they are related to the position of the tooth in the mold used to make the gear. Teeth 1 and 7 are distinguishable; but, teeth located symmetrically about a line drawn through positions 1 and 7 are not, since these positions depend on which face of the gear is upward. Thus, observations for pairs of teeth in a symmetrical position about a line through position 1 and 7 are grouped in Table 7.

Table 7. Gear tooth strength data.

gear position						
1	**2 & 12**	**3 & 11**	**4 & 10**	**5 & 9**	**6 & 8**	**7**
1976	2425	2228	2186	2228	2431	2287
1916	2000	2347	2521	2180	2250	2275
2090	2251	2251	2156	2114	2311	1946
2000	2096	2222	2216	2365	2210	2150
2323	2132	1940	2593	2299	2329	2228
1904	1964	1904	2204	2072	2263	1695
2048	1750	1820	2228	2323	2353	2000
2222	2018	2012	2198	2449	2251	2006
2048	1766	2204	2150	2300	2275	1945
2174		2144	2311	2078	1958	2006
1976		2305	2102	2150	2185	2209
2138		2042	2138	2377		2216
2455		2120	1982	2108		1934
1886		2419	2042	2257		1904
2246		2162	2030	2383		1958
2287		2251	2216	2323		1964
2030		2222	2305	2246		2066
2210			2204	2251		2222
2084			2198	2156		2066
2383			2204	2419		1964
2132			2162	2329		2150
2210			2120	2198		2114
2222			2108	2269		2125
1766			2030	2287		2210
2078			2180	2330		1588
1994			2251	2329		2234
2198			2210	2228		2210
2162			2216			2156
1874			2168			2204
2132			2210			1641
2108			2341			2263
1892			2000			2120
1671			2132			2156

Example. Immigrants to the United States. The data concerning immigrants admitted to the United States summarized by decade as raw frequency distributions in Table 8 were taken from the *2002 Yearbook of Immigration Statistics*, USCIS, (`www.uscis.gov`). Immigrants for whom the country of last residence was unknown are omitted.

Table 8. Region of last residence for immigrants to USA.

	period		
region	**1931–1940**	**1961–1970**	**1991–2000**
Europe	347,566	1,123,492	1,359,737
Asia	16,595	427,692	2,795,672
North America	130,871	886,891	2,441,448
Caribbean	15,502	470,213	978,787
Central America	5,861	101,330	526,915
South America	7,803	257,940	539,656
Africa	1,750	28,954	354,939
Oceania	2,483	25,122	55,845
total	528,431	3,321,634	9,052,999

Example. Cholesterol levels in Guatemalans. This example is taken from Devore and Peck, *Statistics*, 3 ed., (1997), Duxbury, p. 23. The original source is "The Blood Viscosity of Various Socioeconomic Groups in Guatemala" in *The American Journal of Clinical Nutrition*, Nov., 1964, 303–307. The Institute of Nutrition of Central America and Panama measured the serum total cholesterol levels for a group of 49 adult, low–income rural Guatemalans and for a group of 45 adult, high–income urban Guatemalans. The serum total cholesterol levels (in mg/dL) are provided in Table 9.

Table 9. Guatemalan cholesterol data.

Rural group cholesterol levels (in mg/dL).

95	108	108	114	115	124	129	129	131	131
135	136	136	139	140	142	142	143	143	144
144	145	146	148	152	152	155	157	158	158
162	165	166	171	172	173	174	175	180	181
189	192	194	197	204	220	223	226	231	

Urban group cholesterol levels (in mg/dL).

133	134	155	170	175	179	181	184	188	189
190	196	197	199	200	200	201	201	204	205
205	205	206	214	217	222	222	227	227	228
234	234	236	239	241	242	244	249	252	273
279	284	284	284	330					

1.3 Exercises

For each of the examples in Section 1.2 define or identify the following:

1. The unit.

2. The group(s) of interest.

3. The variable(s) and the possible values of the variable(s).

4. The type of variable(s) (nominal qualitative, ordinal qualitative, discrete quantitative, or continuous quantitative).