



1

What is statistics?

The world is ready for the truth; the modern age is here; every year another report appears that examines poverty by means of statistical research rather than romantic claptrap.

(from *The Crimson Petal and the White*, Michael Faber, p. 334)

In this introductory chapter, we give a general description of the topics of statistics and probability theory. Some examples illustrate the purpose and applications of both disciplines, as well as the differences between them. As statistics has more applications in science, industry, and economics than probability theory, statistics is typically given far more attention in degree subjects like business, industrial and bio-science engineering, applied economics, and natural or social sciences. Nevertheless, one should pay some attention to probability theory as well. In fact, both disciplines are strongly connected to each other: it is impossible to understand the working of statistical inference without a sound knowledge of probability theory. Therefore, in this book, we discuss both probability theory and statistics.

1.1 Why statistics?

For many years, statistics has been a subject, often a dreaded one, in several fields of study at universities and colleges. The reason is that quite a few people will, sooner or later, be confronted with problems of data analysis during their professional activities. A sound statistical background not only allows us to analyze the data and to make concrete decisions based on the analysis, but it also provides an advantage in the data collection process.

Nevertheless, statistics is not immediately perceived as useful by most students. This is mainly due to the fact that, during a statistics course, they are still unfamiliar with the sorts of practical decision problems managers, economists, engineers, and

Statistics with JMP: Graphs, Descriptive Statistics, and Probability, First Edition. Peter Goos and David Meintrup.

© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.

Companion Website: wiley.com/go/goosandmeintrup



2 STATISTICS WITH JMP

researchers face on a daily basis. Many students will start realizing the usefulness of statistics when they start to work on their bachelor's or master's thesis. The many examples in this basic course are intended to advance this awareness by several years.

In an introductory statistics course, one often finds a whole series of quotes as an attempt to motivate students. A classic example is “Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.” from the British writer Herbert George Wells (1866–1946). More recent is the judgment by the US quality guru W. Edwards Deming, to whom a large part of the down-right spectacular economic recovery in Japan after World War II is attributed. He claimed that “Statistics is too important to be left to statisticians. The goal is to have many statistically-skilled workers: engineers, scientists, managers...” Hal Varian, chief economist at Google says the following: “I keep saying that the most sexy job in the next 10 years will be statistician. And I’m not kidding.” In Europe, Willy Buysse, former CEO at SN Brussels Airlines, states that too few decisions are made based on data. Only recently, his many years of diligence establishing a research department, where statistical and other quantitative methods are used to address all sorts of problems, has been rewarded.

Another justification for a thorough training in statistical methods can be found in the so-called Six Sigma improvement program. The purpose of this program is to solve concrete problems with a large financial impact both in service and industrial companies, and to reduce the number of faults and defects to 3.4 per million opportunities. The approach is based on statistical methods, as presented in Figure 1.1. The figure shows that the traditional method to solve a practical problem is to immediately search for practical solutions. This approach is typically based on guessing and trial-and-error, so that it will often take a long time to find a final solution to the problem. The Six Sigma improvement program promotes a more thoughtful, scientific approach to problems. First, data is collected in the so-called measurement phase. Then, using statistical methods, the data is carefully examined. This often leads to interesting insights and recommendations to improve existing products, services, or processes. The Six Sigma approach also relies on the use of statistical process control and statistically designed experiments. Hence, statistics helps to find the best possible solution for all kinds of practical problems.

To achieve a successful cooperation between practitioners, on one hand, and statisticians, on the other, some openness is required on both sides. Engineers, economists, or scientists need a solid knowledge of the basic principles and techniques of statistics. Statistics is thus an indispensable skill in the repertoire of

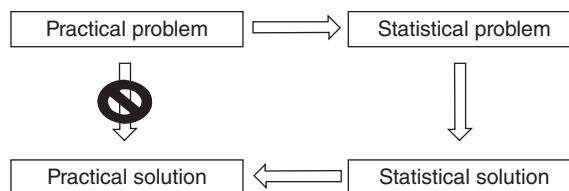


Figure 1.1 Using statistical methods to solve problems.



an effective employee. This explains why statistics is taught not only in the first and second years of many bachelor's degrees in engineering, sciences, and economics, but also later, for example in master's programs.

Finally, a thorough training in statistics is also a prerequisite for students of political and social sciences. They will also be confronted with numerous data sets in their professional careers that are impossible to interpret without a statistical background. For them, statistics is a stepping stone to econometric research methods.

1.2 Definition of statistics

The word statistics may sound familiar to anyone. A statistic usually refers to numerical information, for example, information about

- the population of a country: birth and death rates, immigration and emigration, ... (such statistics are called population statistics),
- the economy: employment and unemployment rates, investments, prices, gross national products (GNP), ... (these statistics are called economic statistics), or
- a company or sector: sales figures, income statements, growth, acquisitions, layoffs, ... (these figures are called business statistics).

More formally, statistics can be defined as the set of methodologies for collecting, representing, analyzing, and interpreting data. This shows that the statistical science is a very general auxiliary science, which plays an important role in almost any environment. Applications of statistics are countless in engineering, medicine, economics, natural sciences, and business management, but statistics is also used in literature, history, political science, criminology, and even musicology.

In our modern society, data is massively present:

- computer files in companies contain sales data, cost data, and customer data (such as addresses, ordered quantities, and order frequencies),
- the financial pages of newspapers contain stock prices, commodity prices, and exchange rates,
- federal and regional authorities regularly publish data on population, trade, and industry, and
- the Internet is a source of numerous data sets.

Companies collect data naturally and actively. Among other things, this takes place by carrying out experiments (e.g., to design new products), in the context of statistical process control, or by measuring all kinds of properties of products, services, and processes. By continuously analyzing data, quality departments of companies attempt to deliver products or services with as few defects as possible and with the highest reliability. In addition, business processes are organized in such a way that waste is minimized, inspections of finished products are reduced to the minimum, and customer requirements are satisfied with minimal costs.



4 STATISTICS WITH JMP

Research agencies collect data via surveys by phone, by post, via the Internet or by street interviews. Such surveys are designed to gather information about the shopping behavior of consumers, about the voting behavior of the population, or public opinion on social issues.

Statistics allows us to turn data into usable information. The role that statistics plays herein may be best illustrated based on some examples.

1.3 Examples

Example 1.3.1 *An airline conducted a study on the behavior of its passengers on intercontinental flights and recorded*

- *the number of passengers with reservations that do not show up (the so-called no-shows),*
- *the weight of the luggage of passengers (often there is a limit of 20 kilograms), and*
- *the time the passengers arrive before the official departure time of the flight (for intercontinental flights, the passengers are asked to be at the airport at least two hours prior to departure).*

The company recorded this data over several months and then made a distinction between passengers in economy class and passengers in business class. The data is analyzed with the aim of instituting appropriate policies. An example may be to allow overbooking, that is, to take more reservations than there are seats on the plane, or to apply more stringent action against passengers carrying too much luggage.

Example 1.3.2 *In the production of coffee, the humidity during production is of crucial importance for the quality of the final product. The humidity is kept under control by a system that does not work flawlessly. Therefore, several measurements of the humidity are taken daily to determine whether it remains within appropriate limits. This approach is referred to as statistical process control.*

Example 1.3.3 *A filling machine for bottles usually has several filling heads, so that many bottles can be filled in parallel. In such a filling process, operators typically weigh a certain number of bottles every hour, to verify that each filling head delivers the desired amount of liquid into the bottles. Another interesting question in this context is whether differences occur between measurements that have been carried out by different operators.*

Example 1.3.4 *Thanks to loyalty cards, supermarkets collect massive data sets. Data that is typically recorded includes*

- *the amount spent per visit at the store, maybe broken down into categories (food, clothing, ...),*



- the number of items sold,
- the payment method (cash, debit card, credit card, or voucher).

Researchers use statistical methods to summarize this huge amount of information and to present it in a way suitable for decision making. Supermarkets exploit this information to send out personalized promotional materials.

Example 1.3.5 Financial analysts are interested in the degree of risk of investing in a particular stock. To this end, they keep track of the monthly return rates of stocks over many years. They take into account price changes, but the dividends as well. Moreover, monthly return rates of the global market, for example the Euro Stoxx 50 index, are tracked. If the return rate of a stock rises or falls to a larger extent than the market, then the share is called risky. In the opposite case, one speaks of a share with little risk. Using statistical methods, one can investigate relations between the return rate of the stock and of the overall market.

1.4 The subject of statistics

In each of the examples in the previous section, the interest is in one or more questions concerning a **population** of objects or elements, or concerning a **process** that generates objects or elements.

The data of the population or process is obtained by recording one or more properties or characteristics of their elements. These properties or characteristics are called **variables**. The name indicates that the value of the property varies from element to element. Therefore, statistics is sometimes referred to as the study of variability.

Usually, it is impossible to include all elements of a population or process in a study. Therefore, one works with a subset of the elements: the **sample**. It is not always easy to collect sample data in a correct way. In any statistical survey, one should pay a lot of attention to the data collection process. In this context, the abbreviation GIGO¹ is often used. This stands for garbage in, garbage out and refers to the fact that any statistical methods can only extract little reliable information from data of poor quality.

Example 1.4.1 For a study of the electoral behavior in European elections, the population can be described easily: all citizens of Europe who are entitled to vote. Variables that could be registered in this context are gender, occupation, political beliefs, age, and so on.

Example 1.4.2 Tossing a die is a process that generates data. A possible sample involves throwing the die 50 times. Variables that could be registered are the number of dots or whether or not the number of dots is even.

In Examples 1.3.2 and 1.3.3, we can consider all times at which the production process is in operation to be the population. At a limited or finite number of points in time,

¹ This abbreviation is a parody of the abbreviations FIFO (first in first out) and LIFO (last in first out), used in accounting for booking items in stock.



6 STATISTICS WITH JMP

measurements or observations are made, for example, measurements of the humidity (Example 1.3.2) or weight (Example 1.3.3). All measurements together form the sample. For the financial analyst in Example 1.3.5, the sample is formed by a finite set of return rates and market indices. In Example 1.3.4, the population of interest for the researcher is the set of all customers of the supermarket. One possible sample consists of all customers that have visited the store during one month and that made use of their loyalty card.

The data collected in a sample can be represented in many ways using tables and graphs. In addition, one can calculate characteristic values or statistics, such as the mean, to generate a clear idea of the collected data. The different ways of presenting sample data are summarized under the term **descriptive statistics**. This topic is covered in Chapters 2 and 3.

In many cases, describing the sample data is only a first step in an investigation. A second phase involves analyzing and interpreting the sample. Analysis and interpretation is necessary in order to find answers to questions about the population or process that were set in advance, to test hypotheses, or to assess the quality of a proposed statistical model. The answers and conclusions obtained from the statistical analysis are generalized to the population or the process. This generalization is called **inference**, which explains the term **inferential statistics**.

The generalization of conclusions from sample data to an entire population or to a process immediately discloses the weakness of statistics: based on sample data, one can never make statements with certainty about the population or process in question. These statements may be considered reliable if statistically valid methods were used for the collection of the sample data. The degree of confidence in a particular statement is expressed by means of a probability, so that a basic knowledge of probability theory is required to be able to understand and apply statistical methods.

1.5 Probability

The words chance and probability sound even more familiar than the term statistics. Intuitively, everyone has a good idea of the meaning of a probability of $1/4$ when participating in a gambling game. Such a probability can be used by virtually everyone to decide whether or not to participate in the game. However, the calculation of such a probability can already raise difficulties.

Probability theory studies processes or experiments in which the outcome is uncertain. Here, the terms process and experiment should be interpreted in their broadest sense. Examples are throwing a die, the price of a share when the stock exchange closes, a mortgage interest rate, the demand for laptop computers of a particular brand, the percentage of defective products in a production line during a certain period, the number of visitors to a website, or drawing a winner from all the participants in a lottery.

The difference between probability and statistics is that, in probability theory, populations and processes are studied directly, while statistics does this through sample data. Probability theory always starts with a set of assumptions about the population or the process. Some examples will illustrate this.



Example 1.5.1 *If the process is throwing a die, then, with the help of probability theory, we can try to figure out the probability of obtaining a six 20 or more times when we toss the die 100 times. This calculation is only possible if we make an important assumption about the die used: the die is fair, or, in other words, the die is completely homogeneous and symmetrical, so that it is equally likely to obtain a one as it is to obtain a two, a three, a four, a five, or a six.*

A statistical question about the die could be to investigate the fairness of the die. The die may be thrown a (large) number of times to collect the required sample data. Based on these data, one can draw a statistical conclusion about the hypothesis that the die is fair.

Example 1.5.2 *In an industrial filling process, one can calculate, based on some assumptions concerning the settings and the accuracy of the filling machine, the probability that a bottle will not be full enough. Another possibility is to calculate the probability that, in a lot with 1,000 bottles, at most 5% of the bottles will not have been filled enough.*

A statistical analysis of the same filling process typically may involve regular weighings of a number of bottles (the sample), in order to verify whether the average content of the bottles is too large or too little, and whether or not the content of the bottles varies too much.

Example 1.5.3 *Using probability theory, one could study the electoral behavior of the European population assuming that 30% will vote for party A, 25% for party B, 20% for party C, and 25% for smaller parties. Probability theory can then calculate that, for every 500 voters, on average 150 will opt for party A, 125 for Party B, 100 for Party C, and 125 will choose other parties.*

Statistics, however, will make a statistical prediction based on a sample of, for example, 2,000 voters. This prediction can also be given with a margin of error.

It is important to realize that statistics works with a limited amount of information obtained from a sample. Therefore, statements about populations and processes can be false. This is the weakness of statistics. Ideally, the probabilities of error are small. The probability for errors can be reduced by collecting a lot of high quality data in a sensible manner.

Probability theory also has a weakness: the assumptions about the studied process or population may be wrong, so that its conclusions are invalid.

1.6 Software

In probability and statistics, a lot of calculations are needed. It is important to create summary tables of all the data in a sample, or to represent the data graphically. This makes the use of a computer and of specialized statistical software necessary. As mentioned in the Preface, in this book, we use the statistical software package JMP[®].