INTRODUCTION TO STATISTICS

1.1 INTRODUCTION

It has become accepted in today's world that in order to learn about something, you must first collect data. *Statistics* is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.

1.2 DATA COLLECTION AND DESCRIPTIVE STATISTICS

Sometimes a statistical analysis begins with a given set of data: For instance, the government regularly collects and publicizes data concerning yearly precipitation totals, earthquake occurrences, the unemployment rate, the gross domestic product, and the rate of inflation. Statistics can be used to describe, summarize, and analyze these data.

In other situations, data are not yet available; in such cases statistical theory can be used to design an appropriate experiment to generate data. The experiment chosen should depend on the use that one wants to make of the data. For instance, suppose that an instructor is interested in determining which of two different methods for teaching computer programming to beginners is most effective. To study this question, the instructor might divide the students into two groups, and use a different teaching method for each group. At the end of the class the students can be tested and the scores of the members of the different groups compared. If the data, consisting of the test scores of members of each group, are significantly higher in one of the groups, then it might seem reasonable to suppose that the teaching method used for that group is superior.

It is important to note, however, that in order to be able to draw a valid conclusion from the data, it is essential that the students were divided into groups in such a manner that neither group was more likely to have the students with greater natural aptitude for programming. For instance, the instructor should not have let the male class members be one group and the females the other. For if so, then even if the women scored significantly higher than the men, it would not be clear whether this was due to the method used to teach them, or to the fact that women may be inherently better than men at learning programming

skills. The accepted way of avoiding this pitfall is to divide the class members into the two groups "at random." This term means that the division is done in such a manner that all possible choices of the members of a group are equally likely.

At the end of the experiment, the data should be described. For instance, the scores of the two groups should be presented. In addition, summary measures such as the average score of members of each of the groups should be presented. This part of statistics, concerned with the description and summarization of data, is called *descriptive statistics*.

1.3 INFERENTIAL STATISTICS AND PROBABILITY MODELS

After the preceding experiment is completed and the data are described and summarized, we hope to be able to draw a conclusion about which teaching method is superior. This part of statistics, concerned with the drawing of conclusions, is called *inferential statistics*.

To be able to draw a conclusion from the data, we must take into account the possibility of chance. For instance, suppose that the average score of members of the first group is quite a bit higher than that of the second. Can we conclude that this increase is due to the teaching method used? Or is it possible that the teaching method was not responsible for the increased scores but rather that the higher scores of the first group were just a chance occurrence? For instance, the fact that a coin comes up heads 7 times in 10 flips does not necessarily mean that the coin is more likely to come up heads than tails in future flips. Indeed, it could be a perfectly ordinary coin that, by chance, just happened to land heads 7 times out of the total of 10 flips. (On the other hand, if the coin had landed heads 47 times out of 50 flips, then we would be quite certain that it was not an ordinary coin.)

To be able to draw logical conclusions from data, we usually make some assumptions about the chances (or *probabilities*) of obtaining the different data values. The totality of these assumptions is referred to as a *probability model* for the data.

Sometimes the nature of the data suggests the form of the probability model that is assumed. For instance, suppose that an engineer wants to find out what proportion of computer chips, produced by a new method, will be defective. The engineer might select a group of these chips, with the resulting data being the number of defective chips in this group. Provided that the chips selected were "randomly" chosen, it is reasonable to suppose that each one of them is defective with probability p, where p is the unknown proportion of all the chips produced by the new method that will be defective. The resulting data can then be used to make inferences about p.

In other situations, the appropriate probability model for a given data set will not be readily apparent. However, careful description and presentation of the data sometimes enable us to infer a reasonable model, which we can then try to verify with the use of additional data.

Because the basis of statistical inference is the formulation of a probability model to describe the data, an understanding of statistical inference requires some knowledge of

the theory of probability. In other words, statistical inference starts with the assumption that important aspects of the phenomenon under study can be described in terms of probabilities; it then draws conclusions by using data to make inferences about these probabilities.

1.4 POPULATIONS AND SAMPLES

In statistics, we are interested in obtaining information about a total collection of elements, which we will refer to as the *population*. The population is often too large for us to examine each of its members. For instance, we might have all the residents of a given state, or all the television sets produced in the last year by a particular manufacturer, or all the households in a given community. In such cases, we try to learn about the population by choosing and then examining a subgroup of its elements. This subgroup of a population is called a *sample*.

If the sample is to be informative about the total population, it must be, in some sense, representative of that population. For instance, suppose that we are interested in learning about the age distribution of people residing in a given city, and we obtain the ages of the first 100 people to enter the town library. If the average age of these 100 people is 46.2 years, are we justified in concluding that this is approximately the average age of the entire population? Probably not, for we could certainly argue that the sample chosen in this case is probably not representative of the total population because usually more young students and senior citizens use the library than do working-age citizens.

In certain situations, such as the library illustration, we are presented with a sample and must then decide whether this sample is reasonably representative of the entire population. In practice, a given sample generally cannot be assumed to be representative of a population unless that sample has been chosen in a random manner. This is because any specific nonrandom rule for selecting a sample often results in one that is inherently biased toward some data values as opposed to others.

Thus, although it may seem paradoxical, we are most likely to obtain a representative sample by choosing its members in a totally random fashion without any prior considerations of the elements that will be chosen. In other words, we need not attempt to deliberately choose the sample so that it contains, for instance, the same gender percentage and the same percentage of people in each profession as found in the general population. Rather, we should just leave it up to "chance" to obtain roughly the correct percentages. Once a random sample is chosen, we can use statistical inference to draw conclusions about the entire population by studying the elements of the sample.

1.5 A BRIEF HISTORY OF STATISTICS

A systematic collection of data on the population and the economy was begun in the Italian city-states of Venice and Florence during the Renaissance. The term *statistics*, derived from the word *state*, was used to refer to a collection of facts of interest to the state. The idea of collecting data spread from Italy to the other countries of Western Europe. Indeed, by the

first half of the 16th century it was common for European governments to require parishes to register births, marriages, and deaths. Because of poor public health conditions this last statistic was of particular interest.

The high mortality rate in Europe before the 19th century was due mainly to epidemic diseases, wars, and famines. Among epidemics, the worst were the plagues. Starting with the Black Plague in 1348, plagues recurred frequently for nearly 400 years. In 1562, as a way to alert the King's court to consider moving to the countryside, the City of London began to publish weekly bills of mortality. Initially these mortality bills listed the places of death and whether a death had resulted from plague. Beginning in 1625 the bills were expanded to include all causes of death.

In 1662 the English tradesman John Graunt published a book entitled *Natural and Political Observations Made upon the Bills of Mortality*. Table 1.1, which notes the total number of deaths in England and the number due to the plague for five different plague years, is taken from this book.

TABLE 1.1 *Total Deaths in England*

Year	Burials	Plague Deaths
1592	25,886	11,503
1593	17,844	10,662
1603	37,294	30,561
1625	51,758	35,417
1636	23,359	10,400

Source: John Graunt, Observations Made upon the Bills of Mortality. 3rd ed. London: John Martyn and James Allestry (1st ed. 1662).

Graunt used London bills of mortality to estimate the city's population. For instance, to estimate the population of London in 1660, Graunt surveyed households in certain London parishes (or neighborhoods) and discovered that, on average, there were approximately 3 deaths for every 88 people. Dividing by 3 shows that, on average, there was roughly 1 death for every 88/3 people. Because the London bills cited 13,200 deaths in London for that year, Graunt estimated the London population to be about

$$13{,}200 \times 88/3 = 387{,}200$$

Graunt used this estimate to project a figure for all England. In his book he noted that these figures would be of interest to the rulers of the country, as indicators of both the number of men who could be drafted into an army and the number who could be taxed.

Graunt also used the London bills of mortality — and some intelligent guesswork as to what diseases killed whom and at what age — to infer ages at death. (Recall that the bills of mortality listed only causes and places at death, not the ages of those dying.) Graunt then used this information to compute tables giving the proportion of the population that

TABLE 1.2 *John Graunt's Mortality Table*

Age at Death	Number of Deaths per 100 Births
0–6	36
6–16	24
16–26	15
26–36	9
36–46	6
46–56	4
56–66	3
66–76	2
76 and greater	1

Note: The categories go up to but do not include the right-hand value. For instance, 0–6 means all ages from 0 up through 5.

dies at various ages. Table 1.2 is one of Graunt's mortality tables. It states, for instance, that of 100 births, 36 people will die before reaching age 6, 24 will die between the age of 6 and 15, and so on.

Graunt's estimates of the ages at which people were dying were of great interest to those in the business of selling annuities. Annuities are the opposite of life insurance in that one pays in a lump sum as an investment and then receives regular payments for as long as one lives.

Graunt's work on mortality tables inspired further work by Edmund Halley in 1693. Halley, the discoverer of the comet bearing his name (and also the man who was most responsible, by both his encouragement and his financial support, for the publication of Isaac Newton's famous *Principia Mathematica*), used tables of mortality to compute the odds that a person of any age would live to any other particular age. Halley was influential in convincing the insurers of the time that an annual life insurance premium should depend on the age of the person being insured.

Following Graunt and Halley, the collection of data steadily increased throughout the remainder of the 17th and on into the 18th century. For instance, the city of Paris began collecting bills of mortality in 1667, and by 1730 it had become common practice throughout Europe to record ages at death.

The term *statistics*, which was used until the 18th century as a shorthand for the descriptive science of states, became in the 19th century increasingly identified with numbers. By the 1830s the term was almost universally regarded in Britain and France as being synonymous with the "numerical science" of society. This change in meaning was caused by the large availability of census records and other tabulations that began to be systematically collected and published by the governments of Western Europe and the United States beginning around 1800.

Throughout the 19th century, although probability theory had been developed by such mathematicians as Jacob Bernoulli, Karl Friedrich Gauss, and Pierre-Simon Laplace, its use in studying statistical findings was almost nonexistent, because most social statisticians

at the time were content to let the data speak for themselves. In particular, statisticians of that time were not interested in drawing inferences about individuals, but rather were concerned with the society as a whole. Thus, they were not concerned with sampling but rather tried to obtain censuses of the entire population. As a result, probabilistic inference from samples to a population was almost unknown in 19th century social statistics.

It was not until the late 1800s that statistics became concerned with inferring conclusions from numerical data. The movement began with Francis Galton's work on analyzing hereditary genius through the uses of what we would now call regression and correlation analysis (see Chapter 9), and obtained much of its impetus from the work of Karl Pearson. Pearson, who developed the chi-square goodness of fit tests (see Chapter 11), was the first director of the Galton Laboratory, endowed by Francis Galton in 1904. There Pearson originated a research program aimed at developing new methods of using statistics in inference. His laboratory invited advanced students from science and industry to learn statistical methods that could then be applied in their fields. One of his earliest visiting researchers was W. S. Gosset, a chemist by training, who showed his devotion to Pearson by publishing his own works under the name "Student." (A famous story has it that Gosset was afraid to publish under his own name for fear that his employers, the Guinness brewery, would be unhappy to discover that one of its chemists was doing research in statistics.) Gosset is famous for his development of the *t*-test (see Chapter 8).

Two of the most important areas of applied statistics in the early 20th century were population biology and agriculture. This was due to the interest of Pearson and others at his laboratory and also to the remarkable accomplishments of the English scientist Ronald A. Fisher. The theory of inference developed by these pioneers, including among others

TABLE 1.3 *The Changing Definition of Statistics*

Statistics has then for its object that of presenting a faithful representation of a state at a determined epoch. (Quetelet, 1849)
Statistics are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man. (Galton, 1889)
Statistics may be regarded (i) as the study of populations, (ii) as the study of variation, and (iii) as the study of methods of the reduction of data. (Fisher, 1925)
Statistics is a scientific discipline concerned with collection, analysis, and interpretation of data obtained from observation or experiment. The subject has a coherent structure based on the theory of Probability and includes many different procedures which contribute to research and development throughout the whole of Science and Technology. (E. Pearson, 1936)
Statistics is the name for that science and art which deals with uncertain inferences — which uses numbers to find out something about nature and experience. (Weaver, 1952)
Statistics has become known in the 20th century as the mathematical tool for analyzing experimental and observational data. (Porter, 1986)
Statistics is the art of learning from data. (this book, 2014)

Karl Pearson's son Egon and the Polish born mathematical statistician Jerzy Neyman, was general enough to deal with a wide range of quantitative and practical problems. As a result, after the early years of the 20th century a rapidly increasing number of people in science, business, and government began to regard statistics as a tool that was able to provide quantitative solutions to scientific and practical problems (see Table 1.3).

Nowadays the ideas of statistics are everywhere. Descriptive statistics are featured in every newspaper and magazine. Statistical inference has become indispensable to public health and medical research, to engineering and scientific studies, to marketing and quality control, to education, to accounting, to economics, to meteorological forecasting, to polling and surveys, to sports, to insurance, to gambling, and to all research that makes any claim to being scientific. Statistics has indeed become ingrained in our intellectual heritage.

Problems

1. An election will be held next week and, by polling a sample of the voting population, we are trying to predict whether the Republican or Democratic candidate will prevail. Which of the following methods of selection is likely to yield a representative sample?

(a) Poll all people of voting age attending a college basketball game.
(b) Poll all people of voting age leaving a fancy midtown restaurant.
(c) Obtain a copy of the voter registration list, randomly choose 100 names, and question them.
(d) Use the results of a television call-in poll, in which the station asked its listeners to call in and name their choice.
(e) Choose names from the telephone directory and call these people.

2. The approach used in Problem 1(e) led to a disastrous prediction in the 1936 presidential election, in which Franklin Roosevelt defeated Alfred Landon by a landslide. A Landon victory had been predicted by the *Literary Digest*. The magazine based its prediction on the preferences of a sample of voters chosen from lists of automobile and telephone owners.

(a) Why do you think the *Literary Digest*'s prediction was so far off?
(b) Has anything changed between 1936 and now that would make you believe that the approach used by the *Literary Digest* would work better today?

3. A researcher is trying to discover the average age at death for people in the United States today. To obtain data, the obituary columns of the *New York Times* are read for 30 days, and the ages at death of people in the United States are noted. Do you think this approach will lead to a representative sample?

4. To determine the proportion of people in your town who are smokers, it has been decided to poll people at one of the following local spots:

   (a) the pool hall;
   (b) the bowling alley;
   (c) the shopping mall;
   (d) the library.

   Which of these potential polling places would most likely result in a reasonable approximation to the desired proportion? Why?

5. A university plans on conducting a survey of its recent graduates to determine information on their yearly salaries. It randomly selected 200 recent graduates and sent them questionnaires dealing with their present jobs. Of these 200, however, only 86 were returned. Suppose that the average of the yearly salaries reported was $75,000.

   (a) Would the university be correct in thinking that $75,000 was a good approximation to the average salary level of all of its graduates? Explain the reasoning behind your answer.
   (b) If your answer to part (a) is no, can you think of any set of conditions relating to the group that returned questionnaires for which it would be a good approximation?

6. An article reported that a survey of clothing worn by pedestrians killed at night in traffic accidents revealed that about 80 percent of the victims were wearing dark-colored clothing and 20 percent were wearing light-colored clothing. The conclusion drawn in the article was that it is safer to wear light-colored clothing at night.

   (a) Is this conclusion justified? Explain.
   (b) If your answer to part (a) is no, what other information would be needed before a final conclusion could be drawn?

7. Critique Graunt's method for estimating the population of London. What implicit assumption is he making?

8. The London bills of mortality listed 12,246 deaths in 1658. Supposing that a survey of London parishes showed that roughly 2 percent of the population died that year, use Graunt's method to estimate London's population in 1658.

9. Suppose you were a seller of annuities in 1662 when Graunt's book was published. Explain how you would make use of his data on the ages at which people were dying.

10. Based on Graunt's mortality table:

    (a) What proportion of people survived to age 6?
    (b) What proportion survived to age 46?
    (c) What proportion died between the ages of 6 and 36?