AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

CHAPTER 1

Introduction

IN THIS CHAPTER

Statistics

Data analysis

Case, variable, data value

Quantitative and qualitative variables

Unit of measurement

Missing value

Statistics are numbers. Statisticians use numbers (or statistics) to expand our knowledge of the universe, if only a very small part of the universe. We are all statisticians when we use numbers in this way. This book is about such use of numbers. It is not intended as a comprehensive manual of statistical techniques, but rather as an introduction to the art of statistical thinking.

> By a **statistic** we mean either a number—a numerical piece of information or datum—or a number calculated from a set of data values.

When practicing the *art of statistics,* we use numerical information to increase our knowledge in some way. Used in this sense, statistics refers to the branch of mathematics dealing with theory and techniques of collecting, organizing, and interpreting numerical information.

> By **statistics** we mean either a collection of numerical information, or the branch of mathematics dealing with theory and techniques of collecting, organizing, and interpreting numerical information.

We may use information from a market analysis to select cities for introducing a new product. Or, we might study racing forms to decide how to place a bet in the next horse race. Perhaps we want to examine individual or team performance in major-league baseball. In each of these cases, we study a collection of information, called a *data set.*

> A **data set** is a collection of information.

When we try to make sense of a data set, we are engaging in *data analysis.*

> By **data analysis** we mean making sense of a data set.

Baseball is extremely conducive to data analysis, since baseball statistics are readily available by player and by team. A baseball fan might study individual variables such as batting average: What is a typical batting average for a player in the major leagues? What is an exceptionally good (or poor) batting average? The fan might also examine relationships between variables: What is the relationship between team batting average and winning percentage? Is this relationship different for the American League than for the National League?

Data analysis involves studying variables and relationships between variables in a collection of information. Often we want to do more. We may want to use a sample of information to learn about a larger population. For instance, we might want to use a sample of the thousands of parts produced in a day to decide whether too much gold is being electroplated onto components used in personal computer hardware. Or, we may want to conduct a taste test of two products in a sample of consumers to make decisions regarding product preference in a larger group of consumers. We might want to compare a new treatment with a standard treatment in patients with a particular form of cancer. In each of these cases, it is impractical to study the entire population (parts electroplated in a day, consumers in a product market, or cancer patients). Instead, we look at a sample or subset of the population. We use the information from the sample to learn about the population. This is *statistical inference.*

By **statistical inference** we mean drawing conclusions about a population based on a sample from that population.

The **population** is the group or collection of interest to us.

A **sample** is a subset of the population. We use the observations in the sample to learn about the population.

Data analysis can aid in statistical inference. Medical researchers routinely study characteristics of patients with a particular form of cancer. They look for relationships among such variables as age, sex, stage of illness, response to treatment, and survival.

Estimation is a part of statistical inference. Investigators might use average survival time for patients in a sample to estimate average survival time for all patients in the population. They might then calculate a range of reasonable values for this average survival time. Interpreting such a range of reasonable values, called a confidence interval, depends on ideas in probability.

Statistical inference also involves hypothesis testing. In testing hypotheses, we compare two statements about the state of nature, such as:

Average survival with the new treatment is the same as for the standard treatment.

Average survival with the new treatment is longer than for the standard treatment.

Which of these two statements does the sample support? To decide, we use ideas in probability.

Both estimation and hypothesis testing use probability. We make probability statements about the population based on what we see in the sample. For these statements to make sense, the sample must be similar to the population, a *representative sample.* Suppose the cancer patients in a sample all have very advanced disease. Then researchers cannot make inferences about a larger population that includes patients with less advanced disease. This leads to the idea of experimental design. We want to collect a sample, or carry out an experiment, so that statistical inferences make sense.

1–1 An Overview of the Book

Our study of statistics begins with data analysis. Though the techniques are fairly simple, they can provide a lot of insight into a collection of information.

Some data analysis tools are tabular. A table can summarize certain types of information in a data set. For instance, we might use a table, called a frequency table, to display the number of baseball players with 1991 salaries in each of several intervals (say, less than $500,000, $500,000 to $1,000,000, and so on). Other tools of data analysis are graphical. A histogram, sometimes called a bar graph, is a graphical tool for displaying the information in a frequency table. We could use such a graph to display the information on numbers of players per salary range, instead of listing these numbers in a table.

Chapter 2 discusses tabular and graphical techniques for studying one variable at a time.

Certain information about a variable in a data set can be summarized with a *descriptive statistic.*

A **descriptive statistic** is a number used to summarize information in a set of data values.

For instance, the average is a summary measure that helps to describe the location or center of a set of values. A baseball player's batting average is a measure of his average performance at bat over one or more seasons. The range, or difference between the largest and smallest values, is a summary measure of the variation or spread in a set of values. Saying that baseball player salaries span a range of $5,000 conveys a very different idea of variation than saying that their salaries span a range of $5,000,000! Chapter 3 covers descriptive statistics as measures of location and measures of variation for a single variable.

We can use tabular and graphical tools to examine relationships between two variables, as we will see in Chapter 4. A scatterplot, for instance, is a graphical tool that could be used to display the relationship between player salary and batting average for the 1991 season. In Chapter 5, we extend these ideas to studying more than two variables at a time. Perhaps we want to look at the relationship between baseball player salary and batting average by playing position, or separately for the American League and the National League.

Data analysis techniques can be applied in a wide variety of situations, including most problems involving statistical inference. We make extensive use throughout the text of the data analysis tools we discuss here in Part I.

After data analysis comes probability, in Part II. Statistical inference involves making some assumptions about the sample observations in order to build a probability model. We then use this probability model to make probability statements about the population.

Chapter 6 provides some background information on probability that allows us to build the probability models we need for statistical inference. This chapter also includes some topics in probability of interest for their own sake: odds ratios, used in public health and gambling; and conditional probabilities, used in assessing the usefulness of medical screening procedures.

Some important probability models for statistical inference are based on counting techniques. Chapter 7 discusses some of these counting techniques and introduces two sets of probability models derived from these techniques: the binomial and hypergeometric models. Gaussian models, the basis for most classical statistical inference, are the subject of Chapter 8.

After probability, we move to statistical inference in Part III. The reasoning involved in statistical inference is more formal than in data analysis. Through definitions and examples, Chapter 9 introduces the traditional concepts of statistical inference, including estimation, confidence intervals, and hypothesis testing. This chapter also emphasizes ideas in experimental design. A well-designed experiment yields much valuable information. A poorly designed experiment can result in no information of value.

Often we are interested in making inferences about one or more averages. Does the average gold thickness on electroplated components exceed the target value? Do athletes run faster on average when competing against themselves or against a rival? Which of several treatments is associated with longest survival time, on average? Chapters 10–13 deal with inferences about averages (also called means or measures of central tendency).

Sometimes questions about the variation in a population are at least as interesting as questions about averages. We may know, for instance, that each of several production processes puts the same amount of dog food in 25-pound bags, *on average.* That is, over many bags filled, the average fill weight is very close to 25 pounds per bag. Our concern is with the variation in the processes. A process that results in a weight close to 25 pounds for each bag is preferable to a process with great variation in bag weights, even though both processes have the same *average* fill weight. We say the first process has less variation (a smaller variance) than the second. Chapter 14 discusses inferences about variances.

Often we want to study relationships between variables. What is the relationship between years in the major leagues and baseball player salary? Is this relationship roughly linear? If so, can we assess the extent of the linear association? If not, can we describe the relationship between the two variables in some other way? Such questions are addressed in Chapter 15.

Sometimes variables have categories rather than numerical values (team, league, and playing position in baseball are examples). Inferences involving such categorical variables are the subject of Chapter 16.

1–2 Data Analysis and the World Bank Data Set

Some fairly rigid requirements must be met for formal procedures in statistical inference to be used correctly. Many data sets do not meet these requirements, but they may nevertheless provide valuable information for learning or decision-making. We can use the techniques of data analysis to study such a data set. These techniques are tools, not ends in themselves. We will emphasize flexibility by showing different uses and formats for a number of techniques. The goal is not to define every possible technique, but to convey the spirit of data analysis. Since no two collections of data are the same, there is no rigid procedure for analysis. Rather, we approach a data set, armed with some techniques and a desire to explore.

It is easy to come out of an introductory statistics course without any idea of what to do with a real data set, because data analysis techniques are often defined as separate entities and illustrated with different examples. They are not presented as a package, each technique with its use in examining a single collection of data. We will try to avoid this dilemma by using a single example, based on a World Bank data set, throughout the next four chapters. For variety, the exercises cover problems from many fields of application.

Our World Bank data set is a collection of indicators of social and economic development for 128 countries with populations of 1 million or more

(World Bank, 1987). We will study these indicators to address such questions as: What is the range of economic development over these World Bank countries? What is the relationship between level of economic development and indicators of quality of life such as calorie supply, percentage of school-age children attending school, and life expectancy? How do birth rates and fertility rates relate to level of education among females and extent of contraception use?

The countries in our World Bank data set are shown on the map in Figure 1-1. The indicators are defined in the first appendix at the end of this chapter. The 128 countries are listed by economic category in Table 1-1. Economic designations, political boundaries, and country names change over time. We are using designations defined by the World Bank in 1987.

Thirty-seven countries are classified as low-income developing nations. Middle-income developing nations include 36 lower-middle-income countries and 23 upper-middle-income countries. Four nations are listed as high-income oil exporters and 19 as industrial market countries. There are 9 countries classified as nonmember nations because of lack of reliable economic information.

The indicators compiled by the World Bank constitute a *data set,* a collection of information. A *case* is an individual sampling unit, the subject of measurement.

A **case** is the individual sampling unit in a data set; it is the basic unit sampled and subject of measurement.

In the World Bank data set, each country is a case because a country is the basic unit for which information is recorded. In a medical study comparing two treatments for lung cancer, case refers to an individual patient participating in the study. If baseball statistics are compiled by team, a team is a case. If statistics are compiled by player, a player is a case. In a quality control setting, a case is an individual object subjected to testing.

A *variable* is a particular piece of information recorded for a case. Many variables may be available for each case. Variables in the World Bank data set include the name of the country, economic category, gross national product per capita, life expectancy, and birth rate. A *data value* is the value a variable takes on for a particular case. For example, one data value for life expectancy is 76 years, the number recorded for the United States.

Variable refers to a particular piece of information recorded for a case.

A **data value** is the value a variable takes on for a particular case.

A *quantitative variable* has numerical values that are measurements or counts. Examples of quantitative variables are life expectancy, population size, number of cities with over 500,000 people, calorie supply per capita, primary school enrollment, birth rate, and percentage of contraception use.

A **quantitative variable** has numerical values that are measurements or counts.

A *qualitative variable,* or *categorical variable,* has values with no intrinsic meaning as numbers. The name of a country is a qualitative variable, as is region of the world. A country is in North America, Africa, Western Europe, and so on. These values are categories, with no numerical meaning.

A **qualitative,** or **categorical, variable** has values that cannot be interpreted as numbers.

The World Bank classifies countries by whether they are oil exporters, exporters of manufactured goods, or highly indebted. Each of these variables

TABLE 1-1 Listing of 128 countries by economic category. Listings are alphabetical within categories.

Low-income countries

Afghanistan
Bangladesh
Benin
Bhutan
Burkina Faso
Burma
Burundi
Central African Republic
Chad
China
Ethiopia
Ghana
Guinea
Haiti
India
Kampuchea
Kenya
Lao PDR
Madagascar
Malawi
Mali
Mozambique
Nepal
Niger
Pakistan
Rwanda
Senegal
Sierra Leone
Somalia
Sri Lanka
Sudan
Tanzania
Togo
Uganda
Viet Nam
Zaire
Zambia

Lower-middle-income countries

Bolivia
Botswana
Cameroon
Chile
Colombia
Congo
Costa Rica
Côte d'Ivoire
Dominican Republic
Ecuador
Egypt
El Salvador
Guatemala
Honduras
Indonesia
Jamaica
Jordan
Lebanon
Lesotho
Liberia
Mauritania
Mauritius
Morocco
Nicaragua
Nigeria
Papua New Guinea
Paraguay
Peru
Philippines
Syria
Thailand
Tunisia
Turkey
Yemen, Arab Republic
Yemen, PDR
Zimbabwe

Upper-middle-income countries

Algeria
Argentina
Brazil
Greece
Hong Kong
Hungary
Iran
Iraq
Israel
Korea, South
Malaysia
Mexico
Oman
Panama
Poland
Portugal
Romania
Singapore
South Africa
Trinidad and Tobago
Uruguay
Venezuela
Yugoslavia

High-income oil exporters

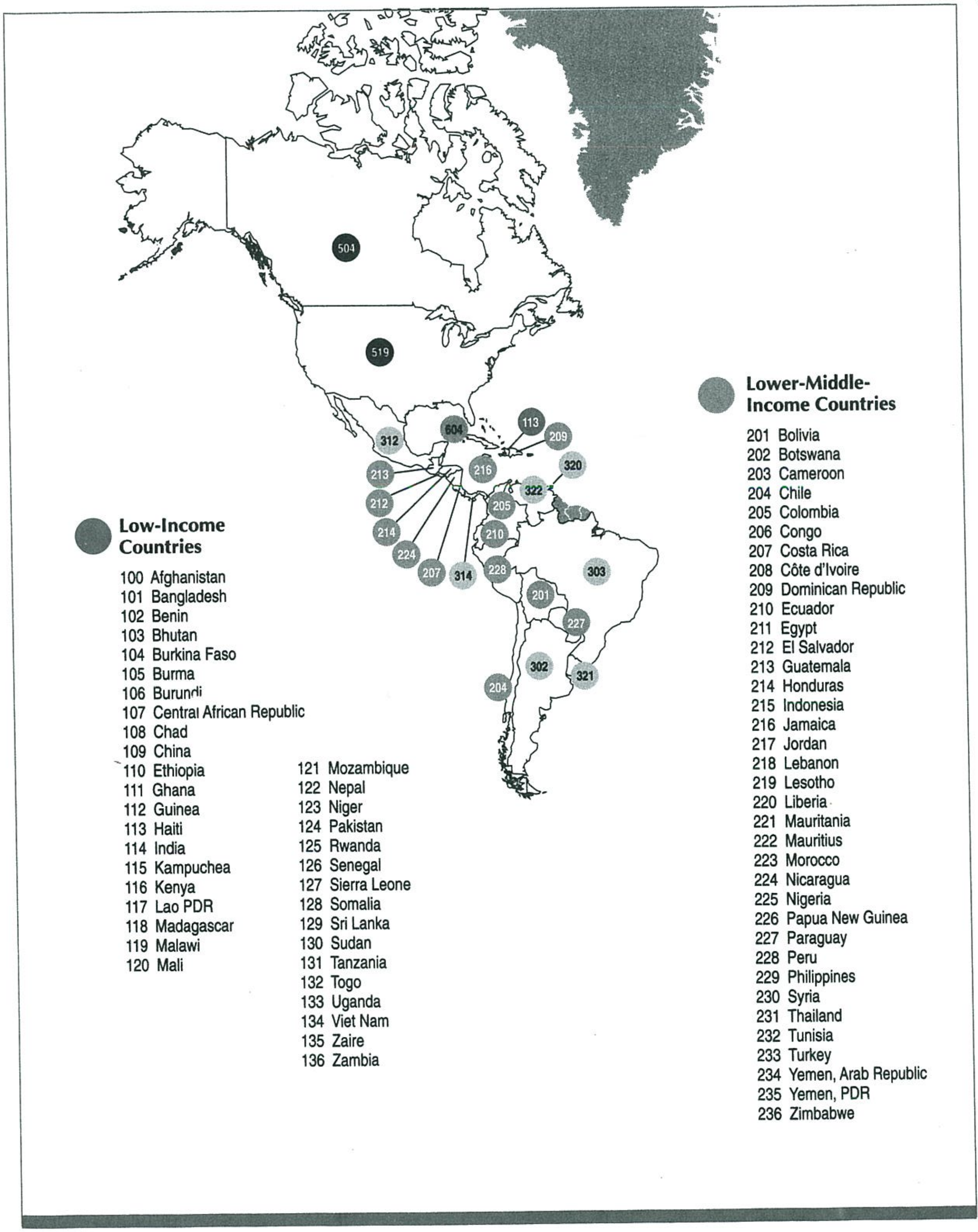
Kuwait
Libya
Saudi Arabia
United Arab Emirates

Industrial market countries

Australia
Austria
Belgium
Canada
Denmark
Finland
France
Germany, West
Ireland
Italy
Japan
Netherlands
New Zealand
Norway
Spain
Sweden
Switzerland
United Kingdom
United States

Nonmember countries

Albania
Angola
Bulgaria
Cuba
Czechoslovakia
Germany, East
Korea, North
Mongolia
USSR



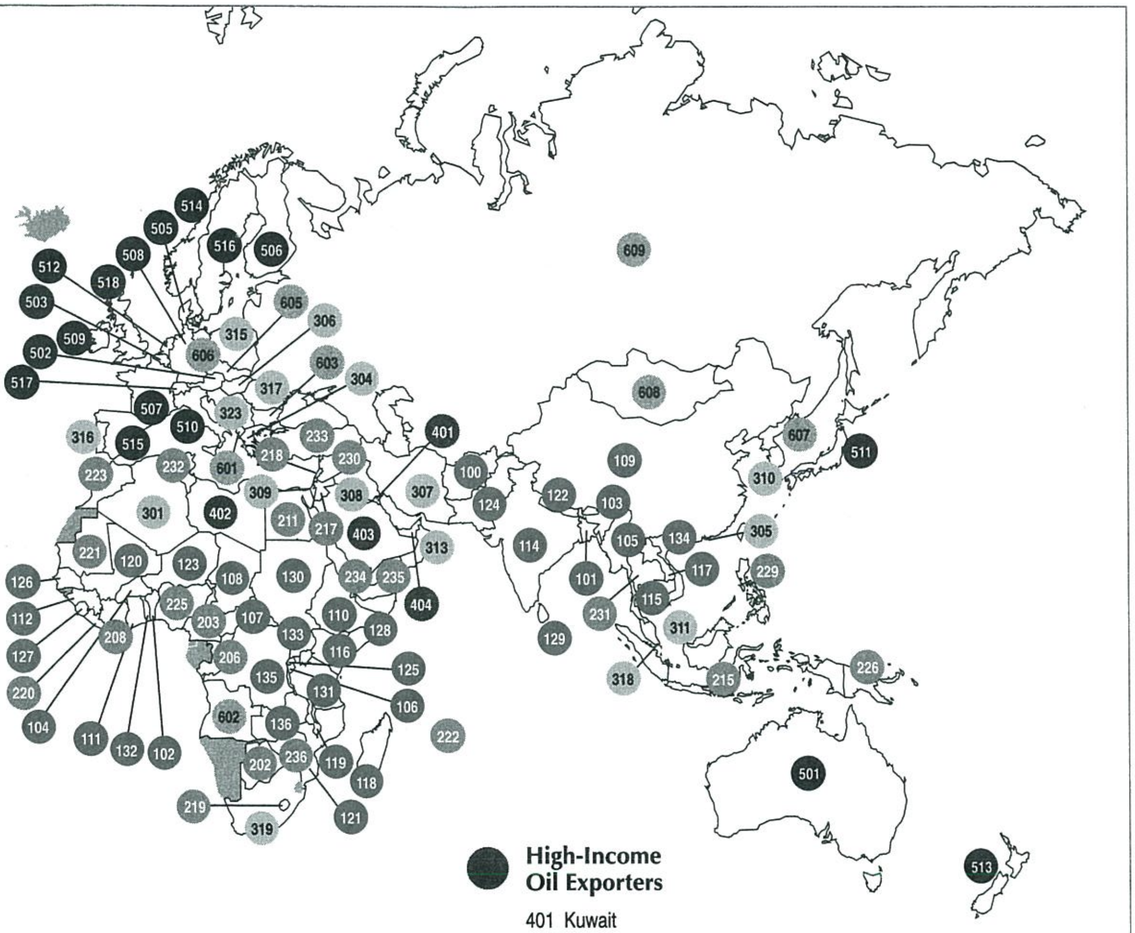
● **Low-Income Countries**

100 Afghanistan
101 Bangladesh
102 Benin
103 Bhutan
104 Burkina Faso
105 Burma
106 Burundi
107 Central African Republic
108 Chad
109 China
110 Ethiopia
111 Ghana
112 Guinea
113 Haiti
114 India
115 Kampuchea
116 Kenya
117 Lao PDR
118 Madagascar
119 Malawi
120 Mali
121 Mozambique
122 Nepal
123 Niger
124 Pakistan
125 Rwanda
126 Senegal
127 Sierra Leone
128 Somalia
129 Sri Lanka
130 Sudan
131 Tanzania
132 Togo
133 Uganda
134 Viet Nam
135 Zaire
136 Zambia

● **Lower-Middle-Income Countries**

201 Bolivia
202 Botswana
203 Cameroon
204 Chile
205 Colombia
206 Congo
207 Costa Rica
208 Côte d'Ivoire
209 Dominican Republic
210 Ecuador
211 Egypt
212 El Salvador
213 Guatemala
214 Honduras
215 Indonesia
216 Jamaica
217 Jordan
218 Lebanon
219 Lesotho
220 Liberia
221 Mauritania
222 Mauritius
223 Morocco
224 Nicaragua
225 Nigeria
226 Papua New Guinea
227 Paraguay
228 Peru
229 Philippines
230 Syria
231 Thailand
232 Tunisia
233 Turkey
234 Yemen, Arab Republic
235 Yemen, PDR
236 Zimbabwe

FIGURE 1-1 Map showing name and economic category for 128 countries with populations of 1 million or more as identified in the *World Development Report* 1987 (World Bank, 1987)



● **Upper-Middle-Income Countries**

301 Algeria
302 Argentina
303 Brazil
304 Greece
305 Hong Kong
306 Hungary
307 Iran
308 Iraq
309 Israel
310 Korea, South
311 Malaysia

312 Mexico
313 Oman
314 Panama
315 Poland
316 Portugal
317 Romania
318 Singapore
319 South Africa
320 Trinidad & Tobago
321 Uruguay
322 Venezuela
323 Yugoslavia

● **High-Income Oil Exporters**

401 Kuwait
402 Libya
403 Saudi Arabia
404 United Arab Emirates

● **Industrial Market Countries**

501 Australia
502 Austria
503 Belgium
504 Canada
505 Denmark
506 Finland
507 France
508 Germany, West
509 Ireland
510 Italy
511 Japan
512 Netherlands
513 New Zealand
514 Norway

515 Spain
516 Sweden
517 Switzerland
518 United Kingdom
519 United States

● **Nonmember Countries**

601 Albania
602 Angola
603 Bulgaria
604 Cuba
605 Czechoslovakia
606 Germany, East
607 Korea, North
608 Mongolia
609 USSR

has two possible values, yes and no, which can be represented by numbers such as 1 and 0. Since the numbers merely represent categories, these are qualitative variables.

Sometimes a qualitative variable has ordered categories. Four economic categories are ordered by level of economic development: low income, lower-middle income, upper-middle income, industrial market. A country might be classified by birth rate as low, moderate, or high. Such categories are ordered, but without exact numerical meaning. Such a variable is often called an *ordinal qualitative variable.*

> An **ordinal qualitative variable** has categories with a natural ordering, but not exact numerical values.

Unit of measurement is the basic unit used for measuring and recording the value of a variable. Life expectancy for the United States is 76 years; year is the unit of measurement for life expectancy. Per capita gross national product is reported in U. S. dollars; U. S. dollar is the unit of measurement for gross national product. Units of measurement should always be included in reports of numerical information.

> The **unit of measurement** of a variable is the basic unit (such as inches, years, dollars) used to measure and record the values of the variable.

With this introduction to the World Bank data set and some terminology, we can talk about how to get started in data analysis.

1–3 Questions to Ask Before Starting Data Analysis

We should ask some questions before starting any data analysis. The answers will help us decide whether an analysis is worthwhile. They will also help to guide us as we begin. Here we address some of these questions in terms of the World Bank data set.

What are the goals of our analysis? What are we trying to learn? There are times when we know very little about the variables and relationships between the variables. Then we might state a vague goal such as: Learn about the variables and about relationships between these variables. It is important to be more specific when possible. We can be more specific in describing the goals of our analysis of the World Bank indicators.

First, we want to study each variable separately, looking at characteristics of the variable for the entire group of countries. Consider the variable life expectancy: How do life expectancies vary across countries? Are the life expectancies clustered around some center value? Are there two or more clusters of countries with similar life expectancies? Are life expectancies scattered across a wide range?

Second, we want to study each variable separately within economic categories. We also want to compare results across economic categories. For life expectancy: How do life expectancies vary across the low-income countries? How do life expectancies for the low-income countries differ from life expectancies among the industrial market countries?

Finally, we want to look at relationships between variables, for all countries and within economic categories. Is there any relationship between female primary school enrollment and birth rates? We want to answer such questions overall and within economic categories.

Can we use the data set to meet the goals of our analysis? This is a question too often overlooked by investigators and data analysts. Many times, the variables recorded cannot legitimately meet the goals of the investigation. Or the data are so poorly collected or full of errors that no analysis is worthwhile. We should always carefully consider what questions our collection of data can really answer, setting limits on interpretations and reported results.

There are many limitations of our data, some clearly stated in the *World Development Report* 1987 (World Bank, 1987). Many indicators are based on information supplied by individual countries. Methods and quality of collecting and reporting vary greatly across countries. Some deliberate misreporting may occur. These problems make comparisons across countries hard to interpret.

Even under the best conditions, some indicators may be difficult to determine with accuracy. It cannot be easy, for example, to estimate the percentage of married women of childbearing age who use contraception.

Our indicators are countrywide averages. Conditions can vary widely within a country. Life expectancy, for example, is reported as 56 years for India and 76 years for the United States. This suggests that lifespans are shorter in India than in the United States. However, there are well-to-do people in India with advantages similar to those available to well-off Americans; life expectancies for these people are probably much longer than 56 years. Likewise, there are Americans who do not have access to these advantages; their life expectancies may be much less than 76 years. Thus, even when we trust the quality of our information, we have to be careful when making interpretations.

Recognizing the limitations of our data, we can set limits on our interpretations of analyses. We will use the indicators to compare countries in a general way. But we will not attach great importance to exact differences. We will also recognize that conditions within countries may vary greatly.

We must also be cautious because there are *missing values* for many variables in the World Bank data set. The worst case is for percentage of married women of childbearing age using contraception in 1984. Thirty-three, or about one-fourth, of the countries have no value recorded for this variable. Three countries—Afghanistan, Kampuchea (Cambodia), and Lebanon—have missing values for most of the variables. We must recognize that any results we obtain apply only to countries with available information. When a variable has many missing values, comparisons across economic categories will be even more suspect and interpretations more tentative than they would be otherwise.

A case has a **missing value** for a variable if no information on the variable is available for that case.

Each of the World Bank indicators corresponds to a single year or short time period. Therefore, we can make no analyses of possible time trends in the variables. All of our results will apply only to the year or time period for which the variables are recorded.

How do we get a data set ready for analysis? Data must be collected and compiled before we can analyze it. Each variable should be measured and recorded correctly. We must avoid mistakes in transcribing information from data collection forms to a computer file. All this care involves proofreading and error checking. Although tedious, proofreading and error checking are important in statistics and data analysis. Analysis may be worthless if there are mistakes in the data.

Since our data come from tables compiled by the World Bank, error checking in collection and compilation is out of our control; all we can do is be aware of possible inadequacies. We should be careful with transcribing, as from tables to computer files, from computer printouts to reports. We will also be cautious when interpreting the results of our analysis.

Some Suggested Reading and a Caution

In the next four chapters, we will use some simple data analysis techniques to study the World Bank data. The chapters describe and report the results of some analyses; other analyses are left to the exercises. Hopefully, these chapters not only introduce some useful concepts in data analysis, but also describe some interesting aspects of the World Bank data we are exploring. For more on using data analysis to learn about a data set, see *Exploratory Data Analysis* (Tukey, 1977); *Data Analysis and Regression* (Mosteller and Tukey, 1977); *Applications, Basics and Computing of Exploratory Data Analysis* (Velleman and Hoaglin, 1981); and *Understanding Robust and Exploratory Data Analysis* (Hoaglin, Mosteller, and Tukey, editors, 1983).

The World Bank data set provides an extended example for use in Chapters 2–5. We look at some World Bank indicators again in Section 15-1 when we consider the correlation coefficient as a descriptive statistic measuring linear association between two quantitative variables. You may notice that we do not mention the World Bank data set anywhere else in Part III, when we discuss statistical inference. The reason is this: Formal statistical inference involves using a carefully selected sample to learn about a larger population. The World Bank data set contains information *on an entire population* of countries: 128 nations with 1 million or more inhabitants. Data analysis is very appropriate for this collection of World Bank indicators, but statistical inference is not

Summary of Chapter 1

Data analysis involves the use of tables, graphs, and descriptive statistics to study variables and relationships between variables in a data set. It is extremely important to formulate as precisely as possible what it is we are trying to learn from a given data set, stating goals and recognizing limitations.

Statistical inference involves more formal reasoning than data analysis. We use a carefully selected sample to make probability statements about a larger population. The validity of these statements depends on certain assumptions being met. Careful experimental design allows us to collect samples of observations that allow valid inferences about the population sampled.

Appendix to Chapter 1: The World Bank Indicators

The data come from the World Development Indicators section of the *World Development Report* 1987 (World Bank, 1987). Unless stated otherwise, variables are for 1985.

Birth rate: Birth rate is the number of live births per 1,000 population. Units are live births per 1,000 population.

Calorie supply: Daily calorie supply per capita is the calorie equivalent of a country's food supply in 1985 divided by its population size. Units are calories per person.

Child death rate: Child death rate is the estimated number of deaths of 1- to 4-year-old children per 1,000 children in this age group. Units are deaths of 1- to 4-year-olds per 1,000 1- to 4-year-old children.

Contraception use: Percentage of married women of childbearing age using contraception in 1984 covers women who are practicing contraception and women whose husbands are practicing contraception. Childbearing age is generally defined as 15–44 years, although some countries use intervals 18–44, 15–49, or 19–49. Unit of measurement is percent.

Death rate: Death rate is the number of deaths per 1,000 population. Units are deaths per 1,000 population.

Fertility rate: Total fertility rate is the number of children born to a woman if she lives to the end of her childbearing years and if the number of children she bears at each age corresponds to the current age-specific fertility rate for her country. Fertility rate estimates the average number of children born per woman in a country. Units are children per woman.

Gross domestic product: Gross domestic product, reported in millions of U. S. dollars, is an estimate of the total market value of all goods and services produced by a country.

Gross national product: Per capita gross national product, reported in U. S. dollars, is an estimate of the average market value of all goods and services produced per person, with an adjustment for income from abroad. The method of calculating per capita gross national product is described in the *World Development Report* 1987, with emphasis on difficulties in comparing values of this indicator across countries.

Higher education enrollment: Number enrolled in higher education as percentage of age group is the number of people enrolled in higher education divided by the number of 20- to 24-year-olds in 1984, times 100. Unit of measurement is percent.

Infant mortality rate: Infant mortality rate is the estimated number of infants who die during their first year of life, per 1,000 live births. The units are deaths of infants in their first year per 1,000 live births.

Life expectancy: Life expectancy at birth is the estimated number of years a baby would be expected to live if mortality patterns prevailing in 1985 continued throughout his or her life. Unit of measurement is year. Male life expectancy is the estimate for males only; female life expectancy is the estimate for females only.

Number of cities of over 500,000 *people in* 1980*:* The unit of measurement is city.

Population growth: Percentage average annual population growth for 1980–1985 is calculated from mid-year population totals. Unit of measurement is percent.

Population size: Population size is the estimated total population of a country in the middle of 1985. Refugees who are not permanently settled are credited to their country of origin in the population counts. Population totals are given in millions of people.

Primary school enrollment: Number enrolled in primary school as percentage of age group is the number of people enrolled in primary school divided by the number of 6- to 11-year-olds in 1984, times 100. Countries differ in ages and duration of primary schooling. It is possible for this variable to have a value greater than 100% if some children in primary school are above or below the standard primary school age. Unit of measurement is percent. Male primary school enrollment is estimated for males only; female primary school enroliment is estimated for females only.