A Guide to Working with Economic Data

This Guide *provides information about various statistical and data-handling techniques that are useful in understanding the text and working the problems at the end of each chapter. Although the* Guide *can be read straight through, the sections are relatively self-contained, so that it may be consulted on particular points as necessary. Cross-references to the* Guide *are provided in the main text and at the beginning of each problem set.*

G.1 Economic Data

G.1.1 VARIABLES

Economic data are represented by variables that name a particular economic concept. For example, in Tables G.1 and G.2 report data for the G-7 countries corresponding to the concept "Real GDP" and "Employment." These may be designated by variables: real GDP by $Y_{i,t}$ and employment by $E_{i,t}$. The subscript i indicates to which unit the data refers (here a G-7 country: Canada, France, etc.), and the subscript t indicates the time period (here 1991, 1992, etc.). For example, $Y_{Japan,1996}$ = \$1,560 billion; $E_{France,\ 2000}$ = 23,262,000. Variables are sometimes written as functions rather than with subscripts; so that $Y(i,t) = Y_{i,t}$. For example, Y(Italy, 1997) = \$426 billion.

Units, of course, need not be countries. They could be states, families, firms, individual people, particular financial assets, or, in fact, any category that can be measured. Time may be divided, as here, into years; but also, commonly, into quarters,

Table G.1
Real GDP in the G-7 Countries
(1996 constant U.S. dollars (chain indexed) at purchasing power parity)

	Canada	France	Germany	Italy	Japan	United Kingdom	United States
1991	277	444	764	422	1,455	474	3,027
1992	274	429	764	425	1,475	463	3,115
1993	281	403	738	400	1,477	467	3,216
1994	299	434	748	402	1,486	492	3,431
1995	310	407	758	411	1,502	510	3,548
1996	316	451	760	416	1,560	528	3,699
1997	336	419	765	426	1,602	555	3,911
1998	355	477	782	438	1,567	577	4,087
1999	376	454	808	451	1,561	594	4,289
2000	401	520	837	462	1,591	617	4,558

Source: Penn-World Tables, 6.1, Table G.2, and author's calculations.

Table G.2
Employment in the G-7 Countries
(thousands)

	Canada	France	Germany	Italy	Japan	United Kingdom	United States
1991	12,916	22,316	37,445	21,595	63,690	26,400	116,877
1992	12,842	21,609	36,940	21,609	64,360	25,812	117,598
1993	13,015	20,705	36,380	20,705	64,500	25,511	119,306
1994	13,292	21,875	36,075	20,373	64,530	25,717	123,060
1995	13,506	20,233	36,048	20,233	64,570	26,026	124,900
1996	13,676	22,311	35,982	20,320	64,860	26,323	126,709
1997	13,941	20,413	35,805	20,413	65,570	26,814	129,558
1998	14,326	22,479	35,860	20,618	65,140	27,116	131,463
1999	14,531	20,864	36,402	20,864	64,623	27,442	133,488
2000	14,910	23,262	36,604	21,225	64,464	27,793	136,891

Source: International Monetary Fund, *International Financial Statistics.*

months, weeks, days, or larger or smaller divisions. Measurements may be made at a particular period (e.g., the last day of the quarter) or as an average across the period.

Months are frequently indicated by the year followed by a colon and a number: 01 for January up to 12 for December. Quarters are frequently indicated similarly: 1 for the first quarter, 2 for the second, and so forth. For example, July 2001 is indicated as 2001:07; while the third quarter of 1992 is indicated as 1992:3.

Typically, an economic variable tracks different units over a single time period or the same unit over different time periods. Data for which time is held fixed and each point refers to a separate unit are called **cross sections**. The fourth line in Table G.1 is the cross section of G-7 real GDP for 1994. When context makes it clear that a cross section is intended, then, to avoid clutter, the variable usually omits the *t* subscript. For example, Y_i is adequate notation for the last cross section, so long as we keep in mind that it applies to 1994.

Data for which time is held fixed and each point refers to a separate unit are called **time series**. The third column in Table G.2 is the time series of employment in Germany. When context makes it clear that a series is intended, then the variable usually omits the *i* subscript. For example, E_t is adequate notation for the last time series, so long as we keep in mind that it applies to Germany.

G.1.2 THE DIMENSIONS OF DATA

Keeping Track of Units

The units in which data are measured are an essential part of economic variables and must be tracked. Often units are abbreviated where it causes no confusion. For example, in the last section, Japanese real GDP in 1996 was reported as $1,560 billion: the stated unit is "billions of dollars," where the true unit is "billions of 1996 constant U.S. dollars (chain indexed) at purchasing power parity." Writing that every time would be clumsy, nevertheless, we must always be careful to make such relevant information available (as it is in Table G.1).

Dimensions follow the same algebra as variables themselves. Suppose that we wish to compute real GDP per capita: $YH_{i,t} = Y_{i,t}/E_{i,t}$. Its units are just the units of real GDP divided by the units of employment, so that $YH_{Canada,1998}$ = $355 billion/14,326 thousand = 0.024780 $\frac{billion\ dollars}{thousand\ people}$. Here, as often happens, the units that follow naturally from the calculation are awkward, and may be restated in a more convenient form. Usually, scientific notation is helpful: $YH_{Canada,1998} = (\$355 \times 10^9)/(14{,}326 \times 10^3$ people) = $0.024780 \times 10^6 \frac{dollars}{person} = 24{,}780 \frac{dollars}{person}$. Or, to put it more naturally, $YH_{Canada,1998}$ = $24,780 per person (or per capita or per head).

Example G.1: What are the units of the price of a Mercedes car and what are the units of the proceeds from selling 10 Mercedes at $50,000 each?
Answer: Units of price: dollars/car; units of proceeds: (dollars/car) × cars = dollars.

Example G.2: If 50,000 workers each work 2,000 hours over the course of a year, what is the rate of total work expressed in correct units?
Answer: 100,000,000 worker-hours per year.

Stocks and Flows

A **stock** is a quantity measured at a point in time, so that its units are the units of the quantity; while a **flow** is a quantity per unit time. As a result a *flow* × *a period of time* has the dimensions of (quantity/time) × time = quantity. Water flowing through a faucet at 5 gallons per minute for 5 minutes adds 25 gallons to the stock of water in a bathtub. The distinction between stocks and flows is particularly important in economics and is discussed in detail in Chapter 2, section 2.2.1.

Annualization and Aggregation

Even when time is measured in units of, say, months or quarters, it is often convenient to convert the results to units of years. Such conversion is called **annualization** and is no different in principle than converting the speed of a sprinter from seconds per quarter mile to miles per hour: e.g., 55 second per quarter mile = 16.36 mile per hour. To take an economic example, real GDP in the United States is measured quarterly, but annualized before it is published: the quarterly number is simply multiplied by 4. For example, real GDP in the fourth quarter of 2004 was $2,748.33 billion per quarter = $10,993.3 billion per year. In the United States, only the annualized figure is published. In some other countries, GDP data is published at quarterly rates.

Example G.3. The officially published nominal GDP for the United Kingdom for the third quarter of 2004 was £290.7 billion per quarter. What is it annualized?
Answer: £1,162.8 billion per year.

Often, we want to know the average flow of a quantity over a period larger than its **frequency of observation**. For example, we may want to know real GDP for 1993 – that is, for the whole year, not for one quarter annualized. Moving from *higher frequency data* (that is, data observed more often) to *lower frequency data* is called **temporal aggregation**. With British data, since it is not annualized, the problem is straightforward: add up each quarter. With American data, each point has already been multiplied by four, so that just adding them up would produce a number four times too big. We must, therefore, add them up and divide by four.

Example G.4. U.S. real GDP for the four quarters of 2004 was: $10,697.5, $10,784.7, $10,891.0, $10,993.3 (each measured in billions at an annual rate). What is real GDP for 2004?
Answer: ($10,697.5 + $10,784.7 + $10,891.0 + $10,993.3)/4 = $10,841.6 billion.

Example G.5. U.K. nominal GDP for the four quarters of 2003 was: £277.6, £276.2, £282.3, £286.5 (each measured in billions at a quarterly rate). What is nominal GDP for 2003?
Answer: £277.6 + £276.2 + £282.3 + £286.5 = £1,122.6 billion.

Percentages and Percentage Points

Percent refers to a fraction or share expressed in hundredths. In fact, the percent sign (%) is a stylization of /100, the denominator of a fraction in hundredths. Any *natural* fraction or decimal can be properly read as a percentage: 0.047 *is* 4.7 percent, just as 12 inches is one foot.

Percentage point refers to the unit one hundredth (1/100). While one may properly read a fraction either as a natural number or a percentage, when making calculations a natural fraction must be multiplied by 100 to convert it into percentage points. Such a conversion is not part of the fundamental calculation but merely a way of

selecting preferred units. (In this book, this unit-conversion step is always omitted. It cannot, however, be omitted in making calculations. [Excel hint: to express fractions as percentages either multiply by 100 or click "Format," "Cell," and then click "Percentage" on the "Number" tab]).

Special care must be taken not to confuse *percent* and *percentage point*. For example, an interest rate is expressed as a percentage of the principal – say, 5.2 percent per year. If the interest rate rises to 6.4 percent per year, it has risen by 1.2 percentage points (= 6.4 – 5.2) but it has risen by 23 percent (= (6.4 – 5.2)/5.2 = 0.23).

G.1.3 SEASONAL ADJUSTMENT

Many economic series vary according to the time of year in a regular way. Sometimes we are interested in this variation in which case we use **non-seasonally adjusted** data. In other cases, this variation is confusing. For example, if we are concerned about whether the economy is entering a recession, then the fact that non-seasonally adjusted GDP slows in the first quarter every year would confuse our judgment. What we really want to know is whether it has slowed *more than usual* this year. **Seasonally adjusted data** transforms the non-seasonally adjusted data to account for the usual seasonal variation. There are many ways of seasonally adjusting data each differing by how they determine what is usual.

Typical seasonal adjustment is displayed in Table G.3, which shows the CPI for food for 2001 in its seasonally adjusted and non-seasonally adjusted forms. The third

G.3
An Illustration of Seasonal Adjustment: Consumer Food Prices in 2001

Month	Seasonally Adjusted	Non-seasonally Adjusted	Seasonal Factor	Percentage Adjustment
January	170.4	170.9	0.997	-0.29
February	171.2	171.3	0.999	-0.06
March	171.6	171.7	0.999	-0.06
April	172.0	171.9	1.001	+0.06
May	172.5	172.5	1.000	0.00
June	173.0	173.0	1.000	0.00
July	173.6	173.5	1.001	+0.06
August	174.0	173.9	1.001	+0.06
September	174.2	174.1	1.001	+0.06
October	174.9	174.9	1.000	0.00
November	174.9	174.6	1.002	+0.17
December	174.7	174.7	1.000	0.00
Total	2077	2077	12.000	-0.01[1]

[1]Does not sum to zero due to rounding error.
Note: CPI-U food (1982-84 = 100).
Source: Bureau of Labor Statistics.

column shows the **seasonal factor** – that is the number that converts non-seasonally adjusted to seasonally adjusted data according to the formula:

(G.1) *Seasonally adjusted variable = Non-seasonally adjusted variable* × *Seasonal factor*.

For example, food and prices tend to be high in January, when many fresh foods are out of season, so that a seasonal factor less than one adjusts the value down. The downward adjustment is 0.29 percent as shown in the fourth column. Similarly, from mid-summer to late autumn (July to November) food prices tend to be low as fresh food production is high, so that the seasonal factors greater than one adjust the values upward. The main differences between different methods of seasonal adjustment are (1) how they estimate the seasonal factors and (2) whether they use multiplicative factors (as most do) or additive adjustments.

Notice that the sum of the seasonally adjusted and non-seasonally adjusted prices in Table G.3 are the same and that the seasonal adjustments factors add to 12, the number of months. Both reflect the fact that seasonal adjustment methods generally reallocate data across a year, but do not change it year to year. The annual sum of prices is not an economically meaningful number, but in other cases (for example, GDP) this is a very important property.

G.2 Graphs

Spreadsheets generally support a large variety of types of graphs. Here we consider only some representative examples. (*Chart*, *figure*, *plot* are among the terms that are commonly used as synonyms for *graph*.)

G.2.1 Cross-sectional Graphs

Univariate Cross-sectional Graphs

A *bar chart* is a typical means of displaying a single (*univariate*) cross section. Figure G.1 shows G-7 employment in 1995; it graphs one line from the Table G.2. Other univariate cross-sectional graphs include the *pie chart*, especially useful for emphasizing relative shares (see, for example, Chapter 2, Figure 2.8). There are a large number of other types, each helpful in stressing particular aspects of the data. [Excel hint: to make a graph highlight the data, then click "Insert," "Chart," and choose from among the menu of chart types.]

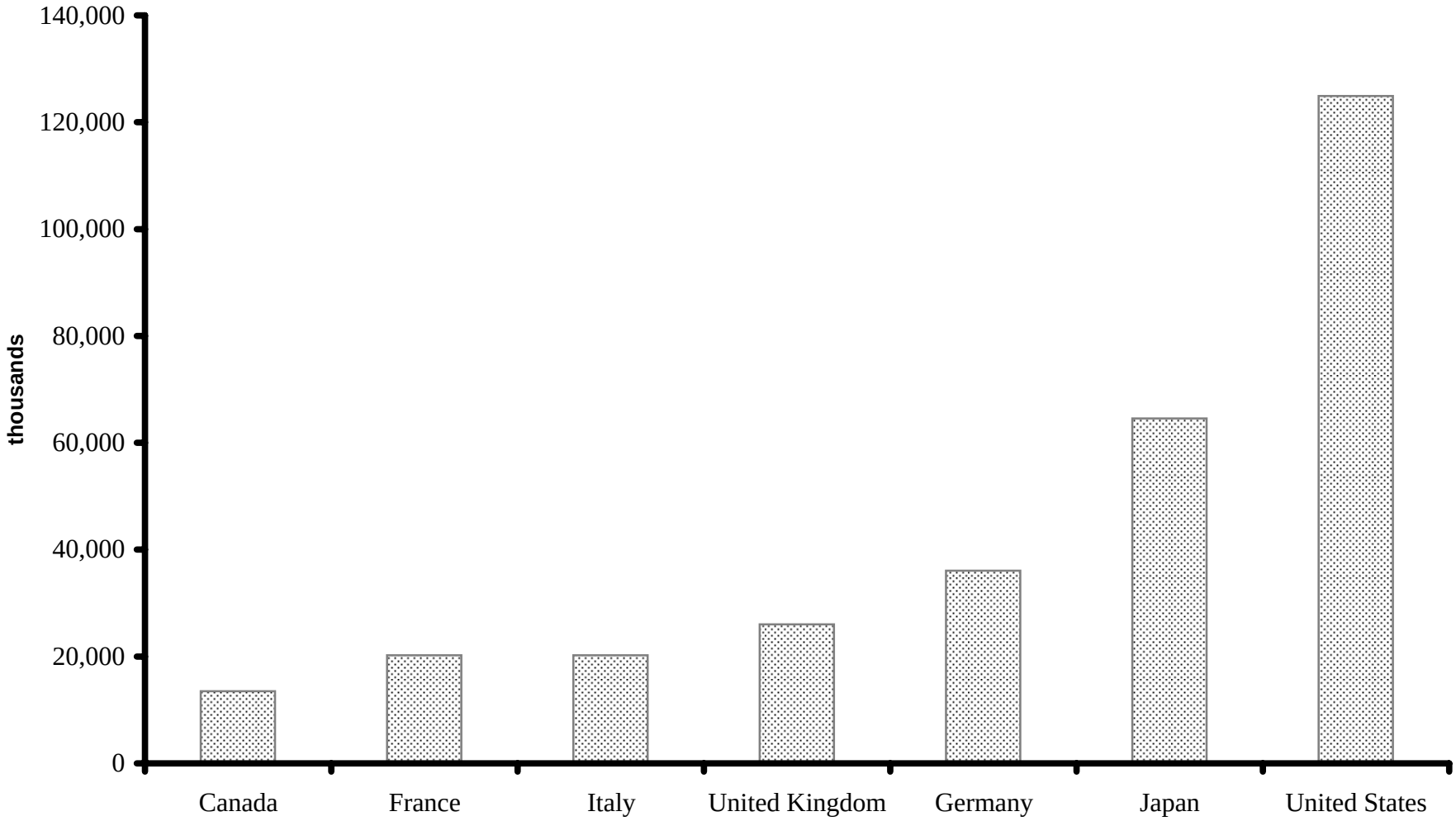
Multivariate Cross-sectional Graphs

"[*T*]*he relational graphic – in its barest form, the scatterplot and its variants – is the greatest of all graphical designs*."[1]

Related cross sections can be displayed in a variety of graphical forms. The most important is the **scatterplot**. Figure G.2 plots the data displayed in Figure G.1 for

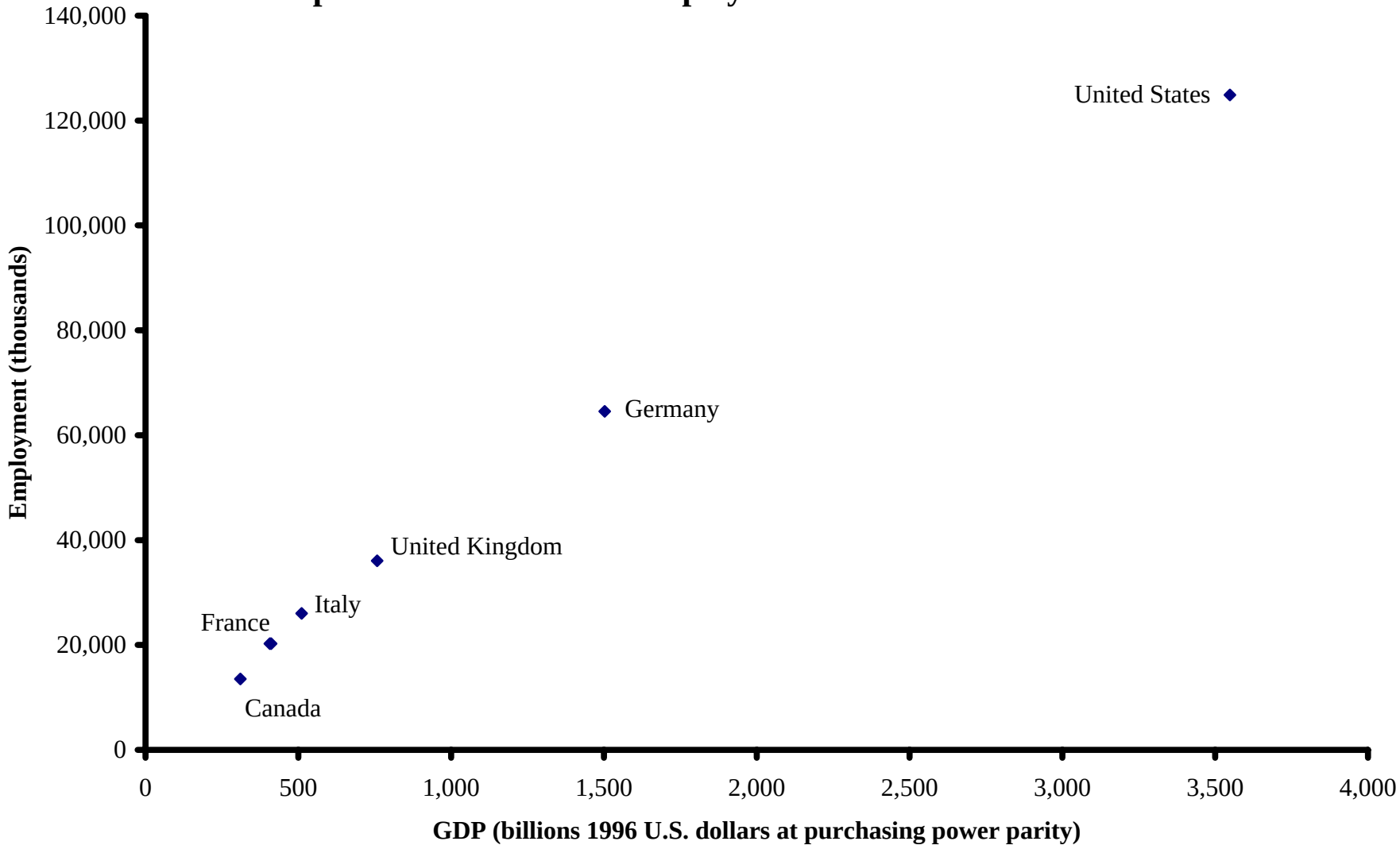
[1] Edward Tufte, *The Visual Display of Quantitative Information*, 2nd edition. Chesire, CT: Graphics Press, 2001, p. 47.

Figure G.1
G-7 Employment in 1995



Source: International Monetary Fund, *International Financial Statistics.*

Figure G.2
A Scatterplot of Real GDP and Employment in the G-7 countries in 1995



Source: International Monetary Fund, *International Financial Statistics.*

employment against real GDP for the G-7 countries in 1995 (i.e., one line from Table G.2 against the corresponding line from Table G.1). The scatterplot not only allows us to see the different levels of real GDP and employment in each country, it also shows clearly that there is a strong upward association: the higher real GDP, the higher the level of employment. Which is cause and which effect is not clear (see section G.12.3).

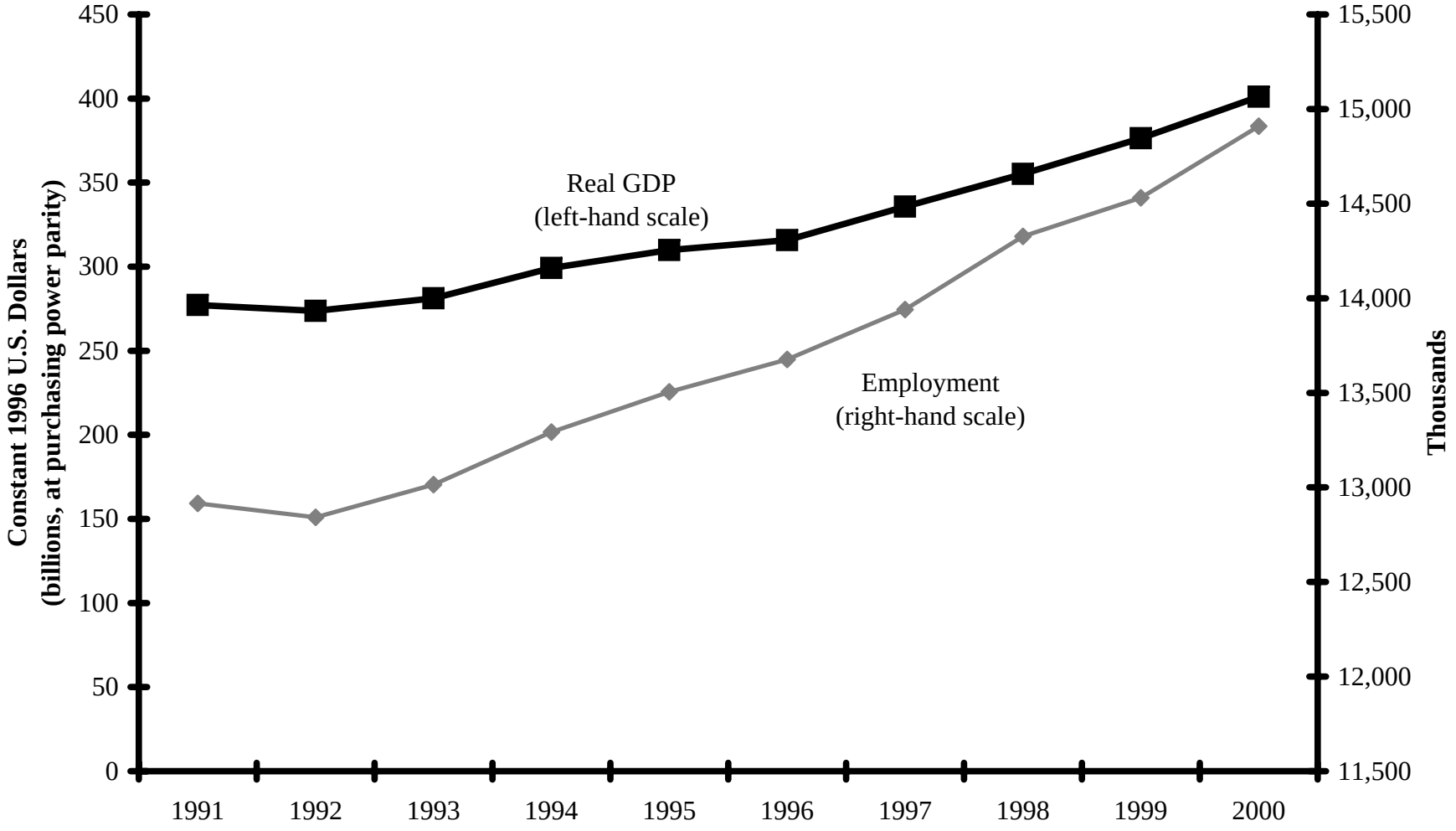
G.2.2 TIME-SERIES GRAPHS

Time-series Plots

The most common way of displaying a time series is as plot with time on the horizontal axis and the time series on the vertical. If time is considered to be a variable then the **time-series plot** is just a scatterplot of a variable against time. A single variable may be plotted or several variables either against a common or separate vertical scales. Figure G.3 plots real GDP and employment for Canada. Since units of measurement are incommensurable, real GDP is plotted against the left-hand vertical scale and employment against the right-hand scale. When two series are plotted against separate scales, their relative changes may remain economically meaningful, but the vertical location (including whether and where they cross) does not: the series can be moved up or down by changing the vertical scale.

It is typical but not required to connect the points on a time-series plot. The squares and diamonds in Figure G.3 mark the only observed data. An information read off of lines that connect these markers is an **interpolation** or conjecture. Frequently, the markers themselves are omitted.

Figure G.3
A Time Series Plot: Canadian Real GDP and Employment



Source: International Monetary Fund, *Internatinal Financial Statistics.*

Time-series Scatterplots

Two time series may also be displayed in a scatterplot. Figure G.4 presents the same data as Figure G.3, but this time employment is plotted against real GDP rather than against time. Time is shown in the labels. The scatterplot again shows clearly that both real GDP and employment are growing over time; yet, rather than emphasizes the time dimension, it highlights the fact that higher employment is associated with higher real GDP (that is, what was true for the G-7 in a single year (see Figure G.2) is also true for Canada over the last decade of the 20th century. Because of the focus on the relationships between the variables, time-series scatterplots frequently omit time labels and lines connecting observed points.

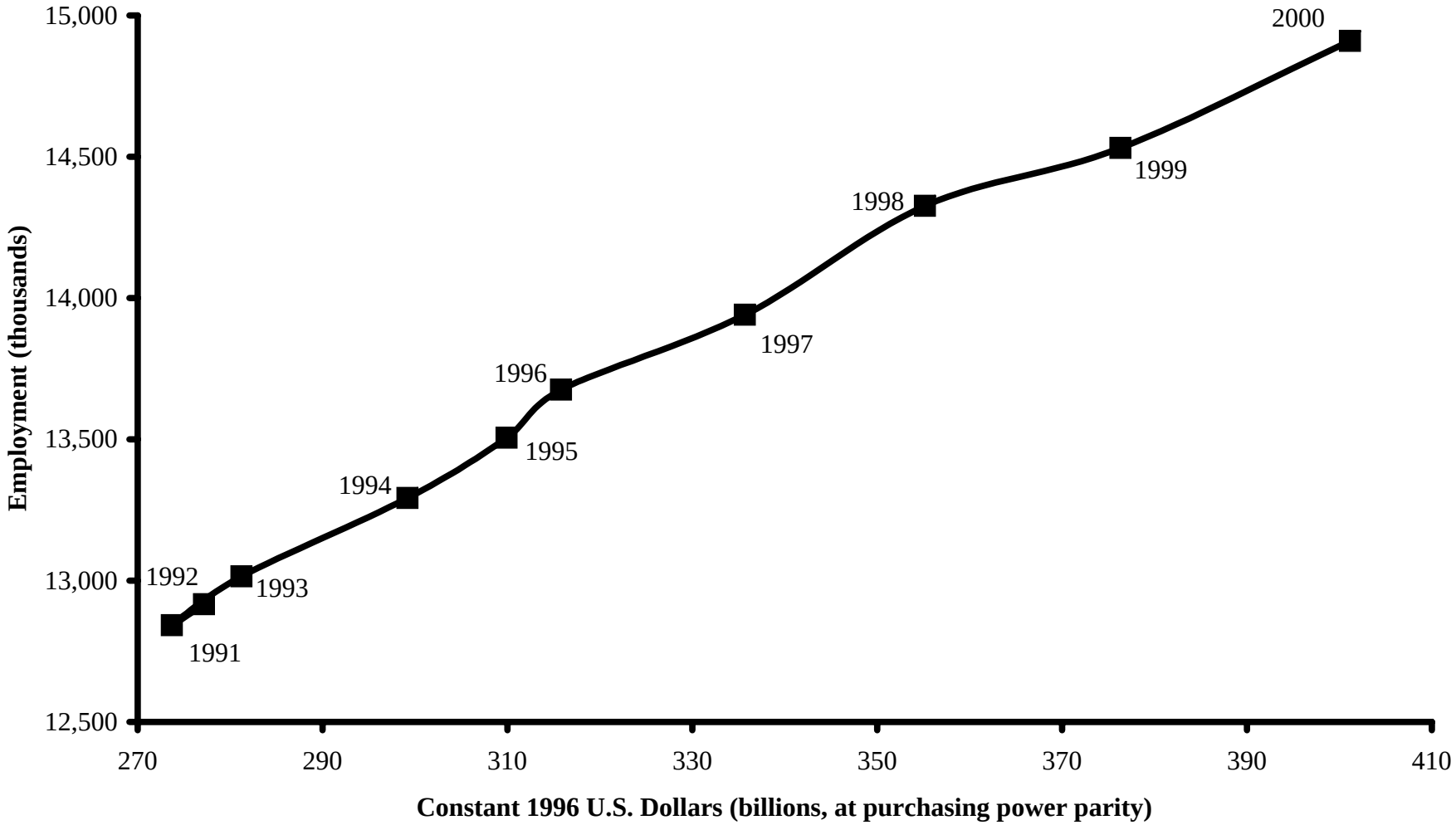
G.2.3 Guide to Good Graphics

Graphs serve two main functions: to communicate and to support empirical analysis. To do these well graphs should be (1) informationally rich; (2) clear and self-contained; and (3) aesthetically pleasing.

A Good Graph is Informationally Rich

Each graph is created for a particular purpose; it has a point of view. The point of view determines in large measure the type of graph created (e.g., scatterplot, pie chart, time-series plot). A good graph includes as much information relevant to that point of view as possible. A good graph not only presents the facts in such a way that it supports a

Figure G.4
Time-Series Scatterplot: Canadian Real GDP and Employment



Source: International Monetary Fund, *International Financial Statistics.*

conclusion that we wish to communicate, it also presents a full enough set of facts that readers can explore the data relationships on their own. In many cases, we create a graph as a tool for data exploration – not because we know what it will show, but precisely because we do not yet know. The more information that is displayed, and the more effectively it is displayed, the more likely we are to see a relationship among data that might have eluded us otherwise.

Of course, Graphs cannot, and should not, contain everything. The point of view guides our choice not only of what to include but of what to omit. A good graph is not dishonest: it does not portray false data; it avoids misleading psychological effects. For example, if the variations of a series in a range between 90 and 100 are of primary interest, a graph that used a scale of 0 to 100 would *appear* to display a flatter line with less variation – a psychologically misleading presentation even if every data point were placed accurately.

A Good Graph is Clear and Self-contained

The nature of the information displayed in a graph should be clear to anyone looking at it without essential reference to an accompanying text.

- Every graph should have a descriptive title.
- Its axes should be clearly labeled with the units of measurement and, where appropriate (as in a scatterplot) the name of the variable to which it refers. (An exception may be made for the horizontal axis in a time-series plot. If the fact that the scale refers to dates is obvious, then labeling the axis with “date” or “years” or other units is usually not necessary.)

- When a graph has multiple lines, they should each be made clear and distinct through the use of different weights, styles (solid, dashed, etc.), and markers. (Before using colors, consider whether the graph is likely to be printed or copied in black-and-white by you or a subsequent user.) Each line should be clearly labeled.
- Graphs should indicate the source of the data.
- Explanatory notes should be kept to minimum, but used where needed to describe the contents of the graph fully and accurately.

A Good Graph is Aesthetically Pleasing

A good graph should not add to the ugliness of the world; there is too much already. There are two aspects: utility and taste.

With respect to utility:

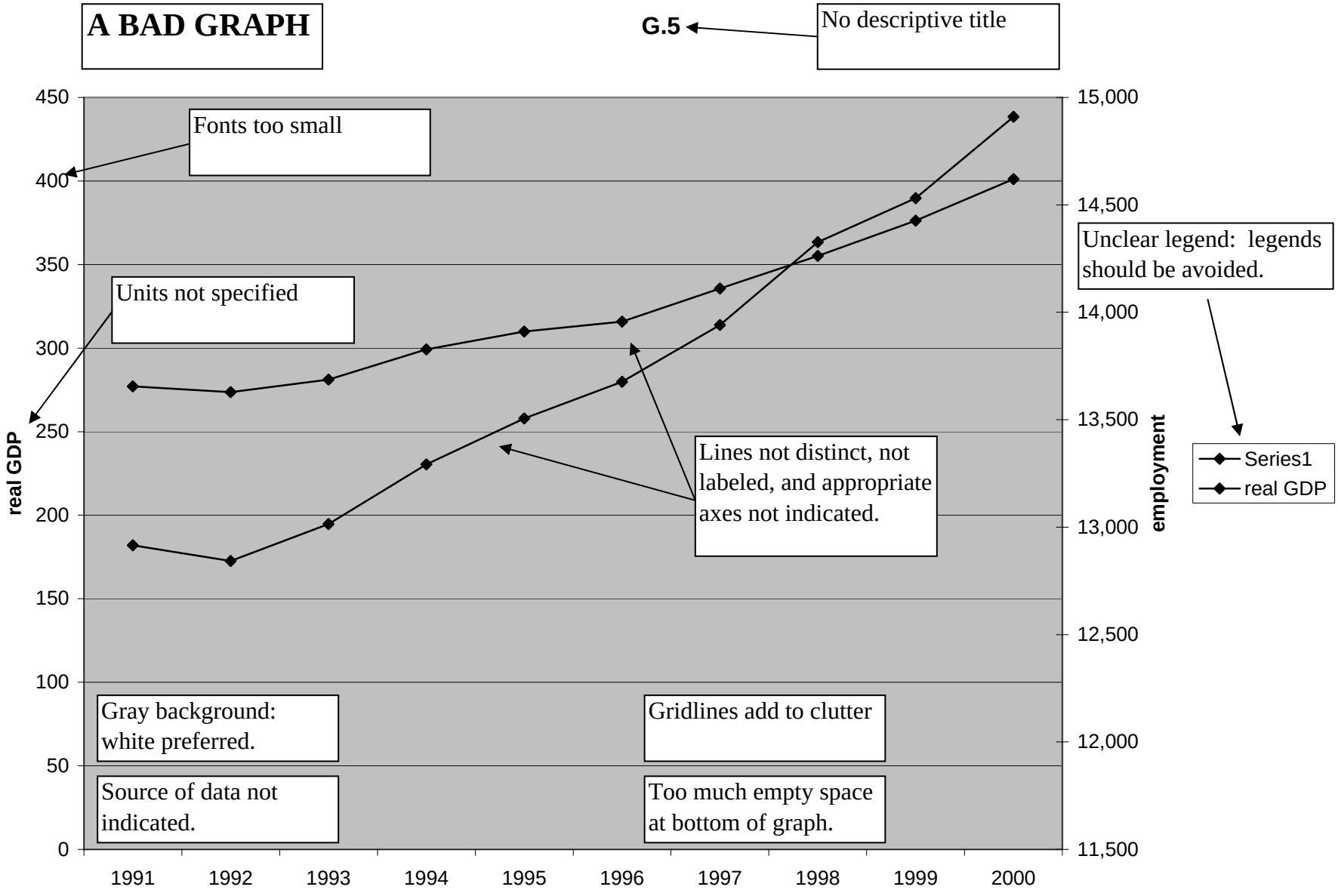
- All lines (not only data plots, but also axes, arrows, and so forth) should be dark and clear;
- Fonts should be large enough to be easily read;
- Legends identifying the data plots outside the graph (an Excel default) should be avoided in favor of labels near each line inside the graph: legends force the reader to continually go back and forth between data plot and legend to assign the correct variable name;

Taste is in the eye of the beholder – though generally some people are held to have better taste than others. My personal taste suggests:

- Use white backgrounds (not the gray backgrounds that are the Excel default);
- Avoid data markers (squares, diamonds, etc.) in time-series plots – especially when the data is dense;
- Avoid gridlines unless essential to an accurate reading of the graph or on logarithmic graphs, where they help to reinforce the reader's appreciation of the nonlinear scale;
- Do not put graphs in a box (i.e., leave the top and right-side (where there is no scale) open;
- Select scales to use available space to show, except in cases where it would be misleading from a reasonable point of view, the maximum variation of a data plot;
- Use proportional, serif fonts (e.g., Times New Roman or Book Antiqua) rather than non-proportional fonts (e.g., Courier) or sans serif fonts (e.g., Arial, the Excel default), as they are easier on the eye.

The Golden Rule of Graphics

No list of rules is complete. What constitutes a good graph depends, in part, on the circumstances and, in part, on the personal taste. Some graphs are clearly better than others. Compare Figure G.5 (which is, more or less, the way an Excel default graph) to the same data displayed in Figure G.3. Which is the better graph? An excellent source of guidance and inspiration on how to make graphs effective and pleasing is found in a series of lovely books by the political scientist Edward Tufte (see Suggested Readings at the end of the chapter for the references).


A BAD GRAPH
G.5
No descriptive title
Fonts too small
Units not specified
Lines not distinct, not labeled, and appropriate axes not indicated.
Unclear legend: legends should be avoided.
Gray background: white preferred.
Gridlines add to clutter
Source of data not indicated.
Too much empty space at bottom of graph.
real GDP
employment
Series1
real GDP
450
400
350
300
250
200
150
100
50
0
15,000
14,500
14,000
13,500
13,000
12,500
12,000
11,500
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000

The Golden Rule of Graphics is, *Be Respectful of the Reader*: create only graphs that you would yourself find useful, effective, and pleasing to the eye. The object is not to follow an arbitrary set of rules, but to communicate and support empirical analysis effectively. Any rule can be broken if it helps it to better serve its purpose.

G.3 A Guide to Good Tables

Many of the rules and considerations discussed in the last section that govern good graphics also apply to good tables. There is no need to repeat them all here. Some rules are particular to tables:

- Make the relationships between headings and subheadings clear.
- Align numbers in columns consistently (on the decimal point if there is one) clearly below the appropriate heading.
- Avoid vertical rules or grids.
- Use horizontal rules sparingly to indicate the beginning and end of tables, and to group headings in a logical manner.
- Units should be indicated in the headings; avoid attaching units directly to entries in the table itself. (For example, if the units are percentages, put that in the heading, and enter, say, 6.8 rather than 6.8%, as data in the table.)

Tables G.1 and G.2, and other tables in this book, provide reasonable models to follow.

G.4 Descriptive Statistics

The main problem with data is that there is too much of it: we cannot see the forest for the trees. Statistics are tools for economizing on information, for describing a mass of data more simply, so that we can see the essential relationships among them.

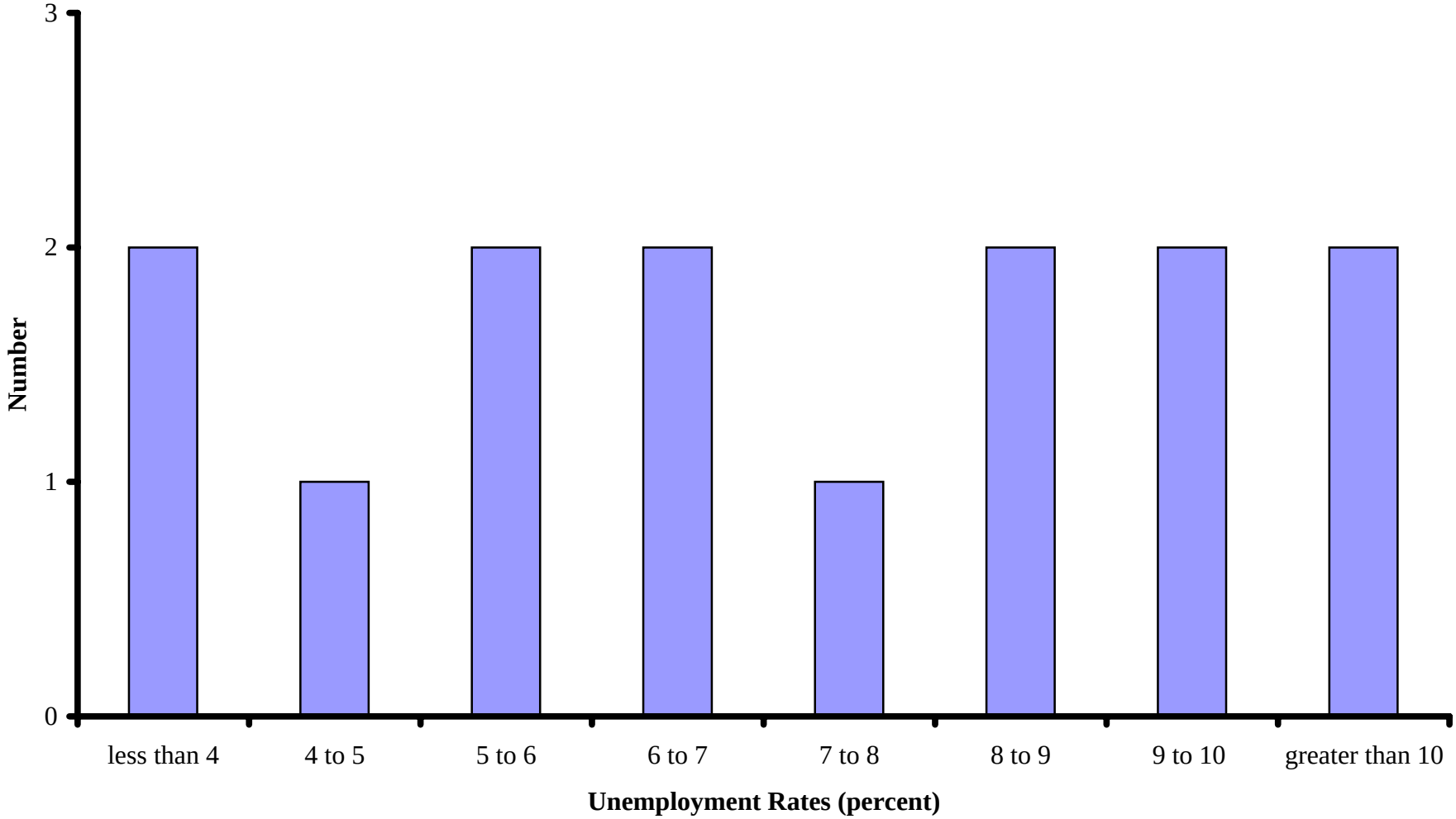
G.4.1 Histograms and Frequency Distributions

Some data are *discrete*; they must take integer values. If we count the number of MMs of each color in a bag, there are only 6 categories (blue, brown, green, orange, red, yellow). We could make a bar chart with each color corresponding to one column and the height of the column to the number of MMs in the bag. Such a chart is called a **histogram**; it displays graphically the **frequency distribution** (how many of each type) of MMs in the bag.

Most economic data is not discrete. For example, look at the Chapter 9, Table 9.2, which reports the unemployment rates for selected countries in 2003. The unemployment rates are *continuous*; it can take any real value between 0 and 100 percent. In fact, every value in Table 9.2 is unique. A histogram running from the minimum value (3.6) to the maximum (19.2) by units of 0.1 would count one entry for 15 values corresponding to each of the 15 countries for which values are reported. Such a histogram would offer no economy of information.

To do better with continuous data, we can divide the possible values into intervals. For example, Figure G.6.A shows a histogram for the data in Table 9.2 that counts the data falling into eight ranges or **bins**: less than 4, between 4 and 5, between 5 and 6, etc.,

Figure G.6.A
A Histogram: Unemployment Rates (selected countries)



Soucrce: Chapter 9, Table 9.2.

up to greater than 11. With more than half as many categories as data points, this histogram is not particularly informative: no bin receives more than two entries. Figure G.6.B shows that we gain economy of description if we reduce the number of bins to five wider bins: less than 4, 4 to 6, 6 to 8, 8 to 10, greater than 10. Fewer, broader bins means that we lose some of the detail in the original data, but the overall pattern is now clearer: most of the unemployment rates fall in the 4 to 10 range with a bias towards the higher side; and very low and very high rates are rarer. In general, histograms or other descriptions of frequency distribution involve a tradeoff between the fine detail of the data and the overall pattern (the big picture).

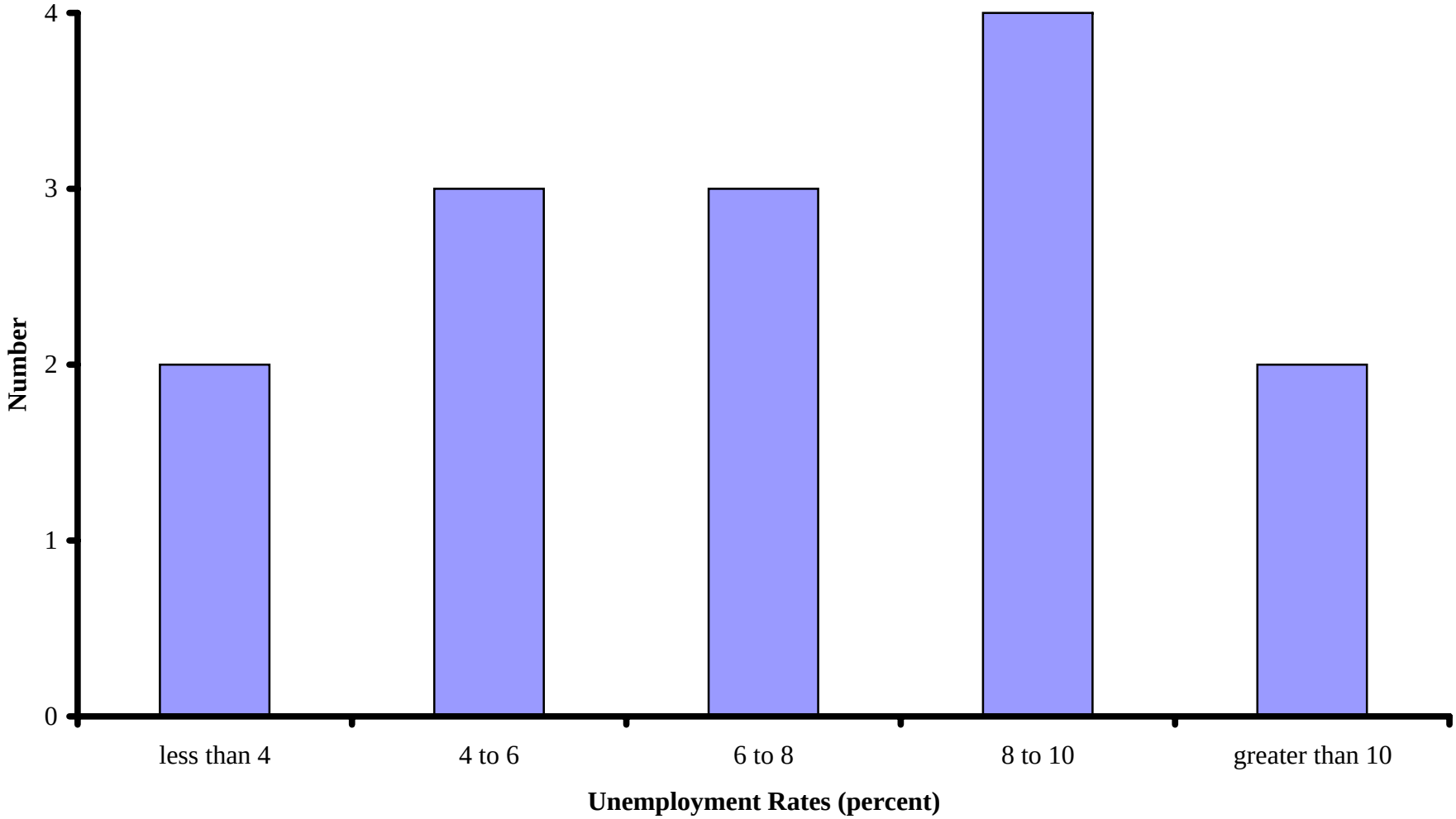
G.4.2 Measures of Central Tendency

Even more economy can be gained if we can summarize the data in a few numbers that characterize its frequency distribution. The most important ones fall into two categories: measures of central tendency and measures of variation (or dispersion).

The (Arithmetic) Mean

The word *average* is sometimes used as a synonym for any measure of central tendency (that is, for a measure of the middle of a frequency distribution). More commonly, it refers to a specific measure: the **arithmetic mean**, usually abbreviated simply as **mean**. The mean starts with a set of data and answers the following question: *if every data point were to be replaced by ones that all took the same value, what value would they have to take to be equal to the same total as the original data?*

Figure G.6.B
A More Informative Histogram: Unemployment Rates (selected countries)



Soucrce: Chapter 9, Table 9.2.

Example G.6. What is the mean of a set of data: 4, 2, 7, 1?
Answer. The total is 14. If every data point took the value 3.5 (= 14/4), then we would have the same total.

The general formula for the mean of a variable X_i, where i indicates one of N data points is:

$$\text{(G.2)} \qquad \overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}.$$

The bar over the variable (without a subscript) indicates the mean.

Example G.7. What is the mean of the unemployment data in Table 9.2?
Answer. $\overline{U} = \frac{\sum_{i=1}^{15} U_i}{15} = \frac{111.6}{15} = 7.4$ percent.

[*Excel* hint: the command for the mean is: =AVERAGE(number1, number2, . . .)]

The Weighted Mean

Sometimes different data points have different degrees of importance. What if we wanted to know the mean unemployment rate for the United States and Canada taken together. Applying equation (G.2) to the data in Table 9.2 yields: 6.8 percent (= (7.6 +6.0)/2). But if we want to use the average unemployment rate as a measure of the chance that a randomly chosen person is from the U.S. and Canada is unemployed, we should take account of the fact that the United States has about ten times the population of Canada (294.0 vs. 31.5 million), so that many more people face the lower

unemployment rate than the higher. We can adjust the formula (G.2) to form a **weighted mean** in which population provides the weights (w_i) :

$$\text{(G.4)} \qquad \overline{X} = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i}^{N} w_i}.$$

Example G.8. What is the weighted mean of U.S. and Canadian unemployment rates weighted by the countries' populations?
Answer. Using formula (G.4), the weighted mean unemployment for the United States and Canada is: $\overline{X} = \frac{31.5 \times 7.6 + 294.0 \times 6.0}{31.5 + 294.0} = 6.1$. This is much closer to the U.S. value than to the Canadian value and lower than the unweighted mean, because the U.S. value receives a far larger weight.

The Median

The arithmetic mean can be thought of as choosing the point that puts an equal weight of the frequency distribution on each side. Visually, imagine the histogram in Figure G.6.B flipped over so that the bars hung down from the horizontal axis. The question the mean answers is: where along the horizontal axis should we attach a string if the histogram is to hang perfectly level.

The mean can be misleading. For example, if we want to know the center of the distribution of income in the State of Washington, then Bill Gates's billions of dollars get a very heavy bar in the histogram with bins representing different income levels on the horizontal axis. We would have to move the string far to the right to make it hang level.

But at that point, the vast majority of Washingtonians are situated far to the left of the mean. The mean does not give us a good idea of what is typical. The **median** asks a different question: *what value divides the distribution into an equal number above and below that value?* In answering this question, despite his billions, Bill Gates counts as one person above the median with an equal weight to a person who is just one above the median.

Example G.9. What is the median of the unemployment data in Table 9.2 and Figure G.6?
Answer. 6.1 percent. Seven countries have unemployment rates higher than this level and seven lower.

Notice that in this last example, the median is actually one of the points in the data set: it is the value for Australia. Whenever there are an odd number of data points (here 15), the median will be an element of the data set itself. By convention, if there are an even number of data points, the median is reported as the arithmetic mean of the highest value in the lower half of the sample and the lowest value in the higher half of the sample.

Example G.10. Using the same data set as in Example G.9 but supplemented with an additional observation for Elbonia with an unemployment rate of 53 percent, what is the median?
Answer. 6.85 percent. Australia (6.1 percent) now has the highest unemployment rate of the lower eight countries and Canada (7.6 percent), the lowest of the highest eight countries: $(6.1 + 7.6)/2 = 6.85$.

When a distribution is *symmetrical* – that is, the left-hand and right-hand sides of the histogram are mirror images of each other, the median and the arithmetic mean coincide. When a distribution is **positively skewed** (that is, its weight is shifted to the right as in Figure G.6.B), *the mean is greater than the median*. When it is **negatively**

skewed (that is, more weight is in the left-hand side of the histogram), *the mean is less than the median*.

Example G.11. Consider 14 people: 1 has an income of $1,000; 5 of $2,000; 4 of $3,000; 2 of $4,000, 1 of $5,000. What are the mean and the median? Is the distribution of income skewed? If so, which way?
Answer. *Mean* = [1(1,000) + 5(2,000) + 4(3,000) + 2(4,000) + 2(5,000)]/14 = $2,929. The highest value of the lowest seven is $3,000 and the lowest value of the highest seven is also $3,000, so that is the *median* is the mean of $3,000 and $3,000. Since the mean is less than the median, the data are *negatively skewed.*

[*Excel* hint: the command for the median is: =MEDIAN(number1, number2, . . .)]

The Geometric Mean

Just as the arithmetic mean is a natural measure for data that are related as a sum, the **geometric mean** is a natural measure for data that are related as a product. The geometric mean starts with a set of data and answers the following question: *if every data point were to be replaced by ones that all took the same value, what value would they have to take to be equal to the same product as the original data?*

The general formula for the geometric mean is:

(G.5) $$\overline{X} = \sqrt[N]{\prod_{i=1}^{N} X_i} = \left(\prod_{i=1}^{N} X_i \right)^{1/N},$$

where $\prod$ is the product sign, which says. "multiply each term together," just as $\sum$ says, "add each term together." We use the bar over the variable (without subscript)

to indicate the geometric mean, relying on context to keep us from confusing it with the arithmetic mean.

Suppose that GDP in the country of Elbonia has grown over three successive years at 16 percent, 5 percent, and 9 percent. A growth rate of 16 percent means that GDP is 1.16 *times* the level of GDP the year before; 5 percent means 1.05 *times*; 9 percent, 1.09 *times*. The total after three years is the *product* of each of these factors *times* the level of GDP the year before the first year of growth. What is the average rate of growth over these three years? Applying the formula: the geometric mean is $(1.16 \times 1.05 \times 1.09)^{1/3} = 1.10$, a growth rate of 10 percent.

Example G.9. You are given a set of data that appear to be growing steadily, but one observation is missing: 1.1, 2, X_3, 8.1, 16.3, where X_3 is the missing observation. What is a good guess for the value of X_3?
Answer. This is a problem of **interpolation**. One strategy is to take the average of the adjacent values: 2 and 3.9. Because the data are growing fairly steadily, and growth is a multiplicative process, the geometric average is preferred. Applying formula (G.5), $\overline{X} = \sqrt[2]{2 \times 8.1} = 4.0$. Notice that the arithmetic mean is 5.0, which is, in context, a very different value.

[*Excel* hint: the command for the median is =GEOMEAN(number1, number2, . . .)]

As with the arithmetic mean, some individual data points may be more important than others, suggesting a **weighted geometric mean**. With an arithmetic mean, the weights enter multiplicatively; with a geometric mean, they enter as powers. The formula is:

(G.6) $$\overline{X} = \sqrt[(\sum w_i)]{\prod_{i=1}^{N} X_i^{w_i}} = \left(\prod_{i=1}^{N} X_i^{w_i} \right)^{(1/\sum w_i)}.$$

Example G.9. A bank offers a special account that earns 7 percent for the first five years and 5 percent thereafter. Interest accrues in the account. If you hold the account for 20 years, what is your average yield?
Answer. Interest acts like a growth rate: 7 percent interest, for instance, implies that each year the account becomes larger by a growth factor of 1.07 each year. Applying formula (G.6) gives us:

$$\textit{average growth factor} = \sqrt[(5+15)]{1.07^5 \times 1.05^{15}} = \sqrt[20]{2.92} = 1.055,$$

which corresponds to an interest rate of 5.5 percent.

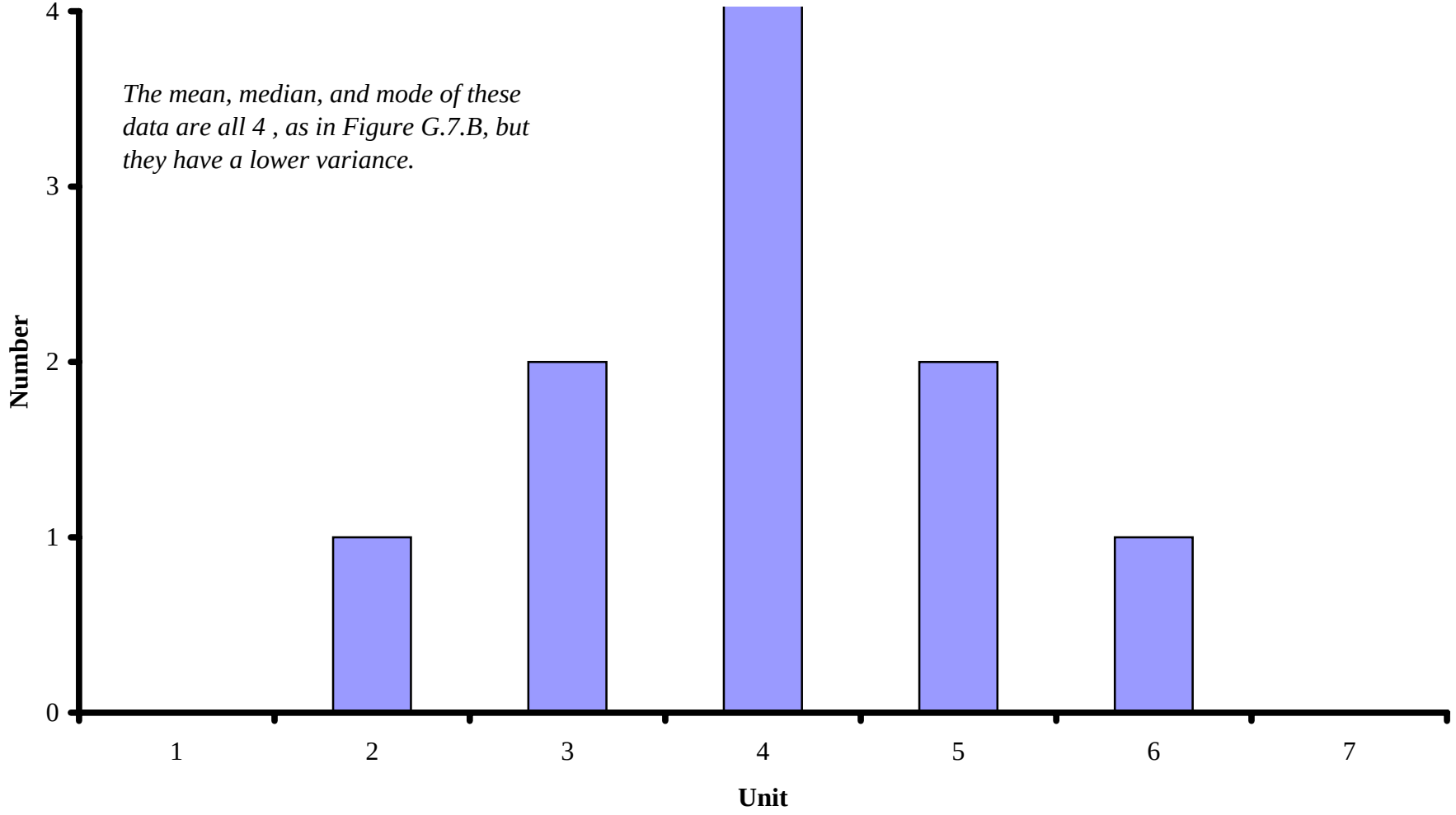
G.4.2 MEASURES OF VARIATION

Variance

The mean, median, and mode are all measures of the center of a frequency distribution. We might also want to know how spread out a distribution is. Figure G.7.A shows a histogram of the dataset A = {2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6} and Figure G.7.B of the dataset B = {1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6 ,6, 7}. Both datasets have the same mean, median, and mode – all 4. Yet, the set B is clearly more spread out than set A.

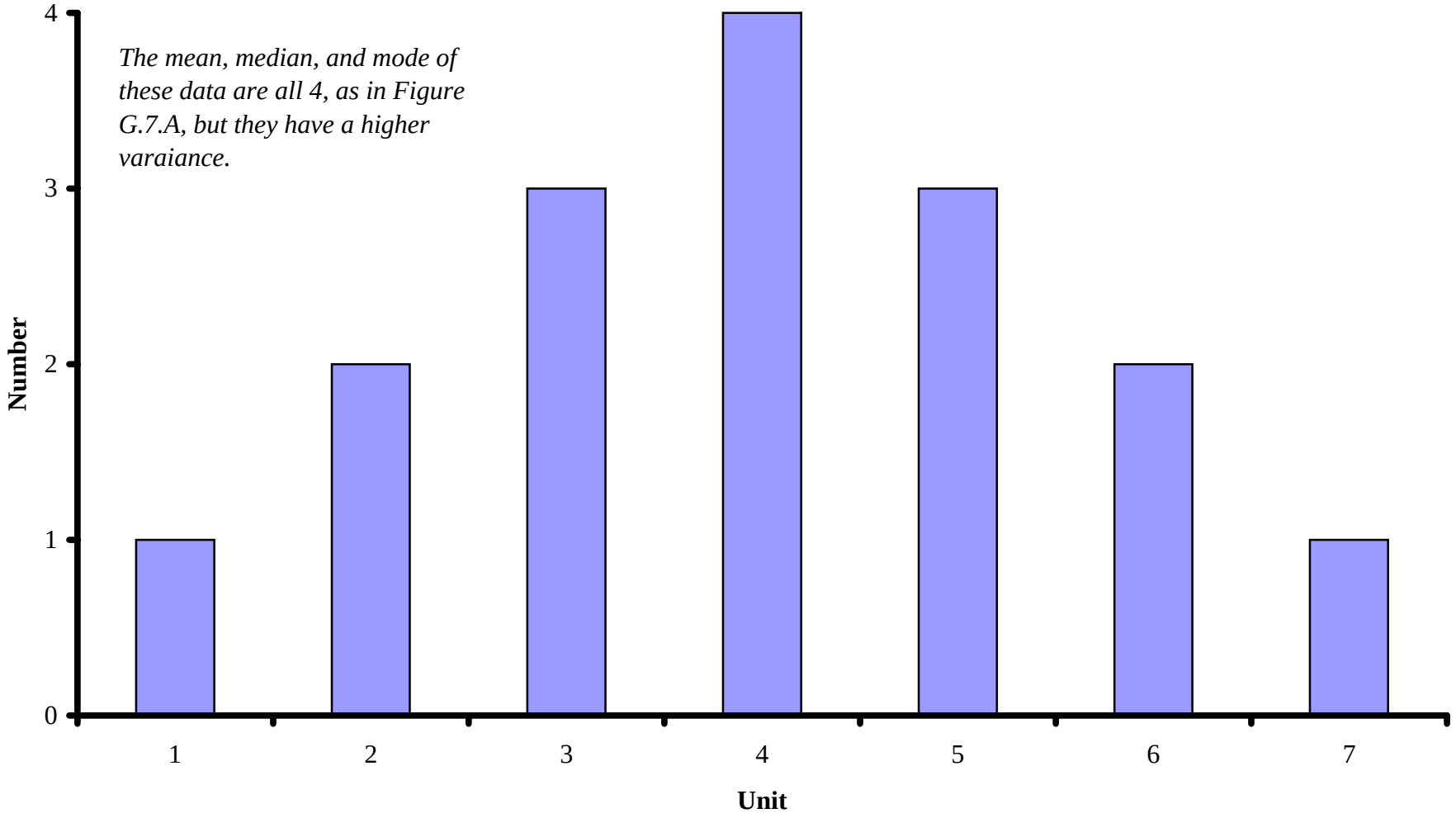
One way of measuring the spread would be to take the mean as the center of the distribution and to measure the deviations from the mean. The mean of these deviations would measure the variation. So, for example, the first deviation in set A would be –2 (= 2 – 4); while the last deviation would be +2 (= 6 – 4). The problem with this idea is that when we add up the deviations to take their mean, the positive and negative ones would cancel out. Since both datasets are symmetrical, the cancellation would be complete, and the measure of variation would be zero for both – not at all capturing the obvious fact that dataset B is more spread out than A. Instead, we could take the absolute value of the

Figure G.7.A
A Histogram with Relatively Low Dispersion



Soucrce: Chapter 9, Table 9.2.

Figure G.7.B
A Histogram of a Data with Greater Dispersion



Soucrce: Chapter 9, Table 9.2.

deviations and add them up. That suggestions works, and is sometimes used as a measure of variation: *mean absolute variation.*

More commonly, variation is measured as the **variance**, calculated as:

$$\text{var}(X) = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1}. \tag{G.7}$$

The term in parentheses on the right-hand side of formula (G.7) defines each deviation from the mean. Because all squared numbers are positive, every deviation contributes positively to the sum. The total is averaged over $N - 1$, rather than over N (as might seem more natural), because the mean is taken to be a reference point, and the mean value itself does not generate a deviation from the mean. Therefore, if we have, say, 10 datapoints, they generate only nine true deviations from the mean. Notice that this false deviation will always be calculated in the numerator of formula (G.7), but that it will always take the value of zero. Dividing by $N - 1$, rather than N, removes its effect is removed from the calculation. (This is known as a *degrees-of-freedom* correction.)

Example G.10. What is the variance of {1, 2, 3, 4, 5}?
Answer. The mean is 3; $N = 5$. The deviations from the mean $(X_i - \overline{X})$ are {–2, –1, 0, +1, +2}. Applying formula (G.7) yields var(X) = $[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2]/4 = 10/4 = 2.5$.

Example G.11. Which is more spread out, dataset A or B in the text above and Figure G.7.A and B?
Answer. Applying formula (G.7), var(dataset A) = 1.200 and var(dataset B) = 2.667, confirming our visual impression that B is more spread out than A.

[*Excel* hint: the command for the variance is: =VAR(number1, number2, . . .)]

Standard Deviation

A disadvantage of the variance is that it has difficult to interpret units. For example, suppose that we calculated the variance of real GDP in Table 9.2 (Figure G.1). The units of the original data are *billions 1996 U.S. dollars at purchasing power parity*. The units of the variance are (*billions 1996 U.S. dollars at purchasing power parity*)^2. It is hard to know what that means. It is, therefore, often convenient to transform the variance into the **standard deviation**, defined as

$$\text{stdev}(X) = \pm\sqrt{\text{var}(X)}\,. \tag{G.8}$$

The standard deviation has the same units as the original data.

Example G.12. What are the standard deviations of datasets A and B (see Example G.11).

Answer. Applying formula (G.8), stdev(dataset A) = $\sqrt{1.200} = 1.095$ and stdev(dataset B) = $\sqrt{2.667} = 1.633$, again confirming our visual impression that B is more spread out than A.

[*Excel* hint: the command for the standard deviation is:

=STDEV(number1, number2, . . .)]

Coefficient of Variation

Sometimes the variance or the standard deviation can be quite misleading about the degree of variability. Imagine a population of mice in which variations of an ounce around the mean weight of a mouse were common. Compare that to a population of elephants in which variations of 5 pounds (80 ounces) were common. The variance and standard deviation of the weight of the elephants is much larger than that of the mice. Yet, there is a clear sense in which *relative to their normal weight*, the mice show more variation. To capture this sense, we can calculate the **coefficient of variation**, which expresses the standard deviation as a share of the mean:

$$\text{cv}(X) = \left|\frac{\text{stdev}(X)}{\overline{X}}\right|. \tag{G.9}$$

We take the absolute value of the share, since means can be positive or negative, and the coefficient of variation is always reported as a positive number.

Example G.13. Relative to their means, which is more variable, real GDP among the G-7 countries or the non-G-7 countries in Table 9.2 (Figure G.1)?
Answer. This table gives the coefficients of variation and the elements from which it is calculated using formula (G.9):

	G-7	**non-G7**
standard deviation	1.878	5.417
mean	7.314	7.550
coefficient of variation	0.257	0.717

With a coefficient of variation of 71.7 percent, real GDP in the non-G-7 countries is nearly three times as variable as in the G-7 countries (25.7 percent).

G.5 Making Inferences from Descriptive Statistics

G.5.1 HOMOGENEITY

We can always calculate various descriptive statistics, but whether they are meaningful depends on how we choose to use them. For example, imagine that plan to take a group of 6th graders on a balloon ride. There are ten children in the group, but the balloon can hold up to 15 children weighing not more than 1,600 lbs. in total. If the mean weight of the 6th graders is 100 lbs., then your group is easily accommodated. Now suppose that you want to add 5 more children from another class: will the balloon accommodate them? If you cannot weigh the children, then you might reason that, since the average weight of your group was 100 lbs., adding five more will on average only add 500 lbs., leaving 100 lbs. as a margin of safety for the balloon. Your reasoning is fine – provided that your children are typical of the children that you are going to add. If they come from a class of 6th graders, this may be a good assumption. But what if them come from a class of 12th graders, who, on average, weigh 135 lbs.? The balloon would then be overloaded ($10 \times 100 + 5 \times 135 = 1{,}675$ or 175 lbs. above the weight limit).

The point of the story is that we often calculate descriptive statistics for the purpose of *projecting* them beyond the dataset that we have observed – that is, to make inferences about unobserved data from observed data. This works only if the observed and unobserved data are, on relevant dimensions, alike or homogeneous.

We may also require **homogeneity** within a dataset as well as between datasets. If we claim that a mean or other descriptive statistic is *typical* of the data, it should be typical of any subset of the data. For example, consider a group of 1st graders on a field trip with their parents – one parent per child. We can calculate the mean weight of the group. But if parents have a mean around 150 lbs. and 1st graders have a mean of around 50 lbs., then the mean for the whole dataset will be around 100 lbs.: not typical of either the parents or the children, and for most purposes of little use.

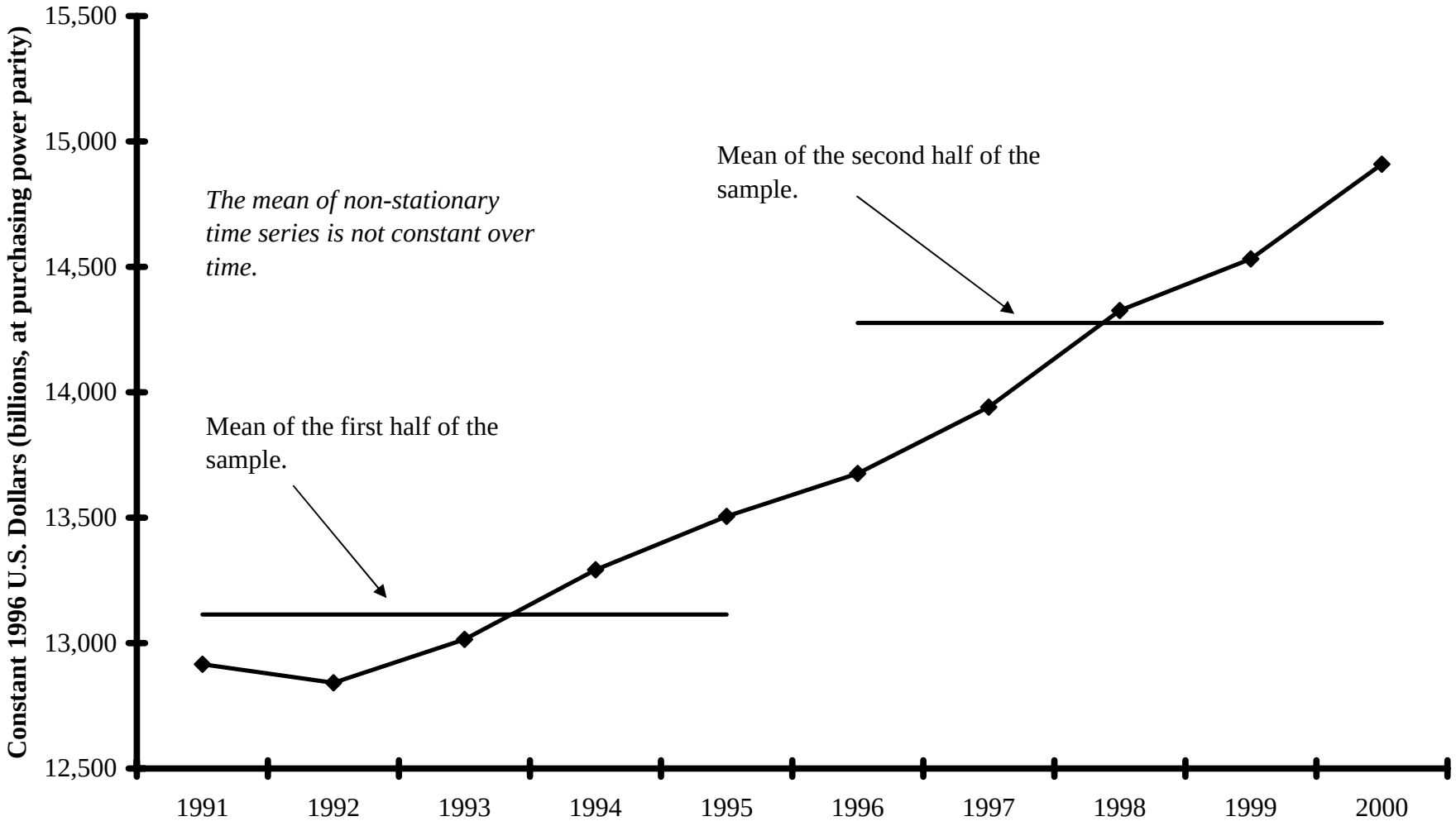
Data are *homogeneous* when *every reasonably sized subset display similar descriptive statistics.* Statisticians have developed a number of formal tests of homogeneity, which are beyond the scope of this book. But frequently, commonsense reflection on a problem and examination of some subsets of the data will give us a good idea of whether the data are homogeneous enough for our purposes.

Example G.14. Are unemployment rates among the G-7 countries in Table 9.2 (Figure G.1) homogeneous?
Answer. There is no unequivocal answer to this question; it depends partly on our purposes. But notice that we can readily divide the G-7 into two subsets: Japan, the United Kingdom, and the United States with low unemployment rates (mean = 5.4 percent); and Canada, France, Germany, Italy with high unemployment rates (mean = 8.7). For many purposes we would reject the homogeneity of G-7 unemployment rates.

G.5.2 STATIONARITY

Figure G.8 shows the same Canadian employment data as Figure G.3. If we divide the data into two halves, the means for each subset (shown as horizontal lines in the figure) are clearly different: the dataset is clearly not homogeneous even though it refers to a single economic unit, Canada, rather than to the different economic units in the last

Figure G.8
A Non-stationary Time Series: Canadian Employment



Source: International Monetary Fund, *International Financial Statistics.*

section. While the order of observations in a cross section is arbitrary, the order in a time series follows the natural succession of dates. Canadian employment, like many economic time series, trends upward: later segments typically have higher means than earlier segments. The Canadian data illustrates a particular type of non-homogeneity. A time series is **stationary** when *every reasonably sized subperiod displays similar descriptive statistics.* The Canadian data are **non-stationary**. Projecting the level of Canadian employment in the later period based on its mean for the earlier period would have led to a substantial underprediction. To do better, we would have to take account of the fact that the mean was constantly increasing.

Statisticians have developed formal tests of stationarity, but these are beyond the scope of this book. Informally, we judge a time series to be stationary when it frequently crosses its sample mean. Obviously, a series that is trending strictly upwards or downwards, such as Canadian employment, will cross its sample mean only once. But a series need not trend strictly in one direction to be nonstationary.

For example, Figure 12.2 in Chapter 12 shows that the average propensity to consume in the United States cross its mean only five times in fifty years and drifts fairly far away from it at times, so, should be regarded as *non-stationary*. In contrast, Figure 12.5 in Chapter 12 shows that the average tax rate crosses its mean twice as often and generally remains quite close to its mean. It could be regarded as *stationary*.

A time series that follows a random walk (see Chapter 5, section 5.3.1), even with a drift (which is a kind of trend), is also a non-stationary time series. Unlike a trending series, a random walk does not typically move dominantly in one direction. Nevertheless,

it crosses its mean infrequently, so that descriptive statistics appropriate to stationary data should not be applied to it.

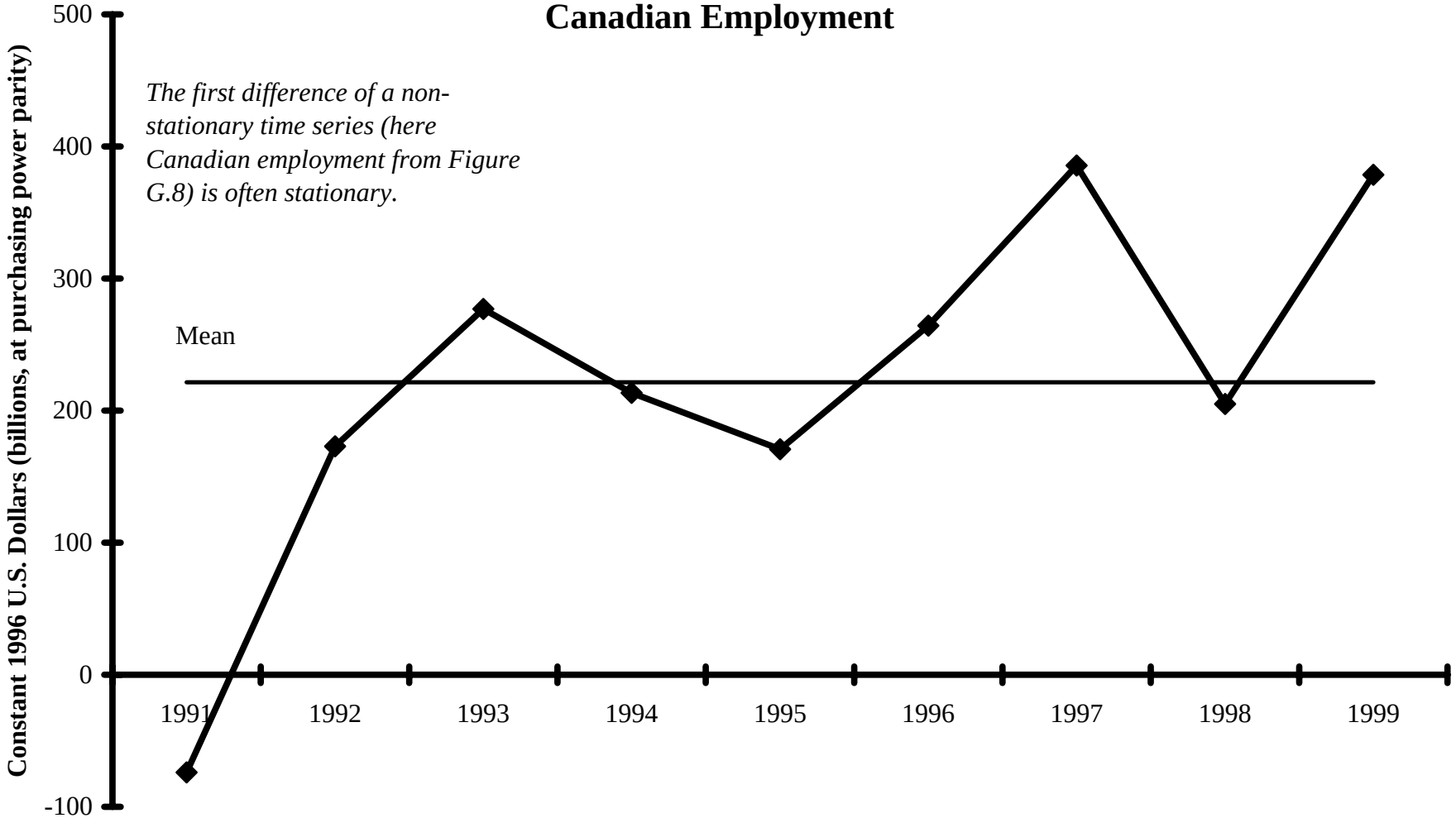
A non-stationary time series may be transformed into a stationary one in various ways. If a series has a clear trend, a detrended version will usually be stationary (see section G.12). Also, the **first-difference** of a non-stationary series ($\Delta X_t = X_t - X_{t-1}$) or the **growth rate** ($\hat{X}_t$) (see section G.10) is frequently stationary.

Example G.15. Can Canadian employment data be transformed into a stationary series?
Answer. Figure G.9 shows the first difference of Canadian employment data and its mean. It is much more like a stationary series than the level data in Figure G.8.

The last example illustrates that it may be hard to judge whether a series is stationary or not in a short sample. In general, we are more likely to perceive stationarity over long time horizons than over short ones. The data for the real rate of interest in Figure 11.9 appear to be stationary over the whole 50 year period of the plot, but would appear to be non-stationary were we to look only at the decade of the 1980s. This is not an illusion or a mistake, but the result of the data adjusting slowly to deviations from the mean – a property called *persistence*. To predict the course of real interest rates from 1986 to 1990, the downward trend of the apparently non-stationary data of the early 1980s would generally be more relevant. But over a longer horizon, say, predicting data for the early 2000s on the basis of data up to 1985, we would usually do better to project the apparently stationary mean of the longer period 1953-1985.

The importance of stationarity becomes even clearer in Section G.14 below.

Figure G.9
Tranformation of a Non-stationary to a Stationary Time Series:
Canadian Employment



Source: International Monetary Fund, *International Financial Statistics.*

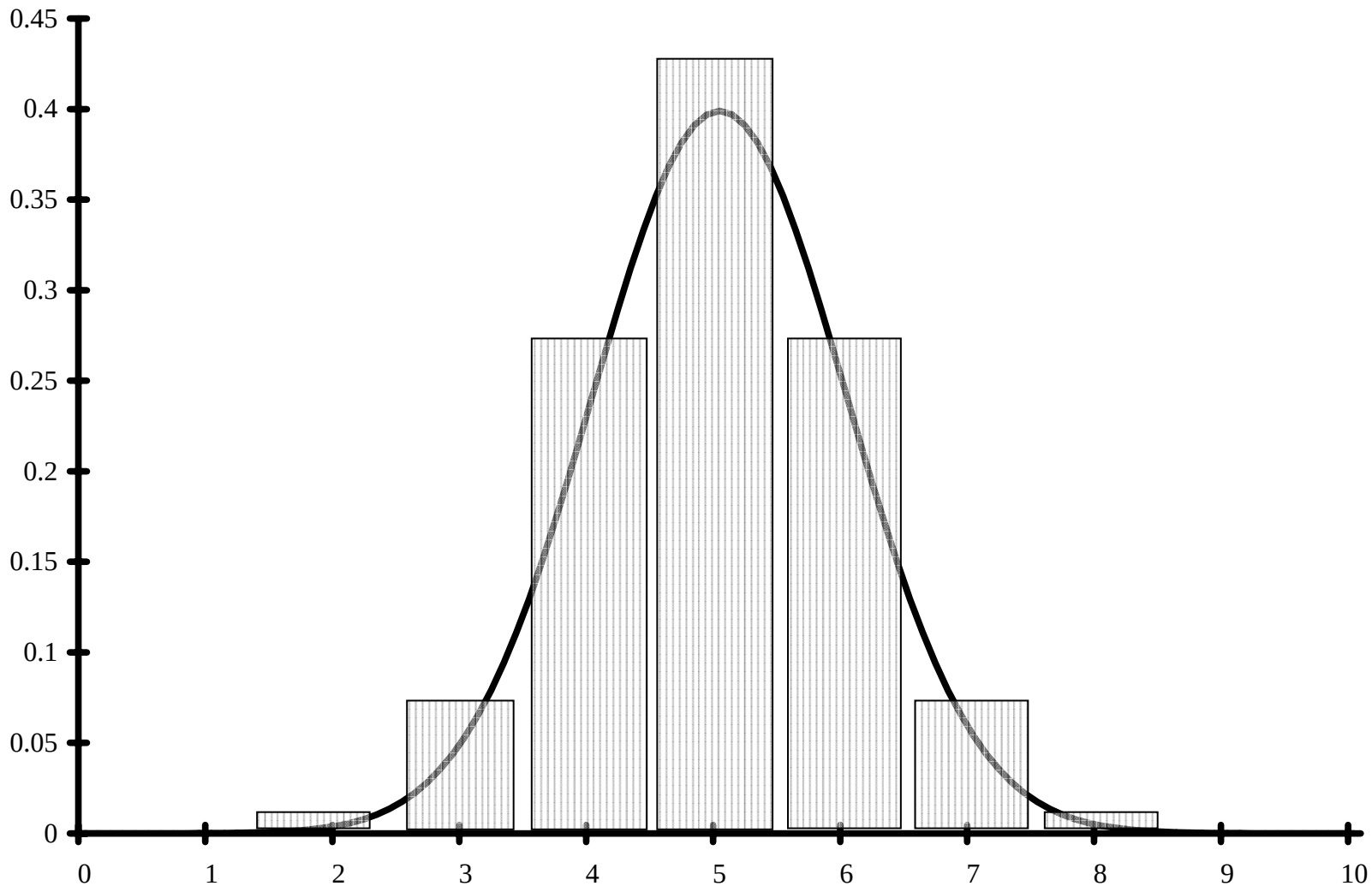
G.6 The Normal Distribution

Many data in the world have (or can be transformed to have) a **normal frequency distribution**. Figure G.10 shows a histogram of normally distributed data. The smooth bell-curve that is superimposed on the histogram shows what the curve would become if we had a large amount data and chose narrower and narrower bins until they were infinitesimally small. Where the histogram is a *discrete* distribution; the bell curve is the *continuous* version of the normal distribution. When a teacher says that students will be "graded on a curve," the usual assumption is that the grades are at least approximately normally distributed.

The normal distribution is common because of a remarkable result in statistics known as the *central limit theorem*. Roughly, the central limit theorem says that the frequency distribution of the average of a number of independent distributions tends to be normal even when the original distributions are very far from normal (i.e., their histograms do not have a bell shape). If data display random variation as the result of a large number of unmeasured, independent causes, then the net effect of those causes tends to have a normal distribution.

The normal distribution has an important property: whatever the mean or standard deviation, about 38 percent of the observations lie within ½ standard deviation of the mean; 68 percent within 1 standard deviation; and about 95 percent lie within 2 standard deviations. These benchmarks – especially the 2 standard deviation equals 95 percent benchmark – are often used as a way of expressing the measure of our uncertainty about random data.

Figure G.10
The Normal Distribution



For example, a political poll may state that 48 percent of the sample approves of the president's job performance with a "margin of error of ±3 points." How do we interpret this information?

Margin of error in the case of opinion polls almost always refers to a two-standard deviation criterion. Suppose that another poll were conducted with a similar number of people in the sample on the same population. Even if 48 percent were the true value, any particular poll would probably come up with a different number. If 100 polls were taken, we would expect 95 of them on average to find a value between 45 and 51 percent. Suppose that a second poll produced a value of 49 percent. We would typically believe that, given the uncertainty of the first poll, that this number is consistent with it and would not show that opinion had shifted. On the other hand, if a poll produced a value of 60 percent (and we did not believe that it was poorly conducted), there would be two possibilities: (a) it could be one of those (5 percent on average) cases in which the true value was 48 but the measured value fell outside the margin of error (more than two standard deviations away from the mean); (b) opinion had really changed, so that the difference was a real one and not random variation. The reason for using a 95 percent criterion, instead of say a 50 percent criterion, is to try to reduce the times when explanation (a) applies. Usually, with such a large difference people would conclude (b), the change was genuine.

G.7 Type I and Type II Error

In the last section, we gave the example of trying to judge when two political polls reflect a genuine shift in opinion. This illustrates a common problem in statistical and non-statistical reasoning. Suppose that we have a rule: *whenever the index of leading economic indicators turns down two months in a row predict a recession within six months.* Is this a good rule?

It would be perfect if every time the index turned down two months in a row, a recession followed *and* every time it did not turn down two months in a row, no recession followed. The rule can fail in two different ways: the index does not turn down, but there is a recession anyway (this is known as a *false negative*) or the index does turn down, but no recession follows (this is a *false positive*).

Statisticians classify these errors in another way. If we have a hypothesis (there is a recession), denying it when it is true (false negative) is called **type I error**. Affirming it when it is false is called **type II error.** Table G.4 shows the relationship between the truth and these two types of error.

Generally, there is a tradeoff between type I and type II error. For example, suppose that a man is on trial. Ideally, if he really committed the crime, we would like to convict; and, if he really did not, we would like to acquit. If he really did commit the crime, but we acquit him we have committed type I error. To minimize it, we could adopt the rule: always convict. Then, no criminals would every get away. But, of course, some noncriminals would be unjustly sent to prison. If he really did not commit the crime, but we convict him anyway, we have committed type II error. To minimize, it we

Table G.4
Type I and Type II Error

	The hypothesis is really	
	False	**True**
We infer that the hypothesis is		
False	Success	Type I Error
True	Type II Error	Success

could adopt the rule: always acquit. No one would go to prison unjustly. But, then again, no criminal would ever go to prison either. To balance the possibilities of type I against type II error, we must choose some intermediate rule. Which way it leans depends, of course, on which type of mistake is more costly in the particular case.

A similar problem arises with the two opinion polls in the last section. What we want to decide is whether the poll that reports 60 percent really reflects the same opinion as the poll that reports 48 percent. The rule is to deny that the polls are the same, if the second poll (60 percent) is outside the margin of error of the first poll (two standard deviations) – which it is. The two-standard-deviation rule essentially says that there is a 5 percent chance of making a type I error following this rule. We could employ an even tighter standard to further reduce the chances of type I error, but then we would increase the chance of type II error: i.e., of treating the polls as the same when they are really different.

The two-standard-deviation rule is a statisticians' rule-of-thumb for balancing type I against type II error. But without further analysis, a precise measure of type II error is not usually known.

G.8 Using Index Numbers

G.8.1 INDEX NUMBERS

An **index number**, transforms a dataset from its natural units into one that uses one of its values as a reference point. For example, suppose that we have a time series that takes the value 5 in 2003, 10 in 2004, and 12.5 in 2005. Index numbers are always stated with

a reference value set to a particular value, usually 100. Here let the **reference value** be 5 in 2003 (this is usually stated 2003 = 100). So the index series takes the values 100 in 2003, 200 in 2004 (since 10 is twice the reference value) and 250 in 2005 (since 12.5 is 2.5 times the reference value).

The general formula for an index number is:

$$XI_t = \left(\frac{X_t}{X_0} \right) 100 , \tag{G.10}$$

where the suffix *I* indicates that a natural dataset has been turned into an index number, and the subscript 0 refers to whichever data point is taken to be the reference value. The coefficient 100 establishes the value of the index at the reference date. It could be chosen to be some other number: 1 and 10 are sometimes used; other values are rare.

The reference value need not be the value of a particular observation. Usually with monthly data, the average value over the 12 months of a year, and with quarterly data, over the 4 quarters of the year, is taken to be the reference value (X_0). Sometimes a longer period is chosen: the reference value for the consumer price index (CPI) is 1982-1984 – the average over 36 months.

Example G.16. Restate French employment data in Table G.2 as index numbers with reference periods of 1991, 1995, and 1998-1999.
Answer. The reference value 1991 is 22,316; for 1995, 22,311; for 1998-99, 21,670.5 (=22,479 + 20,864)/2). The three index numbers are shown in Table G.5.

Index numbers express data as a percentage of the reference value. In effect, they are a way of converting changes into a levels. They are particularly useful when we wish

to emphasize comparisons. Index numbers are most commonly used with time series, but not exclusively: Figure 6.11 (Chapter 6) displays a cross section of national productivity levels, taking the U.S. level as the reference value.

Index numbers are also useful when we want to display a variety of data with incommensurable units: Figure 5.1 (Chapter 5) plots the time series for personal income (units: billions of dollars), employment (units: thousands of people), and industrial production (units: an index with 1997 = 100) all against the same axis by converting each to an index number with 1959:01 = 100.

Once an index number has been calculated, it may be rebased by treating the index itself as the original dataset and applying formula (G.10).

Example G.17. Using the index for French employment (1998-99 = 100) in Table G.5, find the index number for 1993 with the reference period 2000 =100.
Answer. From the table we know that the value for 1991 is 95.5 and for 2000, 107.3. Using formula (G.10), $EmploymentI_{1993} = (95.5/107.3)100 = 89.0$. This is exactly what we would get using the original data: 20,705 for 1993 and 23,262 for 2000: $EmploymentI_{1993} = (20{,}705/23{,}262)100 = 89.0$.

G.8.2 PRICE INDICES

Price indices are a weighted average of individual prices, expressed as an index number. Chapter 4 (section 4.1) provides a detailed discussion about the construction of price indices, concentrating on the choice of the weights for the **price factor**. The price factor takes the place of the term in parentheses in formula (G.10). Here we present the detailed formulae for the price factors for the main indices.

Table G.5
Index Numbers: Employment in France

	Original Data	Index Numbers with Reference Periods		
	thousands	1991=100	1995=100	1998-99=100
1991	22,316	100.0	110.3	103.0
1992	21,609	96.8	106.8	99.7
1993	20,705	92.8	102.3	95.5
1994	21,875	98.0	108.1	100.9
1995	20,233	90.7	100.0	93.4
1996	22,311	100.0	110.3	103.0
1997	20,413	91.5	100.9	94.2
1998	22,479	100.7	111.1	103.7
1999	20,864	93.5	103.1	96.3
2000	23,262	104.2	115.0	107.3

Source: International Monetary Fund, *International Financial Statistics.*

Laspeyres (or Base-weighted) Index

Number each good (j) from 1 to n. Then p_{jt} is the price and q_{jt} is the quantity of good j in period t. Let $t = 0$ indicate the **base period** (that is the period for which the expenditure shares would be calculated). A simple formula for the price factor for Laspeyres index is

(G.11)
$$pf_t^L = \frac{\sum_{j=1}^{n} p_{jt} q_{j0}}{\sum_{j=1}^{n} p_{j0} q_{j0}}.$$

It is easy to show that this formula, in fact, weights the changes in the price of individual goods by their shares in total expenditure in the base period just as in Chapter 4 (section 4.1.1). Notice, first, that the denominator is total expenditure in the base period (that is, base period GDP if the basket is all the final goods and services in the economy). Multiplying and dividing by p_{jt} in the numerator and rearranging yields

$$pf_t^L = \frac{\sum_{j=1}^{n} p_{jt} q_{j0}}{\sum_{j=1}^{n} p_{j0} q_{j0}} = \sum_{j=1}^{n} \left(\frac{p_{jt}}{p_{j0}} \right) \left(\frac{q_{j0} p_{j0}}{\sum_{j=1}^{n} p_{j0} q_{j0}} \right).$$

The first term on the right-hand side of (2) is the price factor for the good j. The numerator of the second term is expenditure on good j in the base period. Consequently, the second term as a whole is the share of expenditure on good j in the base period. The

price factor is, then, the sum of each price change times its share in expenditure in the base period.

Example G.18. What is the Laspeyres price index for 2003, 2004, and 2005 in the couch-potato economy in Table 4.1?

Answer. Take 2003 to be the base year and call tortilla chips good 1 and beer good 2, then the price factor for 2003 is:

$$pf^{L}_{2003} = \frac{\sum_{j=1}^{2} p_{j2003} q_{j2003}}{\sum_{j=1}^{2} p_{j2003} q_{j2003}} = \frac{0.50 \times 5 + 0.75 \times 4}{0.50 \times 5 + 0.75 \times 4} = 1.000\,;$$

for 2004:

$$pf^{L}_{2004} = \frac{\sum_{j=1}^{2} p_{j2004} q_{j2003}}{\sum_{j=1}^{2} p_{j2003} q_{j2003}} = \frac{1.00 \times 5 + 1.25 \times 4}{0.50 \times 5 + 0.75 \times 4} = 1.818\,,$$

for 2005:

$$pf^{L}_{2005} = \frac{\sum_{j=1}^{2} p_{j2005} q_{j2003}}{\sum_{j=1}^{2} p_{j2003} q_{j2003}} = \frac{1.25 \times 5 + 1.40 \times 4}{0.50 \times 5 + 0.75 \times 4} = 2.118\,.$$

Using these price factors and taking 2003 to be the reference year (that is, $p^{L}_{2003} = 100$), what the values of the Laspeyres index for each year are:

$$p^{L}_{2003} = p^{L}_{2003} \times pf^{L}_{2003} = 100 \times 1.000 = 100.0\,,$$

$$p^{L}_{2004} = p^{L}_{2003} \times pf^{L}_{2004} = 100 \times 1.818 = 181.8\,,$$

$$p^{L}_{2005} = p^{L}_{2003} \times pf^{L}_{2005} = 100 \times 2.118 = 211.8\,.$$

These calculations are equivalent to substituting the price factors for the term in parentheses in formula (G.10). They are also agree with the example in Chapter 4 (section 4.1.1).

Paasche (or Current-weighted) Index

The formula for calculating the price factor for the Paasche index is:

(G.12)
$$pf_t^P = \frac{\sum_{j=1}^{n} p_{jt} q_{jt}}{\sum_{j=1}^{n} p_{j0} q_{jt}}.$$

Since the Paasche is a current-weighted index, the base period is now period *t*, and the price factors expresses the ratio of prices in the current period (*t*) to those of some earlier period (0). It is easily shown to be the inverse of the sum of the individual price changes weighted by their shares in current expenditure. Notice that the numerator is total expenditure in the current period.

$$pf_t^P = \frac{\sum_{j=1}^{n} p_{j1} q_{j1}}{\sum_{j=1}^{n} p_{j0} q_{j1}} = \sum_{j=1}^{n} \frac{1}{\left(\frac{p_{j0}}{p_{j1}}\right)\left(\frac{p_{j1} q_{j1}}{\sum_{j=1}^{n} p_{j1} q_{j1}}\right)}.$$

The first term in the denominator on the right-hand side is the individual price factor for good *j* and the second term is the share of expenditure on good *j* in the current period.

Example G.19. What is the Paasche price index for 2003 and 2004 in the couch-potato economy in Table 4.1?
Answer. Taking 2003 as the reference year, then, by definition, the price factor for 2003 is 1.000. For 2004:

$$pf^{P}_{2004} = \frac{\sum_{j=1}^{2} p_{j2004} q_{j2004}}{\sum_{j=1}^{2} p_{j2003} q_{j2004}} = \frac{1.00 \times 4 + 1.25 \times 5}{0.50 \times 4 + 0.75 \times 5} = 1.783 .$$

So, the price indices are:

$$p^{P}_{2003} = p^{P}_{2003} \times pf^{P}_{2003} = 100 \times 1.000 = 100.0 ,$$

$$p^{P}_{2004} = p^{P}_{2003} \times pf^{P}_{2004} = 100 \times 1.783 = 178.3 .$$

which agree with the calculations in Chapter 4 (section 4.1.2).

Since the Paasche index uses new weights for each current period, the price factors and the values of the price index will different if a later year, say 2003, were taken to be the current period. While Laspeyres indices are very common, Paasche indices are used most often as a step in the calculation of chain indices, where the shifting base is a desired feature.

The Fisher-ideal Index

The Fisher-ideal index is the geometric average of the Laspeyres and the Paasche indices as shown in Chapter 4, equation (4.1). The Fisher-ideal index is almost always used as a step in computing chain indices in which the base is updated each period. In this context,

the base period for the Laspeyres component is period $t - 1$, while the base period for the Paasche component is period t. A general formula for the price factor of a Fisher-ideal index is then:

$$\text{(G.13)} \qquad pf_t^F = \sqrt{pf_t^L \times pf_t^P} = \sqrt{\frac{\sum_{j=1}^{N} p_{jt} q_{jt-1}}{\sum_{j=1}^{N} p_{jt-1} q_{jt-1}} \times \frac{\sum_{j=1}^{N} p_{jt} q_{jt}}{\sum_{j=1}^{N} p_{jt-1} q_{jt}}}.$$

Example G.20. What is the Fisher-ideal price index for 2003 and 2004 in the couch-potato economy in Table 4.1?

Answer. Using formula (G.13), the price factor for 2004:

$$pf_t^F = \sqrt{\frac{\sum_{j=1}^{2} p_{j2004} q_{j2003}}{\sum_{j=1}^{2} p_{j2003} q_{j2003}} \times \frac{\sum_{j=1}^{2} p_{j2004} q_{j2004}}{\sum_{j=1}^{2} p_{j2003} q_{j2004}}} = \sqrt{\frac{1.00 \times 5 + 1.25 \times 4}{0.50 \times 5 + 0.75 \times 4} \times \frac{1.00 \times 4 + 1.25 \times 5}{0.50 \times 4 + 0.75 \times 5}}$$

$$= \sqrt{181.8 \times 178.3} = 1.800,$$

which agrees with the results of Chapter 4 (section 4.1.3). If 2003 is the reference year, then by definition $p_{2003}^F = 100$. For 2004, and $p_{2004}^F = p_{2003}^F \times pf_{2004}^F = 100 \times 1.8000 = 180.0$.

G.9 Real and Nominal Magnitudes

G.9.1 Conversions between Real and Nominal Magnitudes

Converting Nominal Data to Real

Fundamentally, a nominal or market value is a number of dollars; a real value is the number of units of a good that those dollars will buy. If Big Macs cost $3.69, the $100 is a nominal value equivalent to the real value of 27.1 (= 100/3.69) Big Macs. Price indices are typically based on a basket of goods. If we knew the dollar cost of the basket, then we could convert any dollar value into a real value measured as a number of baskets. But price indices are generally published only as index numbers and, as we saw in section G.8.1, index numbers are really a way of turning relative changes into levels. We can use index numbers (see Chapter 2, sections 2.4.1 and 2.4.2), to restate the value of a nominal quantity in the dollars of one year into the *constant dollars* of another year. This is similar to creating an index number as in section G.8.1, except that, instead of taking the reference value to be 100, we take the reference value to be the nominal value in the reference period. The general formula for converting any nominal monetary value X to a real value is:

$$\$_R X_t = (\$_t X_t)(p_R / p_t), \tag{G.14}$$

where the subscript R refers to the reference period, and the subscript on the dollar sign indicates in which period's money the quantity is measured.

Example G.21. Mexican exports in 2003 were 165,396 million pesos. If the consumer price index (2000 = 100) was 18.56 in 1990 and 116.80 in 2003, what was the real value of Mexican exports in 2000 pesos? In 1990 pesos?
Answer. Using formula (G.14) and, of course, substituting pesos for dollars, $\text{Pesos}_{2000}\textit{Exports}_{2003} = (\text{Pesos}_{2003}\textit{Exports}_{2003})(p_{2000}/p_{2003}) = 165{,}396(100.0/116.80) = \text{Pesos}_{2000}141{,}606$ million. Similarly, $\text{Pesos}_{1990}\textit{Exports}_{2003} = (\text{Pesos}_{2003}\textit{Exports}_{2003})(p_{1990}/p_{2003}) = 165{,}396(18.56/116.80) = \text{Pesos}_{2000}26{,}282$ million.

Converting Real Data to Nominal

To convert real data to nominal, we simply work this last process backward. Starting with equation (G.14) solve for the nominal value to yield:

(G.15) $$\$_t X_t = (\$_R X_t)(p_t / p_R) .$$

Example G.22. U.S. real GDP in the fourth quarter of 2004 was $10,993.3 billion in constant 2000 dollars. If the implicit price deflator for 2004:4 was 109.06, what was nominal GDP?
Answer: Using formula (G.15),

$$\$_{2004:4}Y_{2004:4} = (\$_{2000}Y_{2004:4})(p_{2004:4} / p_{2000}) = 10{,}993.3(109.06/100.00)$$
$$= \$_{2004:4}11{,}998.9 \textit{ billion.}$$

Converting Real Data of One Reference Period to that of Another Period

If data are already expressed in constant dollars of one reference period ($R1$), we can convert them to the constant dollars of another reference period ($R2$) using the following formula:

(G.15) $$\$_{R2} X_t = (\$_{R1} X_t)(p_{R2} / p_{R1}) .$$

Example G.22. According to Example G.21, the real value of Mexican exports in 2003 was Pesos_{2000}26,282 million in 1990 constant pesos. Using this value and the fact that the Mexican CPI was 18.56 in 1990 and 100 in 2000, what was the real value of Mexican exports in 2000 constant pesos?
Answer. Using formula (G.15) and, of course, substituting pesos for dollars, $\text{Pesos}_{2000}Exports_{2003} = (\text{Pesos}_{1990}Exports_{2003})(p_{2000}/p_{1990}) = 26{,}282\ (100.0/18./56)$ $= \text{Pesos}_{2000}141{,}606$ million. Naturally, this is the same value as when we converted nominal exports in 2003 to 2000 constant pesos; the calculation is different because we started here with real (i.e., constant peso) values rather than nominal (i.e., current values).

G.9.2 Real Values Using Chain-weighted Indices

The construction of chain-weighted indices from underlying Laspeyres and Paasche indices, as well as the calculation use of indices to convert nominal to real values, are described in Chapter 4 (section 4.1.4). Chain-weighted indices provide the best method for comparing the levels of a given series at different times. Unfortunately, because of the shifting weights used in their construction chain-weighted price indices have some undesired properties.

Table G.6 shows nominal and its components, as well as their real counterparts in (chain-weighted)1996 constant dollars for 1991, 1996, and 2001. Notice that using nominal values for any of these years the identity $Y = C + I + G + NEX$ holds exactly. Of course, in 1996 real and nominal values are the same, so that the identity also holds for real values in 1996. But notice that for real values in 1991, $C + I + G + NEX$ = \$6,683.7 billion > \$6,676.4 = Y. The difference is shown as the residual: –7.3. (It is easy to check that the identity also fails in 2001.) The general rule that the whole is the sum of its parts

Table G.6. Nominal and Real (Chain-weighted) GDP and Its Components

	Nominal (billions)					Real (billions chain-weighted 1996 constant dollars)					
	GDP	**Consumption**	**Investment**	**Government Expenditure**	**Net Exports**	**GDP**	**Consumption**	**Investment**	**Government Expenditure**	**Net Exports**	**Residual**
1991	5,986.2	3,971.2	800.2	1,235.5	-20.7	6,676.4	4,466.6	829.5	1,403.4	-15.8	-7.3
1996	7,813.2	5,237.5	1,242.7	1,421.9	-89.0	7,813.2	5,237.5	1,242.7	1,421.9	-89.0	0.0
2001	10,082.2	6,987.0	1,586.0	1,858.0	-348.9	9,214.5	6,377.2	1,574.6	1,640.4	-415.9	22.6

Source: Department of Commerce, Bureau of Economic Analysis.

fails when chain-weighted indices are used to calculate real values. In contrast, fixed-weight indices preserve the rule.

The difference between the real value of GDP and the sum of the real values of its components is reported (as it is in the NIPA tables) as a *residual*. The residual is usually small – often small enough to ignore altogether.

The residual can cause problems. For instance, suppose that we ask, how has the proportion of GDP that is absorbed by government expenditure changed over time? Since both income and expenditure are, in fact, conducted in the current dollars of the day, the correct answer to this question is easily calculated from the nominal (or market) values. The true share for 1991 is Nominal G_{1991}/Nominal Y_{1991} = 1,235.5/5,986.2 = 20.6 percent. Suppose, however, that we tried to calculate this share using (chain-weighted) constant values. Then, the share for 1991 would be Real G_{1991}/Real Y_{1991} = 1,403.4/6,676.4 = 21.0 percent. Table G.7 shows the shares for each of the years in Table G.6 using nominal and real values. In 1996, of course, the shares are the same either way. But in every other year, the real shares are systematically different.

A similar problem occurs in computing the contribution of different components of GDP to the growth rate of GDP (or, equally, the contribution of the components of any series to the growth rate of the whole). Great care must be exercised whenever comparisons are made between different chain-weighted real time series. These calculations become more and more misleading the further away they are from the reference year. In computing shares it is best to use the nominal values. To calculate the correct contributions of real components to the growth of the whole is well understood, but somewhat complex. Fortunately, the NIPA includes supplemental tables that make

Table G.7. Shares in GDP Calculated Using Nominal and Chain-weighted Real GDP

	Percentage of Nominal GDP				Percentage of Chain-weighted Real GDP			
	Consumption	**Investment**	**Government Expenditure**	**Net Exports**	**Consumption**	**Investment**	**Government Expenditure**	**Net Exports**
1991	66.3	13.4	20.6	-0.3	66.9	12.4	21.0	-0.2
1996	67.0	15.9	18.2	-1.1	67.0	15.9	18.2	-1.1
2001	69.3	15.7	18.4	-3.5	69.2	17.1	17.8	-4.5

Source: Table 1.

the necessary calculations appropriately. The NIPA also include supplemental tables that use a variety of reference years to facilitate accurate real comparisons.

G.10 Growth Rates

The main discussion of growth rates is found in Chapter 2, Box 2.2. In this section, we discuss some additional aspects.

G.10.1 When Should Growth Rates Be Compounded?

Compound growth rates make sense when the growth of one period is added to the stock and itself grows the next period. They are, therefore, natural when considering the growth of real GDP, population, employment, and most other macroeconomic variables. However, sometimes the growth of one period is siphoned off and each period starts with the same base.

For example, suppose that you have $100,000 in a bank account that bears 1 percent interest per quarter. Each quarter you receive $1,000. If you left that money in the account and let it grow alongside the original principal, then compounding would be natural. But suppose that each quarter you spend the $1,000. Then what is your *annual* rate of interest. It is just four times your quarterly rate of interest, since in one year you earn $4,000 or 4 percent on your principal. This is an example of **simple annualization**: quarterly rates of growth are multiplied by four, monthly by twelve, and so forth.

The most common application of simple annualization in macroeconomics and finance is the formulae used to price debt and compute payments (e.g., mortgage

payments). Interest rates are general quoted as simple annual rates. To find the monthly rate, for example, they are divided by twelve.

Example G.23. A bank account pays interest at 7.30 percent per year paid each month. If you put $200 in the account and hold it for one year, letting the interest be held in the account: (a) what is the compound rate of interest on your account? and (b) how much will you have at the end of a year?
Answer. The simple monthly rate of interest is 0.608 (= 7.30/12) percent. The compound annual interest $r_{annual}^{compound} = (1 + r_{monthly}^{simple})^{12} - 1 = (1.00608)^{12} - 1 = 7.54$ percent. The value of your holding at the end of the year is 200(1.0754) = $215.09.

The general formula relating the simple to compound annual rates is:

(G.16) $$\hat{X}_{annual}^{compound} = \left(1 + (\hat{X}_{annual}^{simple} / m)\right)^{m} - 1,$$

where m = the frequency of compounding: 4 for quarterly, 12 for montly; 52 for weekly, and so forth.

Example G.24. In example G.23, how would the compound annual rate of interest differed if the bank had compounded daily instead of quarterly?
Answer. Using formula (G.15), the annual rate compounded daily would be $r_{annual}^{compound} = \left(1 + (r_{annual}^{simple} / 365)\right)^{365} - 1 = \left(1 + (0.073/365)\right)^{365} - 1 = 7.57$ percent, which is higher than 7.54 percent, the annual rate compounded quarterly in Example G.23.

G.10.2 EXTRAPOLATION

The compound average annual rate of growth for real GDP in the United Kingdom for the nine years from 1991 to 2000 (see Table G.1) is 2.97 percent (= $(617/474)^{1/9} - 1$). What would real GDP be in 2008 if this rate of growth continued steadily? This is a

problem in **extrapolation**. A general formula for extrapolating time-*t* data *k* periods into the future is:

$$X_{t+k} = X_t(1+\overline{\hat{X}})^k, \tag{G.17}$$

where $\overline{\hat{X}}$ is the average rate of growth *expressed in the same time units as* k (e.g., if *k* is measured in quarters, $\overline{\hat{X}}$ must be the *quarterly* compound average rate). Using the formula, real GDP in the U.K. in 2005 would be

$$Y_{2008} = X_{2000}(1+\overline{\hat{X}})^8 = 617(1+0.0297)^8 = \$780 \text{ billion.}$$

Example G.25. Employment in Italy grew at a 1.54 percent compound annual rate between 1997:4 and 2002:4. Employment in 2002:4 was 21,757,000. If growth continued at a steady rate, what would be the level of employment in 2009:1?
Answer. First, we need the average compound quarterly rate of growth: $\overline{\hat{E}} = (1+0.0154)^{1/4} - 1 = 0.38$ percent. Then, 2009.1 is 33 quarters ahead of 2002:4, so that *k* = 33. Using formula (G.17), we get $E_{2009:1} = E_{2002:4}(1+\overline{\hat{E}})^{33} = 21{,}757{,}000(1.0038)^{33} = 24{,}691{,}529$.

G.11 Logarithms

G.11.1 WHAT ARE LOGARITHMS?

The Concept of the Logarithm

In the days before electronic calculators, personal computers, and spreadsheet programs, logarithms (first developed by Scottish mathematician John Napier (1550-1616)) were an important tool for calculation. Today they remain useful as tools of analysis.

A **logarithm** *of a number* is *the power (or exponent) to which a given base must be raised to produce that number*. So, for example, the base 10 logarithm (notated as $\log_{10}$) of 100 is:

$$\log_{10}(100) = 2, \text{ because } 10^2 = 100$$

Similarly,

$$\log_{10}(1000) = 3, \text{ because } 10^3 = 1000$$

Logarithms are not restricted to integer powers:

$$\log_{10}(43) = 1.633, \text{ because } 10^{1.633} = 43$$

(Check this with a pocket calculator.)

Base 10 logarithms are known as **common logarithms**. Generally the button on a calculator marked "log" computes the common logarithm. There is, however, nothing mathematically special about base 10:

$$\log_2(43) = 5.426, \text{ because } 2^{5.426} = 43$$

The Anti-logarithm

If we know the logarithm, we can calculate the original number by raising the base to the power of the logarithm. So, for example, if we are given a common logarithm of 3.478, we can calculate the original number as

$$10^{3.478} = 3{,}006.076$$

This is referred to as taking the **antilogarithm**. For common logarithms it is sometimes notated as antilog_{10} and sometimes as $\log^{-1}$ (read as "log inverse"), so that

$$\text{antilog}_{10}(3.478) = \log^{-1}(3.478) = 3{,}006.076.$$

The antilogarithm "undoes" or reverses the logarithm and vice versa.

The Natural Logarithm

Common logarithms are convenient for expositional purposes, because everyone is familiar with the powers of 10. Base 10 was the most frequently used base in the past when logarithms were commonly used for calculation; hence the name "common logarithm." In many scientific applications, however, another – and, at first sight, quite strange – base is preferred. This base is given the name *e*, and is defined as $e = 2.71828\ldots$ The ellipsis marks indicate that *e* is an irrational number (like π) that cannot be expressed as a ratio of integers, and, therefore, never terminates or repeats. Logarithms to base *e* follow the same rules as logarithms to base 10 or any other base.

For example, $\log_e(17) = 2.83321$, because $e^{2.83321} = 17$. antilog_e is often written exp, so that $\text{antilog}_e(2.83321) = \exp(2.83321) = e^{2.83321} = 17$.

Logarithms to base *e* are called **natural logarithms**. This name and the explanation for the strange value of *e* arise because the base *e* logarithm is involved naturally in the solution to important problems in mathematics. One such problem arises in the integral calculus: what is the area under the rectangular hyperbola $y = 1/x$ between $x = a$ and $x = b$?[2] Mathematicians have proved that the answer is $\log(b) - \log(a)$. But this answer will be correct numerically only if the logarithms are expressed to the base 2.71828 . . . . In other words, $\log(a) = \log_e(a)$ and so forth.

In the main text of this book, only natural logarithms are used. Natural logarithms are indicated by several different notations. The most common are $\log_e$, ln, and log. Frequently ln is used on calculators and in spreadsheets to avoid confusion with common logarithms. Unfortunately, in many contexts, the lower case "l" in ln can be mistaken for the numeral "1". We shall indicate natural logarithms as log, without any subscript, which has the virtue of suggesting its correct pronunciation.

G.11.2 Calculating with Logarithms

Logarithms convert multiplication problems into addition problems, which are generally easier to solve (at least by hand). To see how this works, consider multiplying 10,000 by

---

[2] For those familiar with the integral calculus the problem is to evaluate $\int_a^b (1/x)dx$.

1,000,000. Of course, this is easy to do, but it illustrates the principle. We can write the problem as

$$10{,}000 \times 1{,}000{,}000 = 10^4 \times 10^6$$

The rules of exponents tell us that we can rewrite the right-hand term as $10^4 \times 10^6 = 10^{4+6} = 10^{10}$; that is, we sum the exponents. So,

$$10{,}000 \times 1{,}000{,}000 = 10^4 \times 10^6 = 10^{10} = 10{,}000{,}000{,}000$$

The exponents are, of course, just the common logarithms. Adding the exponents is just like adding the logarithms. Therefore, another way of approaching the problem would be to say:

$$\log_{10}(10{,}000) = 4$$

$$\log_{10}(1{,}000{,}000) = 6$$

$$4 + 6 = 10$$

$$\text{antilog}_{10}(10) = 10^{10} = 10{,}000{,}000{,}000$$

Logarithms reduce the level of complexity of many calculations: they convert multiplication problems into addition problems; division into subtraction; raising to powers into multiplication; and taking roots into division. The main rules for (natural) logarithms are:

(G.18) If $xy = z$, then $\log(x) + \log(y) = \log(z)$

(G.19) If $x/y = z$, then $\log(x) - \log(y) = \log(z)$

(G.20) If $x^y = z$, then $y\log(x) = \log(z)$

(G.21) If $\sqrt[y]{x} = z$, then $\log(x)/y = \log(z)$

(G.22) If x is a natural number, then

$$\text{antilog}(\log(x)) = \log^{-1}(\log(x)) = \exp(\log(x)) = e^{\log(x)} = x$$

(G.23) If y is a logarithm, then $\log(\exp(y)) = \log(\log^{-1}(y))$

$$= \log(e^y) = y.$$

(G.24) For small x, $\log(1 + x) \approx x$.

Example G.26. Use logarithms to calculate (a) 1,356 × 43,119; (b) 317 ÷ 48; (c) 211^{14}; (d) $\sqrt[4]{12{,}591}$; (e) x, when $\log(x) = 0.57$; (f) y, when $\exp(y) = 8.22$; (g) $\log(1.024)$.

Answer. Using rules (G.18)-(G.24) in turn:

(a) $\log(1{,}356) + \log(43{,}119) = 7.2123 + 10.6717 = 17.8840$; $\exp(17.8840) =$ 58,468,576;

(b) $\log(317) - \log(48) = 1.88770$; $\exp(1.88770) =$ 6.6042;

(c) $\log(211^{14}) = 14\log(211) = 14(5.35186) = 74.92604$; $\exp(74.92604) =$ 3.4671×10^{32};

(d) $\log(\sqrt[4]{12{,}591}) = \log(12{,}591)/4 = 9.44074/4 = 2.36018$; $\exp(2.36018) =$ 10.5929.

(e) $\exp(0.57) =$ 1.0587;

(f) $\log(8.22) =$ 2.1066;

(g) without calculation (applying formula (G.24)) $\log(1.024) \approx$ 0.024; more exactly $\log(1.024) =$ 0.0237.

G.11.3 LOGARITHMS AND GROWTH

The connection between logarithms and growth rates and the algebra of growth rates is addressed informally in Chapter 2 (Box 2.3) and Chapter 7 (Box 7.1). In this section, we amplify that discussion.

Logarithmic Derivatives and Percentage Changes

Start with a fact that is demonstrated in any standard calculus textbook:

$$\text{if } z = \log(x), \text{ then } dz/dx = d\log(x)/dx = 1/x.$$

Recall that a derivative is the rate of change of a function. So, if we want to know how much z changes for a small change in x, we must multiply the *rate* at which z changes for any change in x times the *change* in x itself: $dz = (dz/dx)dx$. In the case of $z = \log(x)$, this is

$$dz = (dz/dx)dx = (d\log(x)/dx)dx = (1/x)dx = dx/x$$

Notice that for small changes $dx \approx \Delta x$, so that

$$dz \approx \Delta x/x = \textit{the percentage change in } x$$

That is: *a small change in the logarithm of a variable (*dz*) is approximately equal to the percentage change in the variable itself.*

Example G.27. Suppose that the price level in one month is $p_t = 123$ and the next month it is $p_{t+1} = 124$, what is the percentage change?
Answer. Using logarithms, $d\log(p)/dt \approx \log(124) - \log(123) = 4.82028 - 4.81218 = 0.0081$ or 0.81 percent. This is the same as direct calculation: $(124/123 - 1) = 0.0081$.

The next example illustrates an important caveat: the logarithmic approximation works *only* for *small* changes:

Example G.28. Suppose that the price level in one month is $p_t = 123$ and the next month it is $p_{t+1} = 275$, what is the percentage change?
Answer. Using logarithms, $d\log(p)/dt \approx \log(275) - \log(123) = 5.54908 - 4.81218 = 0.7369$ or 73.69. But the exact result from direct calculation is very different: $(275/123 - 1) = 0$. 1.2358 or 123.58 percent. The difference in logarithms understates the actual percentage change by almost half.

Logarithms and Growth Rates

Consider a time series x. We can think of x as a function of time t. So, for example, when we plot GDP on a graph, the equation that describes GDP might be $z = Y(t)$. Take the logarithm of each side of this expression: $\log(z) = \log(Y(t))$. What is the derivative of $\log(z)$ with respect to time? Using the rule for the derivative of a logarithmic function and the chain rule:

$$d\log(z)/dt = d\log(Y(t))/dt = [1/Y(t)][d(Y(t))/dt] \tag{G.25}$$

The first term in square brackets is the original logarithmic derivative. The second term in square brackets follows from the chain rule: it is the derivative of the function within the function.

We can interpret the right-hand side of (G.25) in a practical way. Rearranging the terms gives us $\frac{dY(t)/Y(t)}{dt}$. Remembering that for small changes $dx \approx \Delta x$, we see that

(G.26) $$d\log(z)/dt = [1/Y(t)][d(Y(t))/dt] = \frac{dY(t)/Y(t)}{dt} \approx \frac{\Delta Y(t)/Y(t)}{\Delta t}$$

The numerator of the right-hand side is the percentage change in *Y*. (Here time *t* rather than the more common time $t - 1$ is taken to be the base period. This difference will not matter so long as all the changes are small.) The denominator is a small change in time. The whole expression can be read as *the percentage change per unit time*, which is, of course, the definition of the *rate of growth*. The general rule, then, is:

(G.27) For small changes $\frac{d\log(z)}{dt} = \frac{dz/z}{dt} \approx \frac{\Delta z/z}{\Delta t}$

$= \hat{z}$ = the *percentage change in z per unit time*

or *the rate of growth of z*,

where the circumflex or "hat" (^) over a variable indicates the rate of growth.

The derivative of a function measures the slope of its graph. We have, therefore, proved mathematically a result introduced in Chapter 2, Box 2.3: *the slope of a logarithmic time-series graph of a variable is the rate of growth of the variable*.

Similarly, we can demonstrate why there is a strong analogy between the rules for logarithms ((G.18) –G(21)) and the algebra of growth rates in Chapter 2, Box 7.1.

Example G.29. What is the growth rate of $z = xy$?
Answer. $\hat{z} = d\log(z)/dt = d\log(xy)/dt = d\log(x)/dt + d\log(y)/dt = \hat{x} + \hat{y}$, which is just rule (B7.1.1) from Chapter 7, Box 7.1.

The other rules of the algebra of growth rates are derived similarly.

Continuous Compounding

Look again at the formula for compound growth in equation (G.16). The ultimate compound annual growth rate depends on the frequency of compounding. Going from quarterly ($m = 4$) to monthly ($m = 12$) to weekly ($m = 52$) to daily ($m = 365$), the compound growth rate increases as the time interval between compound increments becomes smaller. Any standard calculus book proves that as m becomes infinitely large, so that the time interval between compound increments becomes infinitesimally small, the compound growth formula converges to:

(G.28) $$\hat{X}_{annual}^{compound} = \exp(\hat{X}_{annual}^{simple}) - 1.$$

Example G.30. In Examples G.23 and G.24, how would the compound annual rate of interest differ if the bank had compounded continuously instead of daily or quarterly?
Answer. Using formula (G.28), the annual rate compounded daily would be $r_{annual}^{compound} = \exp(r_{annual}^{simple}) - 1 = \exp(0.073) - 1 = 7.57$ percent, which is higher than the annual rate compounded quarterly 7.54 percent, but the same as the rate compounded daily.

Look again at the extrapolation problem in equation (G.17). If we replace $1 + \overline{\hat{X}}$ with its continuously compounded equivalent, $\exp(\overline{\hat{X}})$, we get

(G.29) $$X_{t+k} = X_t[\exp(\bar{\hat{X}})]^k = X_t k \exp(\bar{\hat{X}}).$$

Taking logs of both sides gives

$$\log(X_{t+k}) = \log(X_t) + k\bar{\hat{X}},$$

and rearranging

(G.30) $$\bar{\hat{X}} = \frac{\log(X_{t+k}) - \log(X_t)}{k}.$$

Equation (G.30) says that we can compute the continuously compounded average rate of growth as the difference in logarithms (a percentage change) divided by the number of periods over which that difference is taken (k). The units of k determine time units of $\bar{\hat{X}}$ -- e.g., if k is measured in years, $\bar{\hat{X}}$ will be an annual rate; if k is measured in months, a monthly rate.

Example G.31. Real GDP in Sweden in 1995:3 was 403.39 constant 1995 krona and in 2004:1, 553.41 krona. What is the average compound rate of growth over this period at (a) a quarterly rate; and (b) an annual rate?
Answer. Use formula (G.30). Measured in quarters the time between the two dates is $k = 34$ quarters; in years, $k = 8.5$ years. So (a)
$\bar{\hat{Y}} = \frac{\log(Y_{2004:1}) - \log(X_{1995:3})}{34} = \frac{\log(553.41) - \log(403.39)}{34} = 0.93$ percent per quarter;

and (b) $\bar{\hat{Y}} = \frac{\log(Y_{2004:1}) - \log(X_{1995:3})}{8.5} = \frac{\log(553.41) - \log(403.39)}{8.5} = 3.72$ percent per year, which is, of course, four times the quarterly rate.

The Rule of 72

Is a particular growth rate fast or slow? It often helps to have a rough-and-ready guide. An easy one is the **rule of 72**: *the doubling time of any growing quantity is 72 divided by its growth rate expressed in percentage points.*

Example G.32. If the inflation rate is 12 percent per year, how long will it take the price level to double?
Answer. The doubling time is approximately 72/12 = 6 years.

Why does the rule of 72 work? Consider a quantity X and ask, how long does it take at a continuously compounded rate of growth for it to double, that is to reach $2X$ at a rate of growth $\bar{\hat{X}}$? Substitute these values into equation (G.30):

$$\bar{\hat{X}} = \frac{\log(2X) - \log(X)}{k} = \frac{\log(2X / X)}{k} = \frac{\log(2)}{k}.$$

Solving for k gives

(G.31) $$k = \frac{\log(2)}{\bar{\hat{X}}} = \frac{0.69}{\bar{\hat{X}}}.$$

The growth rate in equation (G.31) is measured in natural units, but if we want to measure it in percentage points, then it becomes

(G.31') $$k = \frac{100\log(2)}{\hat{\bar{X}}} = \frac{69}{\hat{\bar{X}}}.$$

Using the numbers in Example G.32, k = 69/12 = 5.75 years, which is a more exact doubling time than the 6 years estimated previously. So, the rule of 72 is really a *rule of 69*. Why not state it that way?

The answer goes back to the initial assumption: we need a *rough-and-ready* guide to the speed of growth. The number 69 has four integer factors {1, 3, 23, 69}. In constrast, 72 has twelve integer factors {1, 2, 3, 4, 6, 8, 9, 12, 18, 24, 36, 72}. Seventy-two is, therefore, an easier number to use for mental arithmetic, and the errors will not matter much if we apply it to low growth rates and do not require a precise answer. Whenever it is easier, any other number close to 69 (68, 69, 70) will work just as well or better than 72.

G.12 Detrending

Detrending time series is discussed in Chapter 5, Box 5.1. In this section we add some practical details on estimating trends and detrending time series.

G.12.1 CONSTANT TRENDS

Constant trends correspond to an equation with fixed coefficients, and include linear and exponential trends.

Linear Trends

Excel (and other spreadsheet programs) allow a variety of trend lines to be added to a time series chart: in *Excel* click on Chart, Add Trendline, Choose Trend/Regression Type, and then click on the type you want (for the trends in this book usually Linear or Exponential); click OK and the chart will display a trend. It is usually a good idea to display the equation of the trend line: before clicking OK, click on the Options tab and check the box that says Display Equation on Chart. (Note that the Options tab also allows you to extrapolate a trend forward or backward on your chart.)

The equation is useful in forming the **detrended series** or the deviations from the trend: *deviations from the trend = time series – trend*. Once you have the equation of the trend, add a column to your worksheet that starts with 1 for the first period in the sample over which the trend was formed, 2 for the next, and so on to the end. Say that this is column C, although it could be any column and that 1 is place in cell C2. The equation for a linear trend will be reported on your chart as $y = bx + a$, where b and a are numerical coefficients and x stands for time measured starting from 1, just like your entries in column C. The value of the trend for cell for period 1, then, can be entered into some other cell as "= b*C2 + a". *Excel* does not recognize an equation cut from the chart and pasted into the spreadsheet, so you must copy the numbers for a and b by hand.

Once you have all the trend values, then the detrended values are easily computed. (*A test that your computation is correct*: plot your computed trend on the same chart as your added trend line; they should be indistinguishable.)

Example G.33. Suppose that the equation of a linear trend line is given as $y = 0.4896x + 11.699$ and the value of original time series in period 7 is 23.06507, what is the value of the trend and the detrended series at period 7.
Answer. Assuming that your time data is in column C with C2 = 1, then period 7 corresponds to row 8. The *trend at period* 7 = 0.4896(7) + 11.699 (which is equivalent to entering =0.4896*C8+11.699 into a cell) = 15.1172. The *detrended value of the series at period 7* = 23.06507 – 15.1172 = 7.947874.

Exponential and Other Constant Trends

The procedures for exponential and other constant trends are similar except that the equations describing the trend take a different form.

Example G.34. Using the same original series as in Example G.34, suppose that the equation of an exponential trend line is given as $y = 22.785e^{0.0086x}$, what is the value of the trend and the detrended series at period 7.
Answer. Assuming that your time data is in column C with C2 = 1, then period 7 corresponds to row 8. The *trend at period* 7 = 22.785 exp(0.0086(7)) (which is equivalent to entering =22.785*exp(0.0086*C8) into a cell) = 24.19878. The *detrended value of the series at period 7* = 23.06507 – 24.19878= –1.13371.

Note the big difference between the trend and detrended values in Examples G.33 and G.34. *the method of detrending matters*. In most cases in macroeconomics that involve growing data, exponential trends are preferred to linear trends. If data have been transformed by taking logarithms of the original time series, linear trends are preferred (another example, of how logarithms reduce a more complex problem to a simpler one).

G.12.2 MOVING-AVERAGE TRENDS

Calculating Moving-average Trends

Excel will fit a *trailing moving-average trend*. In this book, however, we use only *centered moving-average trends*. These are easily constructed from *Excel* functions. If the original series is in column B, the a centered moving-average trend with N leading and N lagging terms can be constructed in another column as AVERAGE(Bc-N:Bc+N), where c is the row number of the current cell.

Example G.35. What is the 15-period centered moving average of data in column B at the time corresponding to row 47?
Answer. The current cell $c = 47$. To have a 15-period average, $N = 7$, since the leading terms plus the lagging terms plus the current period equals the total $(2N + 1)$. So, at row 47, the moving average =AVERAGE(B40:B54).

Dealing with the Endpoint Problem

The *endpoint problem* arises with the moving-average trend because we do not have enough data to calculate a $(2N + 1)$-period moving average for any cell less than $N + 1$ cells from the beginning or the end of the sample. There are various "fixes" that might be made. One simple one is adjust the original series before calculating the moving average by adding N terms to the beginning and end of the sample, where each term takes the value of the mean of the first or last N terms of the original sample.

So, in the last example, if the mean of the last 7 terms were 16.5, we would add 7 cells with 16.5 to the end of the sample. Now when we calculate the moving average – stopping on the last row of the *original* data, the AVERAGE function will find the needed seven additional cells to average over 15 altogether. (The approach would be the

same working with the beginning of the sample, except we would use the average of the first 7 rather than the last 7 terms.)

The last 7 terms of the moving average are not then the same as terms in the middle of the sample. Instead they are a weighted average of the current and 7 trailing terms with the average value of the last seven terms. As we get closer to the endpoint, the weights shift more heavily toward the average.

This is not a perfect solution to the endpoint problem, but it should be good enough for many purposes.

G.13 Correlation

Correlation and the interpretation of the correlation coefficient (R) are discussed informally in Chapter 5, Box 5.2. In this section we provide a more detailed account of the correlation coefficient.

G.13.1 COVARIANCE

The **covariance** of two sets of data provides a measure of the degree to which they move together:

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}. \tag{G.32}$$

First, compare this formula to the formula for the variance, equation (G.7). If X and Y took exactly the same values, cov(X,Y) would equal var(X) = var(Y). But in most cases series are different. If X tends to be *above* its mean, when Y is *above* its mean, the products to be summed will tend to be positive on average, and the covariance will be *positive*. If X tends to be *below* its mean when Y is *above* its mean and vice versa, then the products to be summed will tend be negative on average and the covariance will be *negative*. If there is no relationship between the two variables, then the products will sometimes be positive and sometimes negative, so that when they are summed up they will cancel and be zero on average.

G.13.2 The Correlation Coefficient

Like variance, the actual size of the covariance depends, in part, on the units of measurement. Correlation expresses the covariance as a share of the geometric mean of the variances of the two variables in order to produce a unit-free measure of association:

$$R = \text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\dfrac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}}{\sqrt{\dfrac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}\dfrac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1}}}$$

(G.33)

$$= \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}.$$

Formula (G.33) shows that if there were no relationship between the two variables, so that the covariance were zero, then the correlation coefficient (R) would itself be zero. Similarly, if the two series were identical, then both numerator and denominator (most easily seen in the expression after the second equality) would be equal to the variance, so that $R = 1$.

What would value would R take if the two series moved perfectly together but were not identical? For example, let $X_i = \{1, 2, 4, 1, -1, -8, \ldots\}$ and $Y_i = \{0.5, 1, 2, 0.5, -0.5, -4, \ldots\}$, then $Y_i = \frac{1}{2}X_i$. We can see the answer more clearly by considering two variables each with mean zero, so that $\overline{X} = \overline{Y} = 0$, and $Y_i = bX_i$. Then

(G.34)
$$R = \mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} = \frac{\sum_{i=1}^{N}(X_i)(Y_i)}{\sqrt{\sum_{i=1}^{N}(X_i)^2\sum_{i=1}^{N}(Y_i)^2}}$$
$$= \frac{\sum_{i=1}^{N}(X_i)(bX_i)}{\sqrt{\sum_{i=1}^{N}(X_i)^2\sum_{i=1}^{N}(bX_i)^2}} = \frac{b\sum_{i=1}^{N}(X_i)^2}{\sqrt{b^2}\sqrt{\left(\sum_{i=1}^{N}(X_i)^2\right)^2}} = \frac{b}{\sqrt{b^2}}.$$

If b is any positive number then, the last term in equation (G.34) equals one: *if any two series move perfectly directly proportionally to each other, then* $R = 1$. If b is any negative number, then the last term is negative one: *if any two series move perfectly inversely proportionally to each other, then* $R = -1$.

It is messier to demonstrate, but equally true, that the same results hold if the relationship between X and Y is the more general linear function: $Y_i = a + bX_i$ – in which case one or both variables would have different non-zero means.

Generally, two variables will not move perfectly together, so that we rarely find $R = 1$ or -1 unless we have made a mistake or the two variables are connected as identities. But the more closely the movements in variables conform to each other (directly or inversely), the larger the absolute value of R.

Variables that have low, but non-zero correlations may be either (a) truly, but weakly, related or, (b) unrelated. The second case is always possible with random data. For example, if we keep track of the number and heads and tails of a fair coin over many thousands of flips, at any point we will not typically have an equal number of heads and tails, even though the average will get closer and closer 0.5 (calling heads 1 and tails 0) as the number of flips increases. There are formal statistical tests, beyond the scope of this book, for deciding between option (a) or (b). At this level, the soundest procedure is not to use a very low correlation as positive evidence *by itself* for a real, but weak relationship. But if we have *independent evidence* that a relationship is real, then a low correlation does give some idea of its strength. In that case, a low correlation might suggest that there are other important factors that we have not considered.

Excel hint: the function for calculating correlation is CORREL(array1, array2), where array1 and array2 refer to the range of cells for two variables.

G.13.3 Two Important Properties of Correlations

Correlation is Symmetrical

The correlation of *X* with *Y* is exactly the same as the correlation of *Y* with *X*. Look at equation (G.33). If we were to switch *X* and *Y* everyplace they occurred, it is easy to see that the functions would all take the same values. This is easily proved in a spreadsheet as well Calculate the correlation between two series using, e.g., *Excel*'s CORREL function. Now reverse the order of the variables in the function, you should get exactly the same number.

The symmetry of correlation explains, in part, the old adage: correlation does not prove causation. A genuine correlation *does* suggest that variables are causally connected: either (a) one causes the other; or (b) there is some more complex causal connection between them, involving other variables. But a correlation can not tell us whether it is (a) or (b). And, even if it is (a), it cannot tell us which is cause and which is effect. Cause, in contrast to correlation, is an asymmetrical relationship: if *X* causes *Y*, then, in general, *Y* does not cause *X*. (Mutual causation is possible, but it is not the general case.)

Correlation is not Transistive

A relationship such as "larger than" is transitive. If *X* is larger than *Y*, and *Y* is larger than *Z*, then *X* is larger than *Z*. But correlation is not transitive. If *X* is correlated with *Y*, and *Y* is correlated with *Z*, then it does not follow that *X* is correlated with *Z*.

Example G.36. Consider two random sets of data chosen to have a very low correlation: *X* = {0.53, 0.14, 0.2, 0.75, 0.65, 0.22, 0.75, 0.08, 0.75, 0.48} and

Y = {0.8, 0.81, 0.83, 0.75, 0.2, 0.66, 0.57, 0.34, 0.7, 0.82}. A third variable is the sum of the two, $Z = X + Y$ = {1.33, 0.95, 1.03, 1.5, 0.85, 0.88, 1.32, 0.42, 1.45, 1.3}. Use these variables to show that correlation is not transitive.
Answer. It is easy to calculate the correlations is a spreadsheet. Cor(X,Z) = 0.76, which is reasonably high; cor(Z,Y) = 0.59, which is still moderate. So, clearly X is correlated with Z, and Z is correlated with Y. But cor(X,Y) = 0.06, which is very low; so, we can conclude that X is not correlated with Y. Correlation is intransitive.

G.14 Relationships Between Stationary and Nonstationary Time Series

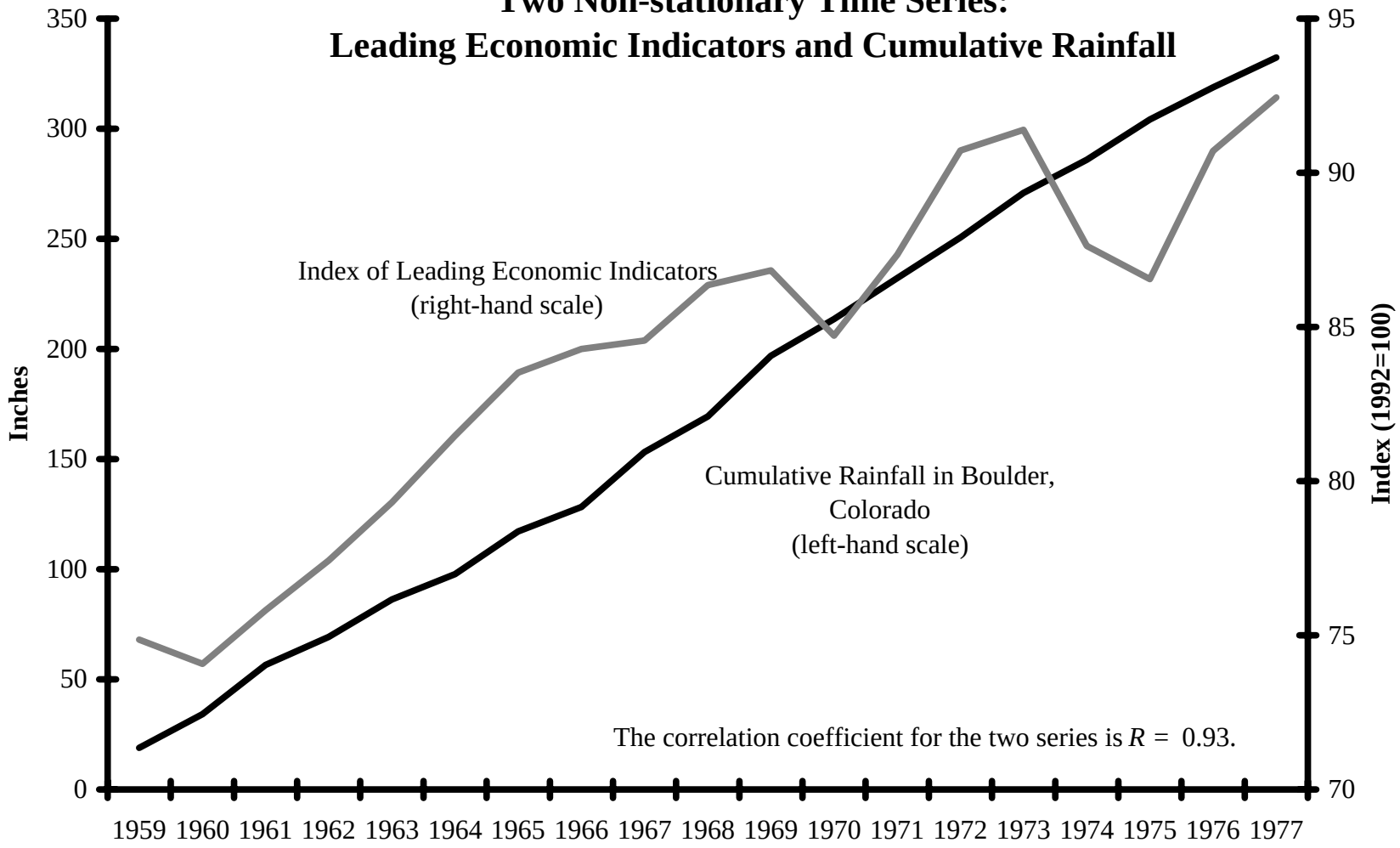
G.14.1 Nonsense Correlations

It is easy to misuse correlation and, as a result, to reach wrong conclusions about the reality and the strength of economic relationships. A simple demonstration illustrates the problem.

Figure G.11 plots the index of leading economic indicators and cumulative rainfall in Boulder, Colorado from 1959 to 1977. Both series trend upwards. The correlation coefficient, R = 0.93, is a very high. The great statistician George Udny Yule referred to such relationships as **nonsense correlations** defined as *a correlation that is high despite an absence of a genuine relationship between the variables*. The example is clearly a nonsense correlation as the history of rainfall in one town can hardly have any deep connection to the leading economic indicators. If they are nonsense, why do such correlations occur?

Notice that the correlation coefficient is based on the means, variances, and a covariance (equation (G.33)). We know from section G.5.2 that these are easily interpreted only for stationary time series. But both the index of leading economic indicators and cumulative rainfall are strongly trending series – that is, they are

Figure G.11
Two Non-stationary Time Series:
Leading Economic Indicators and Cumulative Rainfall



Source: National Weather Service; Conference Board.

nonstationary (they do not cross their own means frequently). The current value of an upwardly trending series is always above its mean, and its mean rises each period. The product terms for two upwardly trending series in the correlation formula (G.33) will both be positive, and so *R* will also necessarily be positive and will get closer to one over time. (If both are negatively trending, *R* will still be positive; if one is downwardly trending, *R* will be negative; but, in every case, *R* will get closer to one in absolute value over time.) These results say nothing for or against a genuine relationship. They occur simply because the data have trends.

The first lesson is: *do not calculate correlation coefficients for trending data; they are meaningless.*

G.14.2 Genuine Relationships Between Nonstationary Time Series

Given the problem of nonsense correlations, how do we distinguish genuine from spurious relationships between non-stationary time series? Relationships may exist in the short run or the long run (roughly the distinction between trend and cycle in Chapter 5). We take each case in turn.

Short-run Relationships

The idea is simple: if we first detrend each series, so that each is stationary, then it is appropriate to apply the correlation coefficient as a measure of their association and to interpret a high correlation as indicative of some (perhaps complicated) relationship.

Example G.37. Is there a genuine short-run relationship between the index of leading economic indicators and cumulative rainfall in Boulder, Colorado?
Answer. There are many ways to detrend (see section G.12). Figure G.12 shows one of them: the first differences of trending data in Figure G.11. These data look stationary. The correlation coefficient between the detrended series is $R = 0.03$, which is very low and suggests that there is no genuine relationship.

While the relationship in this example supports the characterization of nonsense correlation, the next example considers a more reasonable economic relationship.

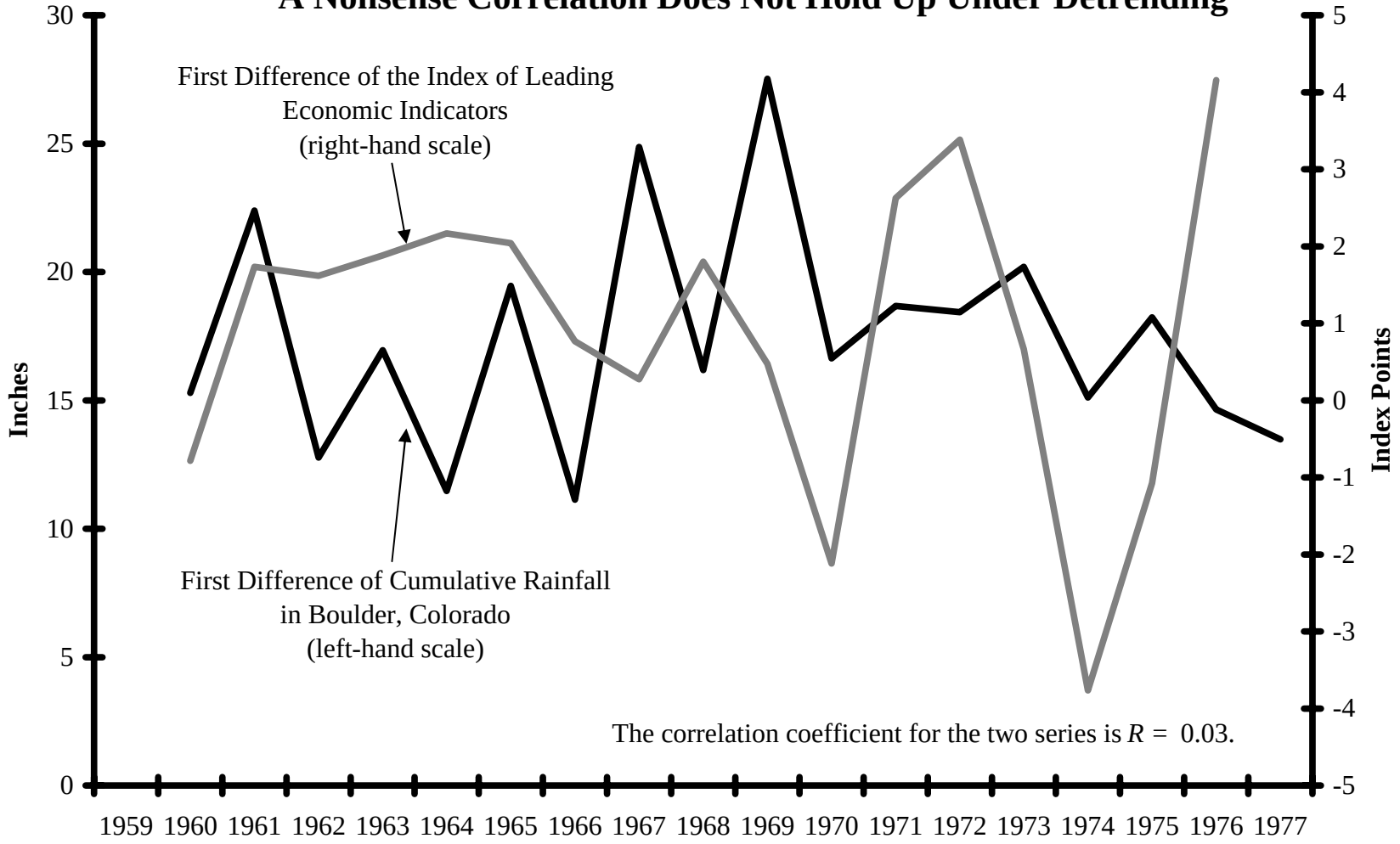
Example G.38. Is there a genuine short-run relationship between the Canadian employment and GDP per capita data in Tables G.1 and G.2?
Answer. Figure G.13 plots the growth rates of the data (see Figure G.3 for the levels). It is hard to make secure judgments of stationarity with short runs of data (here have only 9 periods of data). Nevertheless, these are certainly more stationary than the original data. The correlation coefficient of the two time series is $R = 0.94$, which suggests a genuine tendency to move directly together in the short run.

Long-run Relationships

Any time one transforms data information may also be lost. Just because trending series are vulnerable to nonsense correlations does not mean that all genuine relationships must show up in data transformed through trending or differencing. There may be important economic relationships between the levels of variables. For example, the main reason that consumption is higher in 2000 than in 1800 is that GDP is higher. This would be true whether or not the quarter-to-quarter or year-to-year change in consumption was highly correlated with the quarter-to-quarter or year-to-year change in GDP. The relationship between them is not a nonsense correlation.

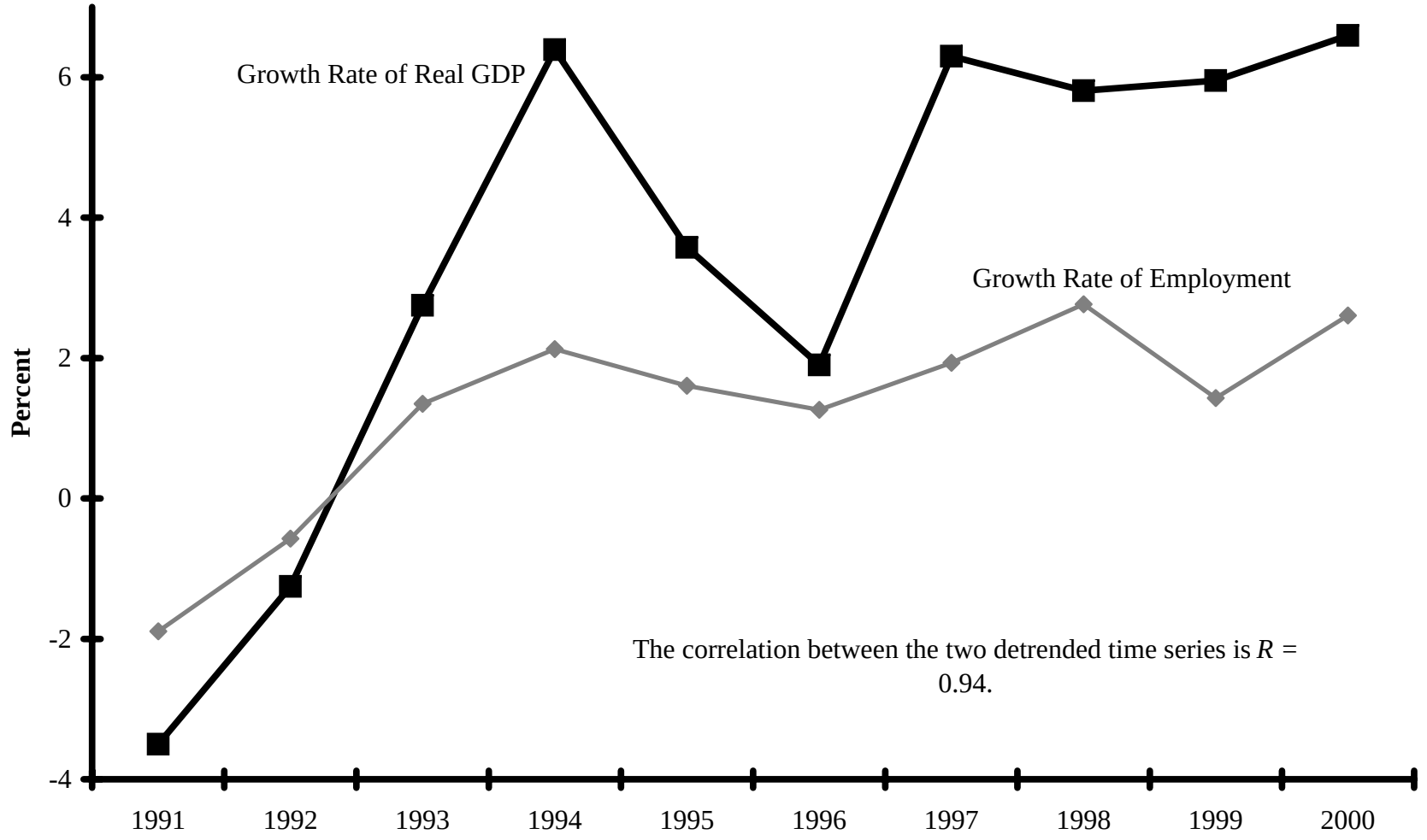
Sometimes the long-term relationship between two time series takes a simple form.

Figure G.12
A Nonsense Correlation Does Not Hold Up Under Detrending



Source: See Figure G.11.

Figure G.13
A Genuine Relationship Between Trending Time Series:
Canadian Real GDP and Employment



Source: International Monetary Fund, *Internatinal Financial Statistics.*

Example G.39. Is there a genuine long-term relationship between taxes and income in the United States?
Answer. Figure 12.5 shows that the average tax rate (i.e., the ratio of taxes to GDP) appears to be a stationary series around a mean of 16.9 percent. This suggests a genuine (and simple) long-term relationship.

What if we apply the same technique to a case of nonsense correlation?

Example G.40. Is there a genuine long term relationship between the index of leading economic indicators and cumulative rainfall in Boulder, Colorado?
Answer. Figure G.14 shows that the ratio of the two time series is itself a trending (non-stationary) series. The ratio provides no evidence in favor of genuine long-term relationship.

The last example reinforces our belief that the rainfall/leading indicator example is nonsense, but that is partly because we have other evidence to support that view. While a stable ratio suggests a genuine relationship, the failure to find a stable ratio could result either from a nonsense relationship or from a genuine, but more complicated relationship.

Example G.41. Is there a genuine long term relationship between Canadian employment and Canadian GDP per capita?
Answer. Figure G.15 shows the ratio of the GDP per capita to employment appears to be trending (non-stationary). Nevertheless, the relationship is likely a genuine one. The trend arises because it is more complicated involving labor productivity as well as the two series on the graph (see Chapter 7).

This example reinforces another important lesson: *statistical calculations alone rarely, if ever, tell us which relationships are genuine or not; they are part of the evidence, but must be applied with economic understanding, taking all the evidence into account.*

Figure G.14
The Ratio of Non-stationary Time Series that are not Genuinely Related is Itself Non-stationary

Source: See Figure G.11.

Figure G.15
The Ratio Between Trending Time Series that are Genuinely Related May Also Be Non-stationary:
Canadian Real GDP and Employment

Source: International Monetary Fund, *Internatinal Financial Statistics.*

G.14.3 DO NOT MIX STATIONARY AND NON-STATIONARY TIME SERIES.

Two stationary series may or may not have a genuine, simple relationship. The same is true for two non-stationary time series. But a stationary and a non-stationary time series certainly do not have a genuine, simple relationship. For example, there can be no genuine, simple relationship between the level of prices in the United States and the unemployment rate. The level of prices (Figure 2.9) has grown almost steadily since World War II, rising by more than 350 percent. And while the level of unemployment (Figure 9.2) may not be technically stationary (it appears to have long periods with higher or lower means), it clearly has kept within a much narrower bound than prices (minimum 2.5 percent; maximum 10.8 percent; mean 5.6 percent), so that it can for some purposes be thought of as stationary about that mean. Suppose that prices were a function of the unemployment rate: $p = f(U)$. Consider a point early in the sample (say, 1960) when the price level was 23.2 and the unemployment rate 5.5 percent. Because unemployment is stationary, it is likely that at some later time the unemployment rate will be at or near 5.5 percent again – as it was, for example, about 1996. The equation, of course, predicts that prices should also return to their earlier level. But of course, at that date, since prices are strongly trending upwards, they are in fact at the much higher level of 109.5. Clearly, no equation of this type can describe the relationship.

This is not to say that there is no relationship between prices and unemployment. In Chapter 16, we derive an important relationship between the change in the inflation rate (the growth rate of prices) and the unemployment rate. But in that case, prices have been transformed into an inflation rate, and the first difference of the inflation rate is

stationary, so we no longer have an impermissible relationship between a stationary and a non-stationary time series.

So the final lesson of this section is: *do not look for a genuine, simple relationship between a stationary and a nonstationary time series.*

G.15 Regression

Regression is discussed in Chapter 8, Box 8.1. In this section we fill in some details and consider two issues about the appropriate use of regression.

G.15.1 Linear Regression

When a computer program or spreadsheet calculates a regression, it finds the "best" values for the coefficients *a* and *b* of the equation

(G.35) $$Y_i = a + bX_i + error_i,$$

where the error terms cannot be observed, but must be estimated depending on the choice of the coefficients, and "best" is defined to mean the values that minimize the variance of the error terms. The formulae for these best estimates are:

(G.36)
$$b^* = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$
$$a^* = \overline{Y} - b^* \overline{X},$$

where the star on the coefficient indicates that it is an estimate.

As we shall see in the next section, regression and correlation are closely related. They are useful for different things. Correlation measures the strength of association between variables. Regression helps us to put a mathematical form on the association by giving us coefficient values that allow us to say how much one variable will change when another changes.

G.15.2 The Direction of Regression

Correlation is symmetrical (section G.13.3); regression is asymmetrical: in general, the regression of Y on X is not equivalent to the regression of X on Y. Consider an equation like (G.35) but with X and Y switched:

(G.37) $$X_i = c + dY_i + error_i.$$

If the error terms were zero, then a little algebra shows that $d = 1/b$. Ordinary equations are symmetrical. But when the error terms are not zero, this is no longer true. The regression estimate of d is

(G.38) $$d^* = \frac{\text{cov}(X, Y)}{\text{var}(Y)}.$$

Obviously, $d^* \neq 1/b^*$.

To visualize this result, Figure G.16 shows the same scatterplot and regression line as in Figure B8.1 in Chapter 8 (Box 8.1). That line reports the regression of labor (l) on the real wage (w/p). In addition, Figure G.16, also shows the reverse regression of the real wage on labor. The two lines are clearly different.

Unlike correlation, regression has a direction. How should we choose which way to run a regression? The rule is simple in principle: *regression equations should be written with causes on the right-hand side (the horizontal axis of a scatterplot) and effects on the left-hand side (the vertical axis).* In that case, the error terms have the interpretation of unobserved (and perhaps unobservable) causes of the dependent variable. This rule is not always easy to apply in practice, because we do not always have good evidence – or intuitions – for which is cause and which is effect. Here, as in other cases, more sophisticated statistics than appropriate to an intermediate course might be helpful.

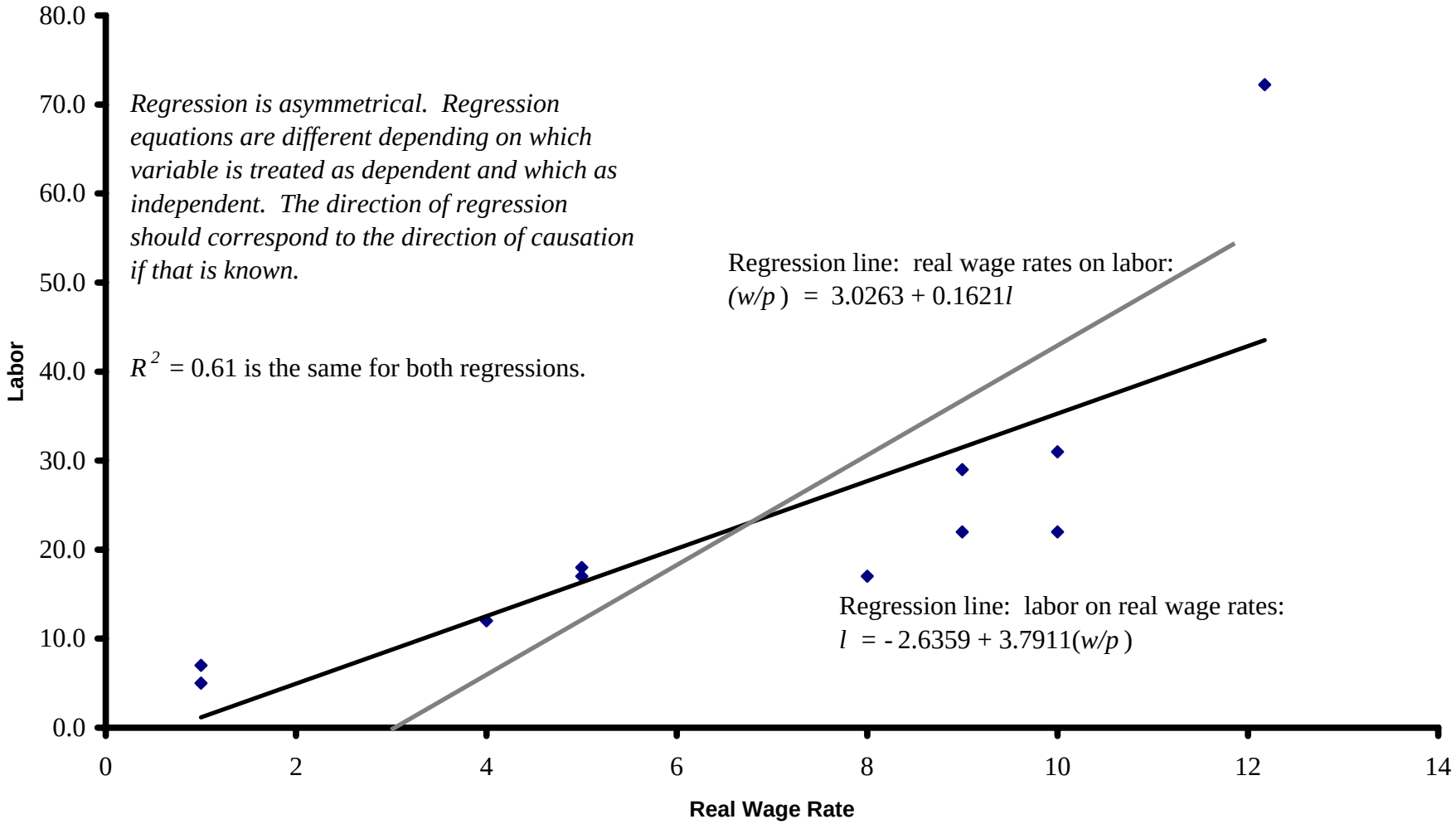
G.15.3 NONSENSE REGRESSION

Regression and correlation are closely connected. Consider the product of b^* and d^* from equations (G.36) and (G.38):

$$b^* d^* = \frac{\text{cov}(X,Y)}{\text{var}(X)} \frac{\text{cov}(X,Y)}{\text{var}(Y)} = \frac{(\text{cov}(X,Y))^2}{\text{var}(X)\,\text{var}(Y)} = R^2 . \tag{G.39}$$

The last step follows from the definition of the correlation coefficient (R) in equation (G.34). The higher the correlation between two variables, the better fitting the regression.

Figure G.16
The Direction of Regression



An important lesson to be drawn from this close connection is that just as we must guard against drawing conclusions from nonsense correlations (section G.14), we must also be on guard for **nonsense regressions**. The regression of one trending variable on another will have a high R^2, but cannot be easily interpreted as expressing a genuine economic connection. Just as with correlations, we must transform data into stationary forms through detrending, differencing, taking growth rates, or ratios before estimating regression equations. And, just as it makes no sense to posit a simple, genuine relationship between a stationary and a non-stationary variable, it makes no sense typically to estimate the regression of a stationary variable on a non-stationary variable or vice versa.

G.16 The Cobb-Douglas Production Function

G.16.1 The Mathematics of the Cobb-Douglas Production Function

The Cobb-Douglas Production Function can be written as

$$Y = AL^{\alpha} K^{1-\alpha}, \tag{G.40}$$

where $0 < \alpha < 1$.

It follows immediately from the formula that if either input (L or K) is zero, output is zero. Graphically, this shows that the Cobb-Douglas production function goes through the origin.

The marginal product of labor is the small extra output that is produced from a small increase in the labor input, holding other inputs constant. Expressed in the terminology of the calculus, it is the derivative, *dY/dL*. Since *Y* is a function of more than one variable, this is properly a problem in multivariate calculus. We can, however, treat the other variables just as we would constant terms, so that we act as if *Y* were a function of *L* only, and then apply the power rule of ordinary calculus:

$$\begin{aligned} mpL = dY/dL &= \alpha A L^{\alpha-1} K^{1-\alpha} = \alpha A L^{\alpha} L^{-1} K^{1-\alpha} \\ &= \alpha (A L^{\alpha} K^{1-\alpha}) / L = \alpha Y / L. \end{aligned} \tag{G.41}$$

Another useful way to write this would be

$$mpL = dY/dL = \alpha A L^{\alpha-1} K^{1-\alpha} = \alpha A K^{1-\alpha} / L^{1-\alpha} = \alpha A (K/L)^{1-\alpha} \tag{G.42}$$

Similarly, the marginal product of capital is

$$\begin{aligned} mpK = dY/dK &= (1-\alpha) A L^{\alpha} K^{(1-\alpha)-1} = (1-\alpha) A L^{\alpha} K^{1-\alpha} K^{-1} \\ &= (1-\alpha) A L^{\alpha} K^{1-\alpha} / K = (1-\alpha) Y / K. \end{aligned} \tag{G.43}$$

And, again, this can be written as

$$mpK = dY/dK = (1-\alpha)AL^{\alpha}K^{(1-\alpha)-1} = (1-\alpha)AL^{\alpha}K^{-\alpha}$$

(G.44)

$$= (1-\alpha)A(L/K)^{\alpha}.$$

Since inputs cannot be negative, except in the case that one or both is zero, the marginal products are necessarily positive. Graphically, this means that the Cobb-Douglas production function slopes up.

Looking at the right-most term of equation (G.42), it is obvious that as *L* becomes larger, *mpL* becomes smaller. In other words, the Cobb-Douglas production function shows *diminishing returns to labor*. Diminishing returns can be demonstrated mathematically by taking the derivative of *mpL*, which shows the rate at which it changes as *L* increases. The first derivative of *mpL* is the same as the second derivative of *Y*. Using the third term in equation (G.42) as the expression for the *mpL*:

(G.45) $$d(mpL)/dl = d^2Y/dL^2 = d(\alpha AL^{\alpha-1}K^{1-\alpha})/dL = (\alpha-1)\alpha AL^{\alpha-2}K^{1-\alpha}$$

Since $\alpha < 1$, the first term on the right-hand side is negative. The rest of the expression is positive, so that the whole expression is negative. Equation (G.45), therefore, says that the marginal product of labor declines as *L* increases. Graphically, the production function is concave toward the labor axis.

Similar results can be derived for capital through the same procedures making appropriate modifications.

Example G.42. Using the data from Chapter 6, section 6.2.4 for 1998: $\alpha = 0.69$, capital is \$19,317 billion, labor is 236,822 million worker-hours per year, and real GDP is \$7,552. What are the marginal products of labor and capital? Prove that they are diminishing.

Answer. Using equation (G.41) $mpL = \alpha(Y/L) = 0.69(7{,}552 \times 10^9)/(236{,}882 \times 10^6)$ = \$22/hour; using equation (G.43) $mpK = (1 - 0.69)(7{,}552 \times 10^9)/(19{,}317 \times 10^9) =$ 0.12 dollars of output/dollar of capital. The value of A can be computed from the other information as in equation (5.15') to be 8.1469. Using equation (G.45),

$$d(mpL)/dl = (\alpha - 1)\alpha AL^{\alpha-2}K^{1-\alpha}$$
$$= (0.69 - 1)0.69(8.1469)(236{,}882 \times 10^6)^{(0.69-2)}(19{,}317 \times 10^9)^{(1-0.69)}$$
$$= -2.88 \times 10^{-11} < 0;$$

the equivalent expression for capital would be

$$d(mpK)/dl = -\alpha(1-\alpha)AL^{\alpha}K^{-\alpha-1}$$
$$= -0.69(1 - 0.69)(8.1469)(236{,}882 \times 10^6)^{0.69}(19{,}317 \times 10^9)^{(-0.69-1)}$$
$$= -4.33 \times 10^{-15} < 0.$$

Since the derivatives of each of the marginal products is negative, each is getting smaller as the input gets larger – that is, returns are diminishing.

Suppose that we increase the inputs of the two factors of production by the same proportion, call it λ. How much will output increase? Replacing K and L in the right-hand side of equation (G.40) with λK and λL gives us:

$$\text{(G.46)} \qquad A(\lambda L)^{\alpha}(\lambda K)^{1-\alpha} = A\lambda^{\alpha}L^{\alpha}\lambda^{1-\alpha}K^{1-\alpha} = \lambda^{\alpha+(1-\alpha)}AL^{\alpha}K^{1-\alpha} = \lambda Y$$

In other words an equiproportional increase in the two factors of production results in an equiproportional increase in output. Mathematicians refer to this property as *homogeneity of degree one*. Economists refer to it in this context as **constant returns to scale.**

G.16.2 ESTIMATING LABOR'S SHARE (α) OF OUTPUT FROM DATA

Labor's share in GDP is, in principle, just the income of workers divided by GDP. A problem arises because the national-income-and-product accounts recognize a variety of classes of income. Wages and salaries are clearly labor income; while profits, interest, and dividends are clearly capital income. But what about proprietors' income. Proprietors are the owners of businesses in which they are, themselves, also among the principal workers, so that proprietor's income is a mixture of labor and capital income. To distribute proprietors' income, we assume that it can contains the same proportion of labor income as the rest of GDP.

The labor share in GDP, excluding proprietors' income is

$$s = \frac{compensation\ of\ \ employees}{GDP - proprietors'\ income - indirect\ \ business\ tax\ \ and\ \ \ nontax\ \ liability}.$$

The term *indirect business tax and nontax liability* is needed because proprietorships, like other businesses, pay taxes at a stage of production before their incomes are reported.

The labor share in GDP, *including imputed proprietors' income*, is

$$\alpha = \frac{compensation\ of\ \ employees + s(proprietors'\ income + indirect\ \ business\ tax\ \ and\ \ \ nontax\ \ liability}{GDP}$$

The share α can be computed each quarter. The value used in the text (0.70) is the average value over the period 1946:1 to 2003:4.

G.17 Financial Tables

G.17.1 Reading Stock Tables

Table G.8 reproduces typical entries from the stock market tables published daily in the *Wall Street Journal* with an explanation of the various entries.

Table G.8. Typical Stock Tables

				D					
YTD	52-WEEK				YLD		VOL		NET
%CHG	HI	LO	STOCK (SYM)	DIV	%	PE	100S	CLOSE	CHG
–13.9	18.81	14.21	DIRECTV **DTV**		…	dd	21303	14.41	–0.8
–0.2	29.99	20.88	DISNEY **DIS**	.24	.9	24	46276	27.75	–0.03

Source: *Wall Street Journal* 25 March 2005, p. C5. Entries refer to the New York Stock Exchange on 24 March 2005.

Key

STOCK = abbreviated name of corporation issuing the stock.

SYM = stock-market ticker symbol for the corporation issuing the stock.

YTD %CHG = percentage change in the price of the stock from 1 January to the current date.

52-WEEK HI and LO = the highest and lowest price (in dollars per share) at which the stock sold over the past year.

DIV = the dividend or other distributions paid on the stock at an annual rate based on the most recent corporate declaration. (A blank means that no dividend was paid.)

YLD % = the dividend or other distributions as a percentage of the price of the share.

PE = the price-earnings ratio calculated by dividing the closing price by the earnings per share for the most recent four quarters.

VOL 100S = the number of shares traded on the reported date in less two zeroes (e.g., Disney traded 4,627,600 traded on 24 March 2005).

CLOSE = the price (in dollars per share) of the last trade in the listed share on the reported date.

NET CHG = the change in the closing price on the date reported from the last date the market was open (in dollars per share).

Footnotes: stock market tables typically use a large number of footnotes to indicate various details not easily listed in the table. E.g., the footnote dd under PE for DirecTV indicates that the company made a loss in the last four quarters, so that no positive earnings are available to compute the price-earnings ratio.

G.17.2 READING BOND TABLES

Table G.9 reproduces entries from four distinct U.S. Government bond tables published daily in the Wall Street Journal followed by explanations of the various entries.

Table G.9. Typical Government Bond Tables

Treasury Bonds, Notes and Bills

RATE	MATURITY MO/YR	BID	ASKED	CHG	ASK YLD
Government Bonds & Notes					
6.500	Aug 05n	101:10	101:11	–1	2.92
4.250	Feb 14n	97:04	97:05	3	4.62
6.250	May 30	118:28	118:29	11	4.93
3.375	Apr 32i	130:02	130:03	13	1.94

U.S. Treasury Strips					
MATURITY	TYPE	BID	ASKED	CHG	ASK YLD
Feb 06	ci	97:01	97:01	1	3.43
Feb 06	bp	97:02	97:02	1	3.38
Feb 06	np	97:02	97:02	1	3.39

Inflation Indexed Securities					
RATE	MAT	BID/ASKED	CHG	*YLD	ACCR PRIN
3.375	01/07	104-25/26	. . .	0.687	1202
2.000	07/14	101-08/09	5	1.849	1011

Treasury Bills					
MATURITY	DAYS TO MAT	BID	ASKED	CHG	ASK YLD
Apr 07 05	13	2.40	2.39	-0.11	2.42
Sep 22 05	181	3.06	3.05	0.01	3.14

Source: *Wall Street Journal* 25 March 2005, p. C11. Entries refer to 24 March 2005.

Key

General

RATE = the coupon rate expressed as a percentage of face value.

MATURITY (or MAT) = date at which the security matures (for bonds and notes, it is given as month and year (MO/YR) either in the format Aug 05 or 01/07 and for bills as month, day, and year.

BID and ASKED = the price offered to buy (BID) or sell (ASKED). For most securities, prices are quoted as dollars per $100 of face value in dollars and 32nd's (e.g., 97:02 means $97 2/32 per $100 face value; 104-25 means $104 25/32 per $100 face value). Prices of Treasury bills are quoted as percentage discounts from face value (e.g., an asked price of 2.39 means that a $10,000 Treasury bill is offered for sale at a price of $9,761, which is 2.39 percent less than $10,000.)

CHG = for most securities, the change in price in dollars from the close of the previous day's business. For Treasury bills, the change in the percentage discount from the previous day's business.

ASK YLD = the yield to maturity based on the asked price.

Treasury Bonds, Notes, and Bills

Treasury bonds are coupon securities originally issued with a maturity of greater than 10 years; notes are coupon securities issued with a maturity of between 1 and 10 years. A suffix n on the maturity date in the table indicates a Treasury note. Bills are pure discount securities issued with a maturity of less than 1 year.

U.S. Treasury Strips

A U.S. Treasury Strip is the repackage interest or principal of a Treasury security sold separately. The entries under TYPE indicate which part of the security is being sold.

ci = coupon interest (in effect, a pure coupon bond);
np = the principal from a Treasury note (in effect, a pure discount bond);
bp = the principal from a Treasury bond (in effect, a pure discount bond).

Inflation Indexed Securities

Inflation-indexed securities are Treasury securities whose principal (originally quoted as $1,000) is increased periodically in line with changes in the consumer price index. When the security pays off it pays both its originally quoted face value and an extra increment to compensate for inflation. Inflation-indexed securities are typically quoted both in their own section of the Treasury securities table and in the section devoted to Bonds, Notes, and Bills.

ACCR PRIN = accrued principal is the face value plus the inflation adjustment from the date of issue to the current date.

*YLD= the yield to maturity based on the accrued principal as the face value. The yield to maturity on an inflation-indexed security can be interpreted as a real rate of interest, since it is the yield after compensation for inflation.

Treasury Bills

DAYS TO MAT = the exact number of days until the bill matures.

Suggestions for Further Reading

B.S. Everitt, *The Cambridge Dictionary of Statistics*, Cambridge: Cambridge University Press, 1998.

Darrell Huff, *How to Lie With Statistics*. New York: Norton, 1993.

Bernard Lindgren, *Statistical Theory*, 3rd ed., New York: Macmillan, 1976.

Edward Tufte, *The Visual Display of Quantitative Information*, 2nd edition. Chesire, CT: Graphics Press, 2001.

Edward Tufte, *Visual Explanations: Images and Quantities, Evidence, and Narrative*. Chesire, CT: Graphics Press, 1997.

Edward Tufte, *Envisioning Information*. Chesire, CT: Graphics Press, 1990.