1 Statistics and Data

What do Numbers have to do with Trees?

In this chapter we define the term **statistics**. We also discuss **data**, the building blocks of statistics, and introduce data collection procedures and measurement scales.

1.1 What is Statistics?

When most people think of *statistics*, they imagine percentages, averages and rates of change, which are displayed in tables, graphs and charts. Statistics give us information about debits and credits, incomes and taxes, births and deaths, home prices, daily temperatures, smokers' mortality rates and so on. In sports, such as hockey, the term statistics refers to the records kept on each player's performance, such as the number of goals, number of shots on goal and number of penalty minutes.

Similarly, in forestry, statistics are collected to summarize tree heights, diameters, volumes, seedling survival rates, bark beetle infestation costs and much more. In sawmills, tables and charts are compiled to indicate the quality of the products produced, such as the distribution of lumber by grade, the strength of the lumber and other important quality characteristics.

We offer two definitions for statistics in this book.

- *Statistics* is the science of collecting, organizing, analysing and interpreting information (in this case, *statistics* is singular).
- *Statistics* are numbers calculated from information (in this case, *statistics* can be singular or plural).

The study of statistics is generally subdivided into two distinct fields: **descriptive statistics** and **statistical inference**.

Consider a large body of information, such as 5000 measurements of tree height collected from a forest management unit. Ordinarily, it is almost impossible to look at a large listing of numbers like this and draw any meaningful conclusions. Using descriptive statistics, we can describe this information with tables, summary numbers, charts and graphs. In this way, an observer (e.g. the forest manager) can very easily and quickly characterize, summarize and communicate the attributes of the forest management unit being measured (tree heights) since it is generally easier to understand the information when it is presented in the form of tables, charts, graphs or summary numbers, the latter often being referred to as *statistics*. The organization or tabulation of such large bodies of information has become a necessary skill for people employed in the forestry sector, from conservation biologists to foresters to wood products manufacturers.

> **Descriptive statistics** deals with the collection, organization and presentation of *information* and the calculations of some measures (*statistics*) that describe the information.

 Introductory Probability and Statistics: Applications for Forestry and Natural Sciences (A. Kozak, R.A. Kozak, C.L. Staudhammer and S.B. Watts)

We may also want to use the information on hand to make predictions about the future or make statements about the larger body of information from which our data were taken. Statistical inference, or inferential statistics, uses information contained in a **sample** to reach conclusions about one or more characteristics of the whole **population**.

> A **population** is the entire collection of items/subjects possessing certain common characteristics about which information is being sought. The characteristics of a population are called **parameters** and are usually denoted with Greek letters (e.g. μ, σ).
>
> A **sample** is a portion or subset of the population. The characteristics of a sample are called **statistics** and are usually denoted with Roman letters (e.g. x, p).

It would be ideal if we could obtain information from every item (subject) in a population. However, populations are quite often very large (e.g. all possible trees in a forest type) and, therefore, it is simply not practical, or even possible, to collect the desired information from each item of the population. In other cases, such as the testing of modulus of rupture (strength) in 2×4s, it is not tenable to observe every item in the population because the item being observed is destroyed in the gathering of this information.

In these situations, we must collect the desired information from a sample. This portion, or subset, of the population is used to draw conclusions (inferences) concerning the whole population. This type of generalization, based on an incomplete set of information, involves a certain amount of **risk**. Therefore, in studying and using inferential statistics, a considerable amount of time is spent quantifying the associated risk. Some theories in **probability** will help us to properly quantify these risks.

> **Inferential statistics** or **statistical inference** is concerned with generalizing from the information obtained in a sample to an entire population. This generalization involves **estimation**, **hypothesis testing**, **determining relationships** and **prediction**.

1.2 Data

Pieces of information collected on subjects or items from a population form the building blocks of statistics and are called data (data is the plural of datum, a piece of information). Data can be collected, organized, analysed and summarized. Table 1.1 shows an example data set, which contains information collected from 50 trees. The trees here are the *elements* (or items or subjects) on which the data were collected. A *variable* is a characteristic of an element that we want to study. Seven *variables* were recorded in this data set: (i) tree identification number; (ii) date of measurement; (iii) species; (iv) crown class; (v) number of neighbouring trees (growing within a 5 m radius); (vi) diameter at breast height (dbh, which is measured at a height of 1.3 m in Canada, or 4.5 ft in the USA); and (vii) height.

Usually, a variable takes on different values from element to element, hence the name. In general, variables whose values are determined by chance are referred to as *random variables*. A set of measurements (such as the seven variables seen in Table 1.1) collected for one element is called an *observation*, and thus Table 1.1 contains 50 observations.

Table 1.1. A data set for 50 trees

Tree number	Date of measurement[a]	Species[b]	Crown class[c]	Number of neighbouring trees[d]	Diameter at breast height (cm)	Total height (m)
1	12	F	C	4	15.3	14.78
2	12	F	D	3	17.8	17.07
3	9	C	D	5	18.2	18.28
4	9	H	S	4	9.7	8.79
5	7	H	I	6	10.8	10.18
6	10	C	I	3	14.1	14.90
7	10	C	C	2	17.1	15.34
8	12	C	D	2	20.6	17.22
9	16	F	C	4	18.2	15.15
10	14	F	I	5	16.1	14.66
11	8	H	D	3	14.2	17.43
12	5	H	D	6	14.8	17.45
13	12	F	I	2	19.1	14.18
14	5	C	I	2	16.7	13.40
15	12	C	S	4	18.9	10.40
16	20	H	S	3	12.4	11.52
17	15	H	C	0	17.3	14.61
18	20	F	D	1	22.7	21.46
19	15	C	C	4	15.1	17.82
20	14	C	I	3	17.7	11.38
21	14	C	S	5	13.4	8.50
22	13	C	I	4	16.2	12.80
23	14	F	D	1	18.5	18.71
24	20	F	I	4	15.0	14.48
25	21	F	C	2	18.8	14.81
26	5	H	I	4	15.8	12.01
27	2	H	I	3	16.1	11.70
28	22	C	C	3	15.4	16.03
29	22	C	I	0	17.8	14.46
30	18	C	S	1	18.5	8.47
31	16	C	I	3	14.1	11.22
32	16	C	C	5	14.8	12.34
33	17	F	C	4	15.5	16.79
34	17	F	I	6	13.8	16.06
35	18	F	S	4	13.0	13.20
36	20	H	C	2	18.2	14.30
37	22	H	C	0	22.3	16.84
38	20	H	I	3	17.8	13.84
39	17	C	I	4	13.1	11.31
40	17	C	I	6	12.8	13.20
41	16	C	C	3	13.3	13.75
42	23	F	C	3	15.6	14.60
43	23	H	C	4	16.6	12.56
44	22	C	I	5	13.0	10.88
45	24	C	I	4	10.2	13.93
46	23	F	I	3	14.4	12.68
47	24	C	S	6	7.7	10.00
48	25	C	S	5	9.9	8.69
49	25	H	D	1	20.4	16.73
50	24	H	D	3	20.9	16.25

[a] Day of the month (March, 2006). [b] C, western red cedar; F, Douglas-fir; and H, western hemlock. [c] D, dominant; C, codominant; I, intermediate; and S, suppressed. [d] Trees within a 5 m radius of the subject tree.

Variables can be classified as either **qualitative** or **quantitative**. **Qualitative variables** are also known as **categorical variables** because they can be placed into distinct categories according to some characteristic. Species and crown class (Table 1.1) are qualitative variables. Other examples of categorical variables include, forest type, level of insect infestation (low, medium and heavy) and field of study (Forestry, Engineering, Agriculture, Education, Arts, Science).

Quantitative variables are numerical and can indicate 'how many' or 'how much' or 'how big' on a numeric scale. For example, dbh, height and number of neighbouring trees (Table 1.1) are quantitative variables. Quantitative variables can be further subdivided into **discrete** and **continuous variables**. Discrete variables, which take on whole numbers only, usually result from counting something such as the number of neighbouring trees (Table 1.1). Continuous variables are those which can take on 'all possible values' over a specific interval and are generally measured, e.g. height and dbh (Table 1.1). Often, 'all possible values' exist only in theory since measurement processes are limited to the precision of measurement devices. For example, current measurement techniques only allow dbh to be measured to the nearest 0.1 cm and tree height to the nearest 0.01 m. This means that a recorded dbh of 15.2 cm includes all possible values between 15.15 and 15.25 cm (not including trees with 15.25 cm dbh).

1.3 Measurement Scales

In analysing variables, the scale of measurement refers to the amount of information contained within the variable and indicates what types of statistical analyses are appropriate. Four common scales are used for measurements: **nominal**, **ordinal**, **interval** and **ratio**.

Nominal scale data can be quantitative or qualitative and are used mainly for identification and classification of items. Examples of quantitative nominal scale data include the tree numbers listed in Table 1.1, numbers on hockey jerseys, zip codes in the USA and telephone numbers (note that the use of numbers here is for identification purposes only). Examples of qualitative nominal scale data are the species identified in Table 1.1, marital status and postal codes in Canada (e.g. V6S 1B9). Even if a variable is quantitative, arithmetic operations (addition, subtraction, multiplication and division) and/or ranking the items by their values are not meaningful for nominal scale data.

The **ordinal scale** is similar to the nominal scale, but in an ordinal scale, the order or rank of the values is valid. For example, crown class (Table 1.1) is in an ordinal scale, as it is known that the dominant trees are taller than the codominant trees within a stand. Again, ordinal scale data can be qualitative or quantitative. Examples of qualitative ordinal scale data are letter grades, levels of insect infestation (light, medium and heavy) and ranking of food quality (excellent, good, medium and poor). Examples of quantitative ordinal scale data are addresses in a block on one side of a street and numeric quality rankings (e.g. 1 for excellent, 2 for good, ..., 5 for poor). While the ranking of items in an ordinal scale is valid and meaningful in interpreting data, arithmetic operations (addition, subtraction, multiplication and division) are not.

The **interval scale** has the same properties as the ordinal scale, but interval scale data are always quantitative and differences between data values are meaningful.

Examples of interval scale data are temperature (in Celsius or Fahrenheit), Scholastic Aptitude Test (SAT) scores and measurement date (Table 1.1). When using an interval scale, zero does not indicate an absence of measurement. For instance, zero degrees is set as the icing point on a Celsius temperature scale; however, zero degrees does not indicate an absence of temperature. Similarly, if the temperature on a given day was 20°C in Vancouver and 10°C in Toronto, the difference of 10°C is meaningful. However, it does not mean that it is twice as warm in Vancouver as in Toronto.

The **ratio scale** is similar to the interval scale, but with two main differences. In the ratio scale, zero means 'none' and, therefore, the ratio of two variables becomes meaningful. Height, dbh and number of neighbouring trees (Table 1.1) are measured in a ratio scale; other examples are weight, distance, height and cost. All arithmetic operations (addition, subtraction, multiplication and division) are valid with ratio scale data. For example, we can say that a 20 m tree is twice as tall as a 10 m tree.

1.4 Data Collection

Data can be collected in many ways. Sometimes, data required for a particular application are available from government or company offices where operational data sets have been historically maintained. Data on forest inventory levels, production quantities and imports and exports are often collected by organizations, such as the United Nations Food and Agriculture Organization (UN/FAO). Employment rates, wage rates and other labour force information can usually be obtained from various government agencies.

If the required data are not available from existing sources, we can turn to some well-known statistical tools for data collection, namely **experimental designs** or **sampling designs** (or a combination of the two). These two techniques are frequently referred to as **experimental** and **observational** studies, respectively, and are discussed in more detail in Chapter 13 of this volume.

In experimental studies, one or more factors affecting the variable(s) of interest are controlled. The objectives of the study are to investigate how these controlled factors affect the variable(s) of interest. For example, to investigate the effect of seeding date on burnt and unburnt seedbeds, the dates and preparation of the seedbeds are controlled and the effects on germination are studied.

In observational studies (sampling), no attempt is made to control the variables of interest; we merely observe a given situation. The main purpose of sampling is to collect data from a subset of the population and to use this data to make predictions or inferences about the entire population. For example, if we would like to estimate the average height of a lodgepole pine plantation, we could randomly select 40 trees from the stand, measure their heights and estimate (with some degree of error) the unknown population average of height. These sorts of sampling designs are often referred to as **sample surveys**.

Exercises

Section 1.1

1.1. Define the word 'statistics'.

1.2. Give 3 examples of how descriptive statistics can be used in your field of interest.

1.3. Give an example of how inferential statistics can be applied in your field of interest.

1.4. Two students are sent out to measure the diameter at breast height (dbh; measured at 1.3 meters above ground level) and height of 75 randomly selected trees in a plantation containing 10,753 trees. Each tree in the plantation has been labeled with a number. Seventy-five random numbers between 1 and 10,753 were generated to indicate the trees to be measured.

a. Describe the population.
b. Describe the selected sample.
c. Give an example of a descriptive statistic that could be used in this study.
d. Give an example of an inferential statistic that will be used in this study.

1.5. A wood science student is working as an intern for a particleboard mill. Every 15 minutes, she is asked to remove one board as it comes off the production line to measure its thickness. These observations will be used to study the quality of the boards being produced as part of a program for statistical quality control.

a. Describe the population.
b. Describe the selected sample.
c. Give an example of a descriptive statistic that could be used in this study.
d. Give an example of an inferential statistic that will be used in this study.

Section 1.2

1.6. Classify each of the following variables as qualitative or quantitative.

a. Number of trees per hectare or land area.
b. Colour of *Rhododendron* flowers.
c. Number of leaders on a weevil-infested Sitka spruce seedling.
d. Outside bark diameter at breast height of a cork oak tree.
e. Fire hazard classification (low, moderate or severe).
f. Thickness of plywood.
g. Grade of lumber (#1, #2 or defective).
h. Snout to vent length of a garden lizard (*Calotes versicolor*).
i. Age of a ponderosa pine tree determined from the number of annual rings.
j. Species.
k. Daily low and high temperatures measured in degrees centigrade.
l. Annual wage (in dollars) earned by 20 graduates from an undergraduate ecology program.
m. The date that each of the above 20 graduates obtained a job.

1.7. Classify the quantitative variables in Exercise 1.6 as discrete or continuous.

Section 1.3

1.8. Classify the variables listed in Exercise 1.6 (Section 1.2) by their scale of measurement (nominal, ordinal, interval or ratio).

Section 1.4

1.9. The effect of chemical treatment on the germination rates of 300 seeds was studied. One hundred seeds were treated with chemical A, 100 seeds were treated with chemical B, and 100 were left untreated. After the treatment, the 300 seeds were monitored and the length of time to germination was measured. Identify and briefly describe the data collection method used in this study.

1.10. Identify and describe the data collection method used in Exercise 1.9 (Section 1.4).

1.11. Modify the study described in Exercise 1.9 so that both experimental design and sampling design are used to obtain the required information.