STATISTICS FOR ECONOMISTS: A BEGINNING

John E. Floyd
University of Toronto

July 2, 2010

PREFACE

The pages that follow contain the material presented in my introductory quantitative methods in economics class at the University of Toronto. They are designed to be used along with any reasonable statistics textbook. The most recent textbook for the course was James T. McClave, P. George Benson and Terry Sincich, *Statistics for Business and Economics*, Eighth Edition, Prentice Hall, 2001. The material draws upon earlier editions of that book as well as upon John Neter, William Wasserman and G. A. Whitmore, *Applied Statistics*, Fourth Edition, Allyn and Bacon, 1993, which was used previously and is now out of print. It is also consistent with Gerald Keller and Brian Warrack, *Statistics for Management and Economics*, Fifth Edition, Duxbury, 2000, which is the textbook used recently on the St. George Campus of the University of Toronto. The problems at the ends of the chapters are questions from mid-term and final exams at both the St. George and Mississauga campuses of the University of Toronto. They were set by Gordon Anderson, Lee Bailey, Greg Jump, Victor Yu and others including myself.

This manuscript should be useful for economics and business students enrolled in basic courses in statistics and, as well, for people who have studied statistics some time ago and need a review of what they are supposed to have learned. Indeed, one could learn statistics from scratch using this material alone, although those trying to do so may find the presentation somewhat compact, requiring slow and careful reading and thought as one goes along.

I would like to thank the above mentioned colleagues and, in addition, Adonis Yatchew, for helpful discussions over the years, and John Maheu for helping me clarify a number of points. I would especially like to thank Gordon Anderson, who I have bothered so frequently with questions that he deserves the status of mentor.

After the original version of this manuscript was completed, I received some detailed comments on Chapter 8 from Peter Westfall of Texas Tech University, enabling me to correct a number of errors. Such comments are much appreciated.

J. E. Floyd
July 2, 2010

Contents

Chapter 1

Introduction to Statistics, Data and Statistical Thinking

1.1 What is Statistics?

In common usage people think of statistics as numerical data—the unemployment rate last month, total government expenditure last year, the number of impaired drivers charged during the recent holiday season, the crime-rates of cities, and so forth. Although there is nothing wrong with viewing statistics in this way, we are going to take a deeper approach. We will view statistics the way professional statisticians view it—as a methodology for collecting, classifying, summarizing, organizing, presenting, analyzing and interpreting numerical information.

1.2 The Use of Statistics in Economics and Other Social Sciences

Businesses use statistical methodology and thinking to make decisions about which products to produce, how much to spend advertising them, how to evaluate their employees, how often to service their machinery and equipment, how large their inventories should be, and nearly every aspect of running their operations. The motivation for using statistics in the study of economics and other social sciences is somewhat different. The object of the social sciences and of economics in particular is to understand how

the social and economic system functions. While our approach to statistics will concentrate on its uses in the study of economics, you will also learn business uses of statistics because many of the exercises in your textbook, and some of the ones used here, will focus on business problems.

Views and understandings of how things work are called *theories*. Economic theories are descriptions and interpretations of how the economic system functions. They are composed of two parts—a logical structure which is tautological (that is, true by definition), and a set of parameters in that logical structure which gives the theory empirical content (that is, an ability to be consistent or inconsistent with facts or data). The logical structure, being true by definition, is uninteresting except insofar as it enables us to construct testable propositions about how the economic system works. If the facts turn out to be consistent with the testable implications of the theory, then we accept the theory as true until new evidence inconsistent with it is uncovered. A theory is valuable if it is logically consistent both within itself and with other theories established as "true" and is capable of being rejected by but nevertheless consistent with available evidence. Its logical structure is judged on two grounds—internal consistency and usefulness as a framework for generating empirically testable propositions.

To illustrate this, consider the statement: "People maximize utility." This statement is true by definition—behaviour is defined as what people do (including nothing) and utility is defined as what people maximize when they choose to do one thing rather than something else. These definitions and the associated utility maximizing approach form a useful logical structure for generating empirically testable propositions. One can choose the parameters in this tautological utility maximization structure so that the marginal utility of a good declines relative to the marginal utility of other goods as the quantity of that good consumed increases relative to the quantities of other goods consumed. Downward sloping demand curves emerge, leading to the empirically testable statement: "Demand curves slope downward." This *theory of demand* (which consists of both the utility maximization structure and the proposition about how the individual's marginal utilities behave) can then be either supported or falsified by examining data on prices and quantities and incomes for groups of individuals and commodities. The set of tautologies derived using the concept of utility maximization are valuable because they are internally consistent and generate empirically testable propositions such as those represented by the theory of demand. If it didn't yield testable propositions about the real world, the logical structure of utility maximization would be of little interest.

Alternatively, consider the statement: "Canada is a wonderful country."

This is not a testable proposition unless we define what we mean by the adjective "wonderful". If we mean by wonderful that Canadians have more flush toilets per capita than every country on the African Continent then this is a testable proposition. But an analytical structure built around the statement that Canada is a wonderful country is not very useful because empirically testable propositions generated by redefining the word wonderful can be more appropriately derived from some other logical structure, such as one generated using a concept of real income.

Finally, consider the statement: "The rich are getting richer and the poor poorer." This is clearly an empirically testable proposition for reasonable definitions of what we mean by "rich" and "poor". It is really an interesting proposition, however, only in conjunction with some theory of how the economic system functions in generating income and distributing it among people. Such a theory would usually carry with it some implications as to how the institutions within the economic system could be changed to prevent income inequalities from increasing. And thinking about these implications forces us to analyse the consequences of reducing income inequality and to form an opinion as to whether or not it should be reduced.

Statistics is the methodology that we use to confront theories like the theory of demand and other testable propositions with the facts. It is the set of procedures and intellectual processes by which we decide whether or not to accept a theory as true—the process by which we decide what and what not to believe. In this sense, statistics is at the root of all human knowledge.

Unlike the logical propositions contained in them, theories are never strictly true. They are merely accepted as true in the sense of being consistent with the evidence available at a particular point in time and more or less strongly accepted depending on how consistent they are with that evidence. Given the degree of consistency of a theory with the evidence, it may or may not be appropriate for governments and individuals to act as though it were true. A crucial issue will be the costs of acting as if a theory is true when it turns out to be false as opposed to the costs of acting as though the theory were not true when it in fact is. As evidence against a theory accumulates, it is eventually rejected in favour of other "better" theories—that is, ones more consistent with available evidence.

Statistics, being the set of analytical tools used to test theories, is thus an essential part of the scientific process. Theories are suggested either by casual observation or as logical consequences of some analytical structure that can be given empirical content. Statistics is the systematic investigation of the correspondence of these theories with the real world. This leads either

to a wider belief in the 'truth' of a particular theory or to its rejection as inconsistent with the facts.

Designing public policy is a complicated exercise because it is almost always the case that some members of the community gain and others lose from any policy that can be adopted. Advocacy groups develop that have special interests in demonstrating that particular policy actions in their interest are also in the public interest. These special interest groups often misuse statistical concepts in presenting their arguments. An understanding of how to think about, evaluate and draw conclusions from data is thus essential for sorting out the conflicting claims of farmers, consumers, environmentalists, labour unions, and the other participants in debates on policy issues.

Business problems differ from public policy problems in the important respect that all participants in their solution can point to a particular measurable goal—maximizing the profits of the enterprise. Though the individuals working in an enterprise maximize their own utility, and not the objective of the enterprise, in the same way as individuals pursue their own goals and not those of society, the ultimate decision maker in charge, whose job depends on the profits of the firm, has every reason to be objective in evaluating information relevant to maximizing those profits.

1.3 Descriptive and Inferential Statistics

The application of statistical thinking involves two sets of processes. First, there is the description and presentation of data. Second, there is the process of using the data to make some inference about features of the environment from which the data were selected or about the underlying mechanism that generated the data, such as the ongoing functioning of the economy or the accounting system or production line in a business firm. The first is called *descriptive statistics* and the second *inferential statistics*.

Descriptive statistics utilizes numerical and graphical methods to find patterns in the data, to summarize the information it reveals and to present that information in a meaningful way. Inferential statistics uses data to make estimates, decisions, predictions, or other generalizations about the environment from which the data were obtained.

Everything we will say about descriptive statistics is presented in the remainder of this chapter. The rest of the book will concentrate entirely on statistical inference. Before turning to the tools of descriptive statistics, however, it is worth while to take a brief glimpse at the nature of statistical

inference.

1.4 A Quick Glimpse at Statistical Inference

Statistical inference essentially involves the attempt to acquire information about a *population* or *process* by analyzing a *sample* of elements from that population or process.

A population includes the set of units—usually people, objects, transactions, or events—that we are interested in learning about. For example, we could be interested in the effects of schooling on earnings in later life, in which case the relevant population would be all people working. Or we could be interested in how people will vote in the next municipal election in which case the relevant population will be all voters in the municipality. Or a business might be interested in the nature of bad loans, in which case the relevant population will be the entire set of bad loans on the books at a particular date.

A process is a mechanism that produces output. For example, a business would be interested in the items coming off a particular assembly line that are defective, in which case the process is the flow of production off the assembly line. An economist might be interested in how the unemployment rate varies with changes in monetary and fiscal policy. Here, the process is the flow of new hires and lay-offs as the economic system grinds along from year to year. Or we might be interested in the effects of drinking on driving, in which case the underlying process is the on-going generation of car accidents as the society goes about its activities. Note that a process is simply a mechanism which, if it remains intact, eventually produces an infinite population. All voters, all workers and all bad loans on the books can be counted and listed. But the totality of accidents being generated by drinking and driving or of steel ingots being produced from a blast furnace cannot be counted because these processes in their present form can be thought of as going on forever. The fact that we can count the number of accidents in a given year, and the number of steel ingots produced by a blast furnace in a given week suggests that we can work with finite populations resulting from processes. So whether we think of the items of interest in a particular case as a finite population or the infinite population generated by a perpetuation of the current state of a process depends on what we want to find out. If we are interested in the proportion of accidents caused by drunk driving in the past year, the population is the total number of accidents that year. If we are interested in the effects of drinking on driving, it is the

infinite population of accidents resulting from a perpetual continuance of the current process of accident generation that concerns us.

A sample is a subset of the units comprising a finite or infinite population. Because it is costly to examine most finite populations of interest, and impossible to examine the entire output of a process, statisticians use samples from populations and processes to make inferences about their characteristics. Obviously, our ability to make correct inferences about a finite or infinite population based on a sample of elements from it depends on the sample being *representative* of the population. So the manner in which a sample is selected from a population is of extreme importance. A classic example of the importance of representative sampling occurred in the 1948 presidential election in the United States. The Democratic incumbent, Harry Truman, was being challenged by Republican Governor Thomas Dewey of New York. The polls predicted Dewey to be the winner but Truman in fact won. To obtain their samples, the pollsters telephoned people at random, forgetting to take into account that people too poor to own telephones also vote. Since poor people tended to vote for the Democratic Party, a sufficient fraction of Truman supporters were left out of the samples to make those samples unrepresentative of the population. As a result, inferences about the proportion of the population that would vote for Truman based on the proportion of those sampled intending to vote for Truman were incorrect.

Finally, when we make inferences about the characteristics of a finite or infinite population based on a sample, we need some measure of the reliability of our method of inference. What are the odds that we could be wrong. We need not only a prediction as to the characteristic of the population of interest (for example, the proportion by which the salaries of college graduates exceed the salaries of those that did not go to college) but some quantitative measure of the degree of uncertainty associated with our inference. The results of opinion polls predicting elections are frequently stated as being reliable within three percentage points, nineteen times out of twenty. In due course you will learn what that statement means. But first we must examine the techniques of descriptive statistics.

1.5 Data Sets

There are three general kinds of data sets—*cross-sectional*, *time-series* and *panel*. And within data sets there are two kinds of data—*quantitative* and *qualitative*. Quantitative data can be recorded on a natural numerical scale. Examples are gross national product (measured in dollars) and the consumer price index (measured as a percentage of a base level). Qualitative data cannot be measured on a naturally occurring numerical scale but can only be classified into one of a group of categories. An example is a series of records of whether or not the automobile accidents occurring over a given period resulted in criminal charges—the entries are simply yes or no.

Table 1.1: Highest College Degree of Twenty Best-Paid Executives

Rank	Degree	Rank	Degree
1	Bachelors	11	Masters
2	Bachelors	12	Bachelors
3	Doctorate	13	Masters
4	None	14	Masters
5	Bachelors	15	Bachelors
6	Doctorate	16	Doctorate
7	None	17	Masters
8	Bachelors	18	Doctorate
9	Bachelors	19	Bachelors
10	Bachelors	20	Masters

Source: *Forbes*, Vol. 155, No. 11, May 22, 1995.

Table 1.1 presents a purely qualitative data set. It gives the highest degree obtained by the twenty highest-paid executives in the United States at a particular time. Educational attainment is a qualitative, not quantitative, variable. It falls into one of four categories: None, Bachelors, Masters, or Doctorate. To organize this information in a meaningful fashion, we need to construct a summary of the sort shown in Table 1.2. The entries in this table were obtained by counting the elements in the various categories in Table 1.1—for larger data sets you can use the spreadsheet program on your computer to do the counting. A fancy bar or pie chart portraying the information in Table 1.2 could also be made, but it adds little to what can be

Table 1.2: Summary of Table 1.1

Class (Highest Degree)	Frequency (Number of Executives)	Relative Frequency (Proportion of Total)
None	2	0.1
Bachelors	9	0.45
Masters	5	0.25
Doctorate	4	0.2
Total	20	1.0

Source: See Table 1.1

gleaned by looking at the table itself. A bachelors degree was the most commonly held final degree, applying in forty-five percent of the cases, followed in order by a masters degree, a doctorate and no degree at all.

The data set on wages in a particular firm in Table 1.3 contains both quantitative and qualitative data. Data are presented for fifty employees, numbered from 1 to 50. Each employee represents an *element* of the data set. For each element there is an *observation* containing two *data points*, the individual's weekly wage in U.S. dollars and gender (male or female). Wage and gender are *variables*, defined as characteristics of the elements of a data set that vary from element to element. Wage is a quantitative variable and gender is a qualitative variable.

As it stands, Table 1.3 is an organised jumble of numbers. To extract the information these data contain we need to enter them into our spreadsheet program and sort them by wage. We do this here without preserving the identities of the individual elements, renumbering them starting at 1 for the lowest wage and ending at 50 for the highest wage. The result appears in Table 1.4. The lowest wage is \$125 per week and the highest is \$2033 per week. The difference between these, \$2033 − \$125 = \$1908, is referred to as the variable's *range*. The middle observation in the range is called the *median*. When the middle of the range falls in between two observations, as it does in Table 1.4, we represent the median by the average of the two observations, in this case \$521.50. Because half of the observations on the variable are below the median and half are above, the median is called the *50th percentile*. Similarly, we can calculate other percentiles of the variable—90 percent of the observations will be below the 90th percentile and 80 percent will be below the 80th percentile, and so on. Of particular

Table 1.3: Weekly Wages of Company Employees in U.S. Dollars

No.	Wage	Gender	No.	Wage	Gender
1	236	F	26	334	F
2	573	M	27	600	F
3	660	F	28	592	M
4	1005	M	29	728	M
5	513	M	30	125	F
6	188	F	31	401	F
7	252	F	32	759	F
8	200	F	33	1342	M
9	469	F	34	324	F
10	191	F	35	337	F
11	675	M	36	1406	M
12	392	F	37	530	M
13	346	F	38	644	M
14	264	F	39	776	F
15	363	F	40	440	F
16	344	F	41	548	F
17	949	M	42	751	F
18	490	M	43	618	F
19	745	F	44	822	M
20	2033	M	45	437	F
21	391	F	46	293	F
22	179	F	47	995	M
23	1629	M	48	446	F
24	552	F	49	1432	M
25	144	F	50	901	F

Table 1.4: Weekly Wages of Company Employees in U.S. Dollars: Sorted into Ascending Order

No.	Wage	Gender		
1	125	F		
2	144	F		
3	179	F		
4	188	F		
5	...	...		
...				
11	324	F		
12	334	F		
13	337	F		
			340.5	1st (Lower) Quartile
14	344	F		(25th Percentile)
15	346	F		
16	...	...		
...				
23	469	F		
24	490	M		
25	513	M		
			521.50	Median
26	530	M		(50th Percentile)
27	548	F		
28	552	F		
29	...	...		
...				
35	675	M		
36	728	M		
37	745	F		
			748	3rd (Upper) Quartile
38	751	F		(75th Percentile)
39	759	F		
40	776	F		
41	...	...		
...				
48	1432	M		
49	1629	M		
50	2033	M		

interest are the 25th and 75th percentiles. These are called the *first quartile* and *third quartile* respectively. The difference between the observations for these quartiles, $\$748 - \$340.5 = \$407.5$, is called the *interquartile range*. So the wage variable has a median (mid-point) of \$521.50, a range of \$1908 and an interquartile range of \$407.5, with highest and lowest values being \$2033 and \$125 respectively. A quick way of getting a general grasp of the "shape" of this data set is to express it graphically as a histogram, as is done in the bottom panel of Figure 1.1.

An obvious matter of interest is whether men are being paid higher wages than women. We can address this by sorting the data in Table 1.3 into two separate data sets, one for males and one for females. Then we can find the range, the median, and the interquartile range for the wage variable in each of the two data sets and compare them. Rather than present new tables together with the relevant calculations at this point, we can construct histograms for the wage variable in the two separate data sets. These are shown in the top two panels of Figure 1.1. It is easy to see from comparing horizontal scales of the top and middle histograms that the wages of women tend to be lower than those paid to men.

A somewhat neater way of characterising these data graphically is to use box plots. This is done in Figure 1.2. Different statistical computer packages present box plots in different ways. In the one used here, the top and bottom edges of the box give the upper and lower quartiles and the horizontal line through the middle of the box gives the median. The vertical lines, called whiskers, extend up to the maximum value of the variable and down to the minimum value.[1] It is again obvious from the two side-by-side box plots that women are paid less than men in the firm to which the data set applies. So you can now tell your friends that there is substantial evidence that women get paid less than men. Right?[2]

The wage data can also be summarised in tabular form. This is done in Table 1.5. The range of the data is divided into the classes used to draw

[1] The box plot in Figure 1.2 was drawn and the median, percentiles and interquartile range above were calculated using XlispStat, a statistical program freely available on the Internet for the Unix, Linux, MS Windows (3.1, 95, 98, NT, XP, Vista and 7) and Macintosh operating systems. It is easy to learn to do the simple things we need to do for this course using XlispStat but extensive use of it requires knowledge of object-oriented-programming and a willingness to learn features of the Lisp programming language. Commercial programs such as SAS, SPSS, and Minitab present more sophisticated box plots than the one presented here but, of course, these programs are more costly to obtain.

[2] Wrong! First of all, this is data for only one firm, which need not be representative of all firms in the economy. Second, there are no references as to where the data came from—as a matter of fact, I made them up!

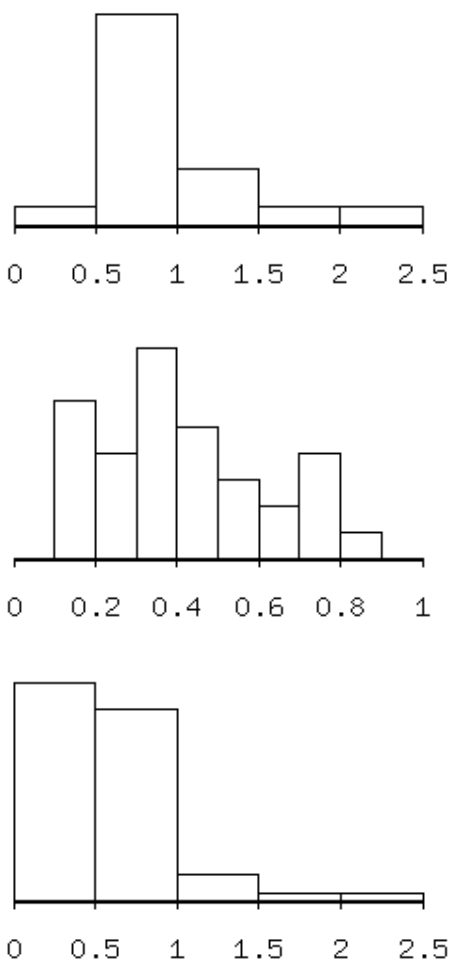


Figure 1.1: Histogram of weekly wages for male (top), female (middle) and all (bottom) employees. The horizontal scale is thousands of U.S. dollars.

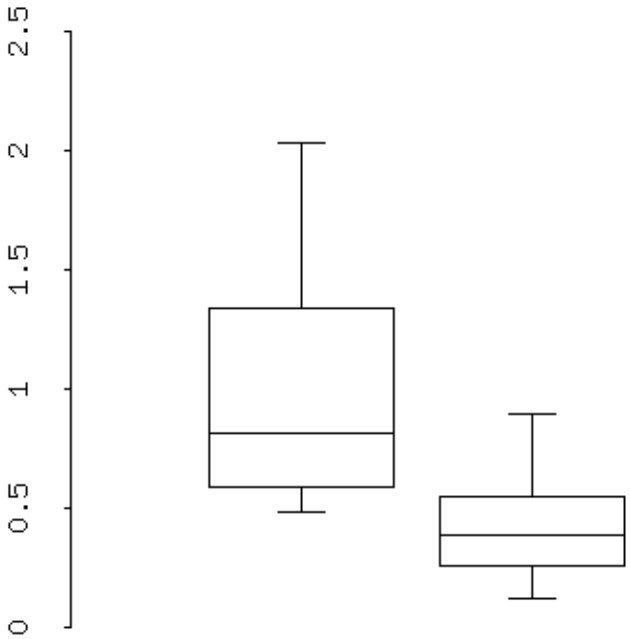


Figure 1.2: Box plot of weekly wages for males (left) and females (right). The vertical scale is thousands of U.S. dollars.

Table 1.5: Frequency Distributions From Table 1.3

Class	Frequency			Relative Frequency		
	M	F	Total	M	F	Total
0.0 – 0.5	1	23	24	.06	.70	.48
0.5 – 1.0	10	10	20	.58	.30	.40
1.0 – 1.5	4	0	4	.24	.00	.08
1.5 – 2.0	1	0	1	.06	.00	.02
2.0 – 2.5	1	0	1	.06	.00	.02
Total	17	33	50	1.00	1.00	1.00

the histogram for the full data set. Then the observations for the wage variable in Table 1.3 that fall in each of the classes are counted and the numbers entered into the appropriate cells in columns 2, 3 and 4 of the table. The observations are thus 'distributed' among the classes with the numbers in the cells indicating the 'frequency' with which observations fall in the respective classes—hence, such tables present *frequency distributions*. The totals along the bottom tell us that there were 17 men and 33 women, with a total of 50 elements in the data set. The relative frequencies in which observations fall in the classes are shown in columns 5, 6 and 7. Column 5 gives the proportions of men's wages, column 6 the proportions of women's wages and column 7 the proportions of all wages falling in the classes. The proportions in each column must add up to one.

All of the data sets considered thus far are *cross-sectional*. Tables 1.6 and 1.7 present time-series data sets. The first table gives the consumer price indexes for four countries, Canada, the United States, the United Kingdom and Japan, for the years 1975 to 1996.[3] The second table presents the year-over-year inflation rates for the same period for these same countries. The inflation rates are calculated as

$$\pi = [100(P_t - P_{t-1})/P_{t-1}]$$

where π denotes the inflation rate and P denotes the consumer price index. It should now be obvious that in time-series data the elements are units of time. This distinguishes time-series from cross-sectional data sets, where all observations occur in the same time period.

A frequent feature of time-series data not present in cross-sectional data is *serial correlation* or *autocorrelation*. The data in Tables 1.6 and 1.7 are plotted in Figures 1.3 and 1.4 respectively. You will notice from these plots that one can make a pretty good guess as to what the price level or inflation rate will be in a given year on the basis of the observed price level and inflation rate in previous years. If prices or inflation are high this year, they will most likely also be high next year. Successive observations in each series are serially correlated or autocorrelated (i.e., correlated through time) and hence not statistically independent of each other. Figure 1.5 shows a time-series that has no autocorrelation—the successive observations were generated completely independently of all preceding observations using a computer. You will learn more about correlation and statistical independence later in this chapter.

[3]Consumer price indexes are calculated by taking the value in each year of the bundle of goods consumed by a typical person as a percentage of the monetary value of that same bundle of goods in a base period. In Table 1.6 the base year is 1980.

Table 1.6: Consumer Price Indexes for Selected Countries, 1980 = 100

	Canada	U.S.	U.K.	Japan
1975	65.8	65.3	51.1	72.5
1976	70.7	69.0	59.6	79.4
1977	76.3	73.5	69.0	85.9
1978	83.1	79.1	74.7	89.4
1979	90.8	88.1	84.8	92.8
1980	100.0	100.0	100.0	100.0
1981	112.4	110.3	111.9	104.9
1982	124.6	117.1	121.5	107.8
1983	131.8	120.9	127.1	109.8
1984	137.6	126.0	133.4	112.3
1985	143.0	130.5	141.5	114.6
1986	149.0	133.0	146.3	115.3
1987	155.5	137.9	152.4	115.4
1988	161.8	143.5	159.9	116.2
1989	169.8	150.4	172.4	118.9
1990	177.9	158.5	188.7	122.5
1991	187.9	165.2	199.7	126.5
1992	190.7	170.2	207.2	128.7
1993	194.2	175.3	210.4	130.3
1994	194.6	179.9	215.7	131.2
1995	198.8	184.9	223.0	131.1
1996	201.9	190.3	228.4	131.3

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

Table 1.7: Year-over-year Inflation Rates for Selected Countries, Percent Per Year

	Canada	U.S.	U.K.	Japan
1975	10.9	9.1	24.1	11.8
1976	7.5	5.7	16.6	9.4
1977	8.0	6.5	15.9	8.2
1978	8.9	7.6	8.2	4.1
1979	9.2	11.3	13.5	3.8
1980	10.2	13.6	17.9	7.8
1981	12.4	10.3	11.9	4.9
1982	10.8	6.2	8.6	2.7
1983	5.8	3.2	4.6	1.9
1984	4.3	4.3	5.0	2.2
1985	3.9	3.6	6.1	2.0
1986	4.2	1.9	3.4	0.6
1987	4.4	3.6	4.2	0.1
1988	4.0	4.1	4.9	0.7
1989	5.0	4.2	7.8	2.3
1990	4.8	5.4	9.5	3.1
1991	5.6	4.2	5.8	3.3
1992	1.5	3.0	3.7	1.7
1993	1.8	3.0	1.6	1.3
1994	0.2	2.6	2.4	0.7
1995	2.2	2.8	3.4	-0.1
1996	1.6	2.9	2.4	0.1

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

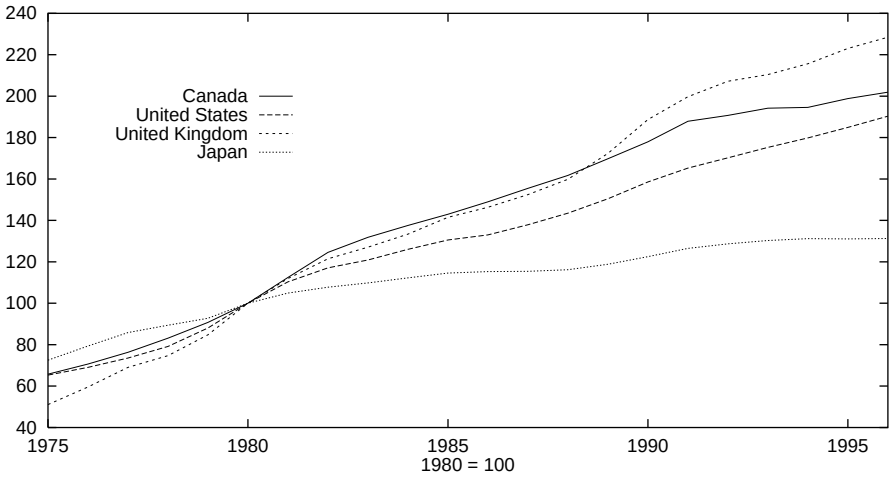


Figure 1.3: Consumer price indexes of selected countries

.

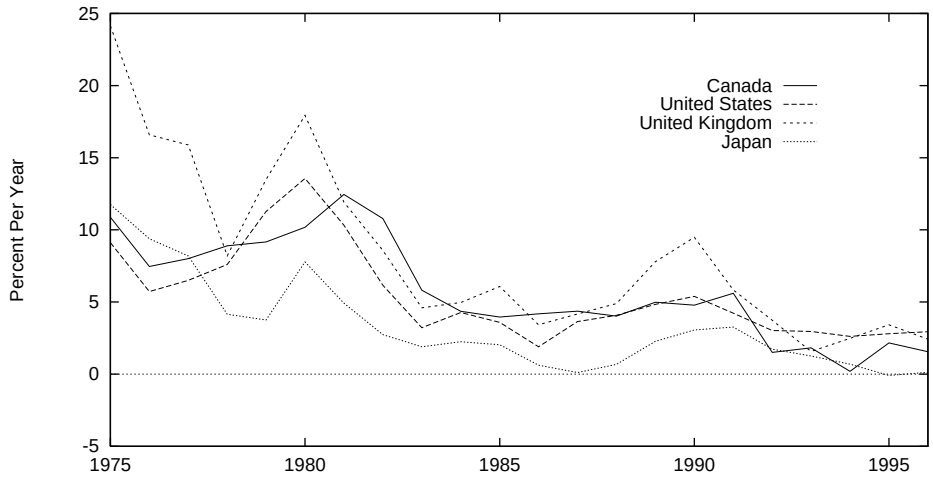


Figure 1.4: Year-over year inflation rates of selected countries

.

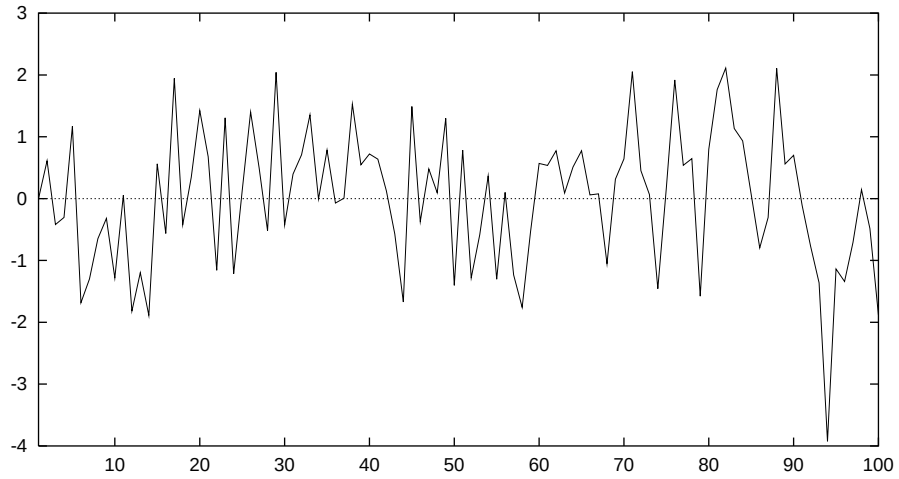


Figure 1.5: A time-series devoid of autocorrelation

.

Some data sets are both time-series and cross-sectional. Imagine, for example a data set containing wage and gender data of the sort in Table 1.3 for each of a series of years. These are called *panel data*. We will not be working with panel data in this book.

1.6 Numerical Measures of Position

Although quite a bit of information about data sets can be obtained by constructing tables and graphs, it would be nice to be able to describe a data set using two or three numbers. The median, range, interquartile range, maximum, and minimum, which were calculated for the wage data in the previous section and portrayed graphically in Figure 1.2 using a box plot, provide such a description. They tell us where the centre observation is, the range in which half of the observations lie (interquartile range) and the range in which the whole data set lies. We can see, for example, that both male and female wages are concentrated more at the lower than at the higher levels.

There are three types of numerical summary measures that can be used to describe data sets. First, there are measures of position or central tendency. Is the typical wage rate paid by the firm in question, for example, around \$500 per week, or \$1500 per week, or \$5000 per week? The median provides one measure of position. Second, there are measures of variability

or dispersion. Are all the weekly wages very close to each other or are they spread out widely? The range and the interquartile range provide measures of variability—the bigger these statistics, the more dispersed are the data. Finally, there are measures of skewness. Are wages more concentrated, for example, at the lower levels, or are they dispersed symmetrically around their central value? In this section we will concentrate on numerical measures of position. Measures of variability and skewness will be considered in the subsequent two sections.

The median is a measure of position. In the case of the wage data, for example, it tells us that half the wages are below \$521.50 and half are above that amount. Another important measure of position is the *mean* (or, more precisely, the *arithmetic mean*), commonly known as the average value. The mean of a set of numbers $X_1, X_2, X_3, \ldots, X_N$ is defined as

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{1.1}$$

where $\bar{X}$ is the arithmetic mean and

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + X_3 + \ldots + X_N. \tag{1.2}$$

The sum of the weekly wage data (including both males and females) is \$30364 and the mean is \$607.28. The mean wages of males and females are, respectively, \$962.24 and \$424.42. It follows from equation (1.1) that the sum of the observations on a particular quantitative variable in a data set is equal to the mean times the number of items,

$$\sum_{i=1}^{N} X_i = N\bar{X}, \tag{1.3}$$

and that the sum of the deviations of the observations from their mean is zero,

$$\sum_{i=1}^{N}(X_i - \bar{X}) = \sum_{i=1}^{N} X_i - N\bar{X} = N\bar{X} - N\bar{X} = 0. \tag{1.4}$$

When a set of items is divided into classes, as must be done to create a frequency distribution, the overall mean is a weighted average of the means

of the observations in the classes, with the weights being the number (or frequency) of items in the respective classes. When there are k classes,

$$\bar{X} = \frac{f_1\bar{X}_1 + f_2\bar{X}_2 + f_3\bar{X}_3 + \ldots + f_k\bar{X}_k}{N} = \frac{\sum_{i=1}^{k} f_i\bar{X}_i}{N} \quad (1.5)$$

where $\bar{X}_i$ is the mean of the observations in the ith class and f_i is the number (frequency) of observations in the ith class. If all that is known is the frequency in each class with no measure of the mean of the observations in the classes available, we can obtain a useful approximation to the mean of the data set using the mid-points of the classes in the above formula in place of the class means.

An alternative mean value is the *geometric mean* which is defined as the anti-log of the arithmetic mean of the logarithms of the values. The geometric mean can thus be obtained by taking the anti-log of

$$\frac{\log X_1 + \log X_2 + \log X_3 + \ldots + \log X_N}{N}$$

or the nth root of $X_1X_2X_3 \ldots X_N$.[4] Placing a bar on top of a variable to denote its mean, as in $\bar{X}$, is done only to represent means of samples. The mean of a population is represented by the Greek symbol μ. When the population is finite, μ can be obtained by making the calculation in equation 1.1 using all elements in the population. The mean of an infinite population generated by a process has to be derived from the mathematical representation of that process. In most practical cases this mathematical data generating process is unknown. The ease of obtaining the means of finite as opposed to infinite populations is more apparent than real. The cost of calculating the mean for large finite populations is usually prohibitive because a census of the entire population is required.

The mean is strongly influenced by extreme values in the data set. For example, suppose that the members of a small group of eight people have the following annual incomes in dollars: 24000, 23800, 22950, 26000, 275000, 25500, 24500, 23650. We want to present a single number that characterises

[4]Note from the definition of logarithms that taking the logarithm of the nth root of $(X_1X_2X_3 \ldots X_N)$, which equals

$$(X_1X_2X_3 \ldots X_N)^{\frac{1}{N}},$$

yields

$$\frac{\log X_1 + \log X_2 + \log X_3 + \ldots + \log X_N}{N}.$$

how 'well off' this group of people is. The (arithmetic) mean income of the group is \$55675.[5] But a look at the actual numbers indicates that all but one member of the group have incomes between \$23000 and \$26000. The mean does not present a good picture because of the influence of the enormous income of one member of the group.

When there are extreme values, a more accurate picture can often be presented by using a *trimmed* mean. The 50 percent trimmed mean, for example, is the (arithmetic) mean of the central 50 percent of the values—essentially, the mean of the values lying in the interquartile range. This would be \$24450 in the example above. We could, instead, use an 80 (or any other) percent trimmed mean. The median, which is \$24250 is also a better measure of the central tendency of the data than the mean. It should always be kept in mind, however, that extreme values may provide important information and it may be inappropriate to ignore them. Common sense is necessary in presenting and interpreting data. In the example above, the most accurate picture would be given by the following statement: Seven of the eight members of the group have incomes between \$22950 and \$26000, with mean \$24342, while the eighth member has an income of \$275000.

Another measure of position of the *mode*, which is defined as the most frequently appearing value. When the variable is divided into equal-sized classes and presented as a histogram or frequency distribution the class containing the most observations is called the *modal class*. In the wage data, using the classes defined in Table 1.5, the modal class for females and for all workers is \$0–\$500, and the modal class for males is \$500–\$1000. Using the classes defined in the middle panel of Figure 1.1 the modal class for female wages is \$300–\$400.

Sometimes there will be two peaks in a histogram of the observations for a variable. A frequent example is the performance of students on mathematics (and sometimes statistics) tests where the students divide into two groups—those who understand what is going on and those to do not. Given that there is variability within each group there will typically be two humps in the histogram—one at a high grade containing the students who understand the material and one at a low grade containing the students who do not understand the material. In such situations the data are referred to as *bimodal*. Figure 1.6 gives examples of a bimodal and a unimodal or hump-shaped distribution. We could imagine the horizontal scales as representing the grade achieved on a mathematics test.

[5]The arithmetic mean is generally referred to as simply the mean with the geometric mean, which is rarely used, denoted by its full name. The geometric mean of the eight

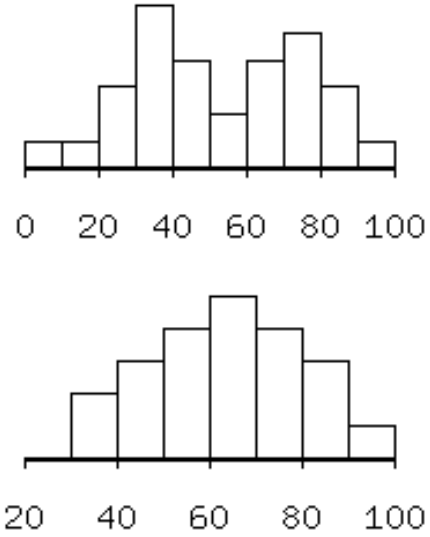


Figure 1.6: Bimodal distribution (top) and unimodal or humped-shaped distribution (bottom).

1.7 Numerical Measures of Variability

The range and interquartile range are measures of variability—the bigger these are, the more dispersed are the data. More widely used measures, however, are the *variance* and *standard deviation*. The variance is, broadly, the mean or average of the squared deviations of the observations from their mean. For data sets that constitute samples from populations or processes the calculation is

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}, \tag{1.6}$$

where s^2 denotes the sample variance. An approximation can be calculated from a frequency distribution of the sample using

$$s^2 = \frac{\sum_{i=1}^{S} f_i(\bar{X}_i - \bar{X})^2}{N-1}, \tag{1.7}$$

where S is the number of classes, f_i is the frequency of the ith class, $\bar{X}_i$ is the mean of the ith class, $\bar{X}$ is the mean of the whole sample and the total

observations above is $32936.

number of elements in the sample equals

$$N = \sum_{i=1}^{S} f_i.$$

The population variance is denoted by σ^2. For a finite population it can be calculated using (1.6) after replacing $N-1$ in the denominator by N. $N-1$ is used in the denominator in calculating the sample variance because the variance is the mean of the sum of squared *independent* deviations from the sample mean and only $N-1$ of the N deviations from the sample mean can be independently selected—once we know $N-1$ of the deviations, the remaining one can be calculated from those already known based on the way the sample mean was calculated. Each sample from a given population will have a different sample mean, depending upon the population elements that appear in it. The population mean, on the other hand, is a fixed number which does not change from sample to sample. The deviations of the population elements from the population mean are therefore all independent of each other. In the case of a process, the exact population variance can only be obtained from knowledge of the mathematical data-generation process.

In the weekly wage data above, the variance of wages is 207161.5 for males, 42898.7 for females and 161893.7 for the entire sample. Notice that the units in which these variances are measured is dollars-squared—we are taking the sum of the squared dollar-differences of each person's wage from the mean. To obtain a measure of variability measured in dollars rather than dollars-squared we can take the square root of the variance—s in equation (1.6). This is called the *standard deviation*. The standard deviation of wages in the above sample is \$455.15 for males, \$207.12 for females, and \$402.36 for the entire sample.

Another frequently used measure of variability is the *coefficient of variation*, defined as the standard deviation taken as a percentage of the mean,

$$C = \frac{100s}{\bar{X}}, \tag{1.8}$$

where C denotes the coefficient of variation. For the weekly wage data above, the coefficient of variation is 47.30 for males, 48.8 for females and 66.28 for the entire sample.

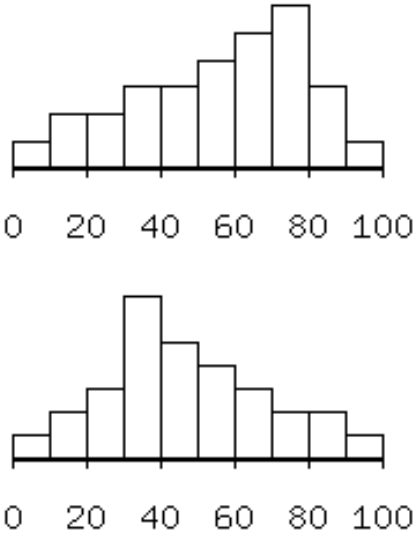


Figure 1.7: Left-skewed distribution (top—mean = 55.1 , median = 58, mode = 75) and right-skewed distribution (bottom —mean = 46.4, median = 43.5, mode = 35).

1.8 Numerical Measures of Skewness

Skewed quantitative data are data for which a frequency distribution based on equal classes is not symmetrical. For example, the wage data presented Figure 1.1 are not symmetrical—the right tail is longer than the left tail, which is non-existent in the bottom panel. These data are described as *skewed right*—the skew is in the direction of the longer tail. This skewness appears in the box plots in Figure 1.2 as a longer upper whisker than lower whisker. Notice that in the wage data the mean is always larger than the median and the median larger than the mode. The means, medians and modes (taken as the mid-points of the modal classes) are respectively \$962, \$822.5 and \$750 for males, \$424, \$391 and \$350 for females and \$607, \$521 and \$200 for all workers. The mean will always exceed the median and the median will always exceed the mode when the data are skewed to the right. When the skew is to the left the mean will be below the median and the median below the mode. This is shown in Figure 1.7. The rightward

(leftward) skew is due to the influence of the rather few unusually high (low) values—the extreme values drag the mean in their direction. The median tends to be above the mode when the data are skewed right because low values are more frequent than high values and below the mode when the data are skewed to the left because in that case high values are more frequent than low values. When the data are symmetrically distributed, the mean, median and mode are equal.

Skewness can be measured by the average cubed deviation of the values from the sample mean,

$$m^3 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^3}{N-1}. \tag{1.9}$$

If the large deviations are predominately positive m^3 will be positive and if the large deviations are predominately negative m^3 will be negative. This happens because $(X_i - \bar{X})^3$ has the same sign as $(X_i - \bar{X})$. Since large deviations are associated with the long tail of the frequency distribution, m^3 will be positive or negative according to whether the direction of skewness is positive (right) or negative (left). In the wage data m^3 is positive for males, females and all workers as we would expect from looking at figures 1.1 and 1.2.

1.9 Numerical Measures of Relative Position: Standardised Values

In addition to measures of the central tendency of a set of values and their dispersion around these central measures we are often interested in whether a particular observation is high or low relative to others in the set. One measure of this is the percentile in which the observation falls—if an observation is at the 90th percentile, only 10% of the values lie above it and 90% percent of the values lie below it. Another measure of relative position is the *standardised value*. The standardised value of an observation is its distance from the mean divided by the standard deviation of the sample or population in which the observation is located. The standardised values of the set of observations $X_1, X_2, X_3 \ldots X_N$ are given by

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{1.10}$$

for members of a population whose mean μ and standard deviation σ are known and

$$Z_i = \frac{X_i - \bar{X}}{s} \tag{1.11}$$

for members of a sample with mean $\bar{X}$ and sample standard deviation s. The standardised value or z-value of an observation is the number of standard deviations it is away from the mean.

It turns out that for a distribution that is hump-shaped—that is, not bimodal—roughly 68% of the observations will lie within plus or minus one standard deviation from the mean, about 95% of the values will lie within plus or minus two standard deviations from the mean, and roughly 99.7% of the observations will lie within plus or minus three standard deviations from the mean. Thus, if you obtain a grade of 52% percent on a statistics test for which the class average was 40% percent and the standard deviation 10% percent, and the distribution is hump-shaped rather than bimodal, you are probably in the top 16 percent of the class. This calculation is made by noting that about 68 percent of the class will score within one standard deviation from 40—that is, between 30 and 50—and 32 percent will score outside that range. If the two tails of the distribution are equally populated then you must be in the top 16% percent of the class. Relatively speaking, 52% was a pretty good grade.

The above percentages hold almost exactly for *normal distributions*, which you will learn about in due course, and only approximately for hump-shaped distributions that do not satisfy the criteria for normality. They do not hold for distributions that are bimodal. It turns out that there is a rule developed by the Russian mathematician P. L. Chebyshev, called *Chebyshev's Inequality*, which states that a fraction no bigger than $(1/k)^2$ (or $100 \times (1/k)^2$ percent) of any set of observations, no matter what the shape of their distribution, will lie beyond plus or minus k standard deviations from the mean of those observations. So if the standard deviation is 2 at least 75% of the distribution must lie within plus or minus two standard deviations from the mean and no more than 25% percent of the distribution can lie outside that range in one or other of the tails. You should note especially that the rule does *not* imply here that no more than 12.5% percent of a distribution will lie two standard deviations above the mean because the distribution need not be symmetrical.

1.10 Bivariate Data: Covariance and Correlation

A data set that contains only one variable of interest, as would be the case with the wage data above if the gender of each wage earner was not recorded, is called a *univariate* data set. Data sets that contain two variables, such as wage and gender in the wage data above, are said to be *bivariate*. And the consumer price index and inflation rate data presented in Table 1.6 and Table 1.7 above are *multivariate*, with each data set containing four variables—consumer price indexes or inflation rates for four countries.

In the case of bivariate or multivariate data sets we are often interested in whether elements that have high values of one of the variables also have high values of other variables. For example, as students of economics we might be interested in whether people with more years of schooling earn higher incomes. From Canadian Government census data we might obtain for the population of all Canadian households two quantitative variables, household income (measured in $) and number of years of education of the head of each household.[6] Let X_i be the value of annual household income for household i and Y_i be the number of years of schooling of the head of the ith household. Now consider a random sample of N households which yields the paired observations (X_i, Y_i) for $i = 1, 2, 3, \ldots, N$.

You already know how to create summary statistical measures for single variables. The sample mean value for household incomes, for example, can be obtained by summing up all the X_i and dividing the resulting sum by N. And the sample mean value for years of education per household can similarly be obtained by summing all the Y_i and dividing by N. We can also calculate the sample variances of X and Y by applying equation (1.6).

Notice that the fact that the sample consists of *paired* observations (X_i, Y_i) is irrelevant when we calculate summary measures for the individual variables X and/or Y. Nevertheless, we may also be interested in whether the variables X and Y are related to one another in a systematic way. Since education is a form of investment that yields its return in the form of higher lifetime earnings, we might expect, for example, that household income will tend to be higher the greater the number of years of education completed by the head of household. That is, we might expect high values of X to be paired with high values of Y—when X_i is high, the Y_i associated with it should also be high, and vice versa.

Another example is the consumer price indexes and inflation rates for

[6] This example and most of the prose in this section draws on the expositional efforts of Prof. Greg Jump, my colleague at the University of Toronto.

pairs of countries. We might ask whether high prices and high inflation rates in the United States are associated with high prices and inflation rates in Canada. One way to do this is to construct scatter plots with the Canadian consumer price index and the Canadian inflation rate on the horizontal axes and the U.S. consumer price index and the U.S. inflation rate on the respective vertical axes. This is done in Figure 1.8 for the consumer price indexes and Figure 1.9 for the inflation rates. You can see from the figures that both the price levels and inflation rates in the two countries are positively related with the relationship being 'tighter' in the case of the price levels than in the case of the inflation rates.

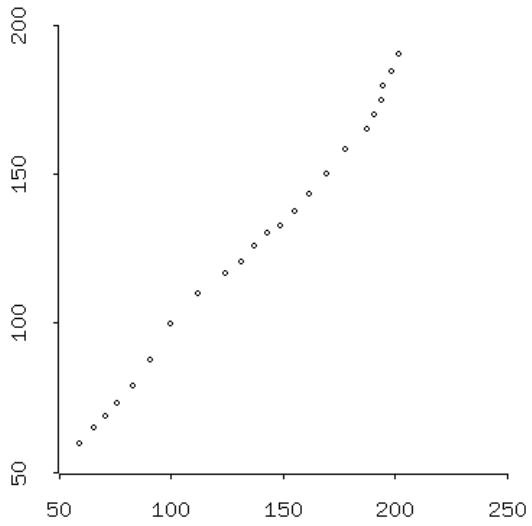


Figure 1.8: Scatterplot of the Canadian consumer price index (horizontal axis) vs. the U.S. consumer price index (vertical axis).

We can also construct numerical measures of covariability. One such measure is the *covariance* between the two variables, denoted in the case of sample data as $s_{x,y}$ or $s_{y,x}$ and defined by

$$\begin{aligned} s_{x,y} &= \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1} \\ &= \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{N-1} = s_{y,x}. \end{aligned} \tag{1.12}$$

When X and Y represent a population we denote the covariance between

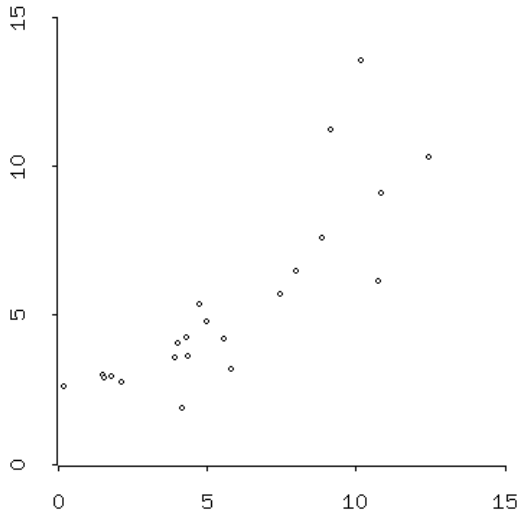


Figure 1.9: Scatterplot of the Canadian year-over-year inflation rate (horizontal axis) against the U.S. year-over-year inflation rate (vertical axis).

them by $\sigma_{x,y}$ or $\sigma_{y,x}$. It can be calculated using (1.12) with the $N-1$ in the denominator replaced by N in the case where an entire finite population is used in the calculation. In an infinite population generated by a process, the covariance can only be obtained from knowledge of the mathematics of the data generation process. Notice that the value of the covariance is independent of the order of the multiplicative terms within the summation sign. Note also that $s_{x,y}$ is measured in units of X times units of Y—in our annual household income and years of schooling of household head example, $s_{x,y}$ would be expressed in terms of "dollar-years" (whatever those might be).

For any sample of paired variables X and Y, $s_{x,y}$ has a single numerical value that may be positive, negative or zero. A positive value indicates that the observed values for X and Y are *positively related*—that is, they tend to rise and fall together. To put it somewhat differently, a positive value for $s_{x,y}$ indicates that X_i tends to be above (below) its mean value $\bar{X}$ whenever Y_i is above (below) its mean value $\bar{Y}$. Similarly, the variables X and Y are *negatively related* whenever $s_{x,y}$ is negative in sign. This means that X_i tends to be below (above) its mean value $\bar{X}$ whenever Y_i is above (below)

its mean value $\bar{Y}$. When there is no relationship between the variables X and Y, $s_{x,y}$ is zero.

In our household income and education example we would expect that a random sample would yield a positive value for $s_{x,y}$ and this is indeed what is found in actual samples drawn from the population of all Canadian households.

Note that equation (1.12) could be used to compute $s_{x,x}$—the covariance of the variable X with itself. It is easy to see from equations (1.12) and (1.6) that this will yield the sample variance of X which we can denote by s_x^2. It might be thus said that the concept of *variance* is just a special case of the more general concept of *covariance*.

The concept of covariance is important in the study of financial economics because it is critical to an understanding of 'risk' in securities and other asset markets. Unfortunately, it is a concept that yields numbers that are not very 'intuitive'. For example, suppose we were to find that a sample of N Canadian households yields a covariance of $+1,000$ dollar-years between annual household income and years of education of head of household. The covariance is positive in sign, so we know that this implies that households with highly educated heads tend to have high annual incomes. But is there any intuitive interpretation of the magnitude 1000 dollar-years? The answer is no, at least not without further information regarding the individual sample variances of household income and age of head.

A more intuitive concept, closely related to covariance, is the *correlation* between two variables. The *coefficient of correlation* between two variables X and Y, denoted by $r_{x,y}$ or, equivalently, $r_{y,x}$ is defined as

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} = r_{y,x} \tag{1.13}$$

where s_x and s_y are the sample standard deviations of X and Y calculated by using equation (1.6) above and taking square roots.

It should be obvious from (1.13) that the sign of the correlation coefficient is the same as the sign of the covariance between the two variables since standard deviations cannot be negative. Positive covariance implies positive correlation, negative covariance implies negative correlation and zero covariance implies that X and Y are uncorrelated. It is also apparent from (1.13) that $r_{x,y}$ is independent of the units in which X and Y are measured—it is a unit-free number. What is not apparent (and will not be proved at this time) is that for any two variables X and Y,

$$-1 \leq r_{x,y} \leq +1.$$

That is, the correlation coefficient between any two variables must lie in the interval $[-1, +1]$. A value of plus unity means that the two variables are perfectly positively correlated; a value of minus unity means that they are perfectly negatively correlated. Perfect correlation can only happen when the variables satisfy an exact linear relationship of the form

$$Y = a + bX$$

where b is positive when they are perfectly positively correlated and negative when they are perfectly negatively correlated. If $r_{x,y}$ is zero, X and Y are said to be perfectly uncorrelated. Consider the relationships between the Canadian and U.S. price levels and inflation rates. The coefficient of correlation between the Canadian and U.S. consumer price indexes plotted in Figure 1.8 is .99624, which is very close to +1 and consistent with the fact that the points in the figure are almost in a straight line. There is less correlation between the inflation rates of the two countries, as is evident from the greater 'scatter' of the points in Figure 1.9 around an imaginary straight line one might draw through them. Here the correlation coefficient is .83924, considerably below the coefficient of correlation of the two price levels.

1.11 Exercises

1. Write down a sentence or two explaining the difference between:

   a) Populations and samples.

   b) Populations and processes.

   c) Elements and observations.

   d) Observations and variables.

   e) Covariance and correlation.

2. You are tabulating data that classifies a sample of 100 incidents of domestic violence according to the Canadian Province in which each incident occurs. You number the provinces from west to east with British Columbia being number 1 and Newfoundland being number 10. The entire Northern Territory is treated for purposes of your analysis as a province and denoted by number 11. In your tabulation you write down next to each incident

the assigned number of the province in which it occurred. Is the resulting column of province numbers a quantitative or qualitative variable?

3. Calculate the variance and standard deviation for samples where

a) $n = 10$, $\Sigma X^2 = 84$, and $\Sigma X = 20$. (4.89, 2.21)

b) $n = 40$, $\Sigma X^2 = 380$, and $\Sigma X = 100$.

c) $n = 20$, $\Sigma X^2 = 18$, and $\Sigma X = 17$.

Hint: Modify equation (1.6) by expanding the numerator to obtain an equivalent formula for the sample variance that directly uses the numbers given above.

4. Explain how the relationship between the mean and the median provides information about the symmetry or skewness of the data's distribution.

5. What is the primary disadvantage of using the range rather than the variance to compare the variability of two data sets?

6. Can standard deviation of a variable be negative?

7. A sample is drawn from the population of all adult females in Canada and the height in centimetres is observed. One of the observations has a sample z-score of 6. Describe in one sentence what this implies about that particular member of the sample.

8. In archery practice, the mean distance of the points of impact from the target centre is 5 inches. The standard deviation of these distances is 2 inches. At most, what proportion of the arrows hit within 1 inch or beyond 9 inches from the target centre? Hint: Use $1/k^2$.

a) 1/4

b) 1/8

c) 1/10

d) cannot be determined from the data given.

e) none of the above.

9. Chebyshev's rule states that 68% of the observations on a variable will lie within plus or minus two standard deviations from the mean value for that variable. True or False. Explain your answer fully.

10. A manufacturer of automobile batteries claims that the average length of life for its grade A battery is 60 months. But the guarantee on this brand is for just 36 months. Suppose that the frequency distribution of the life-length data is unimodal and symmetrical and that the standard deviation is known to be 10 months. Suppose further that your battery lasts 37 months. What could you infer, if anything, about the manufacturer's claim?

11. At one university, the students are given z-scores at the end of each semester rather than the traditional GPA's. The mean and standard deviations of all students' cumulative GPA's on which the z-scores are based are 2.7 and 0.5 respectively. Students with z-scores below -1.6 are put on probation. What is the corresponding probationary level of the GPA?

12. Two variables have identical standard deviations and a covariance equal to half that common standard deviation. If the standard deviation of the two variables is 2, what is the correlation coefficient between them?

13. Application of Chebyshev's rule to a data set that is roughly symmetrically distributed implies that at least one-half of all the observations lie in the interval from 3.6 to 8.8. What are the approximate values of the mean and standard deviation of this data set?

14. The number of defective items in 15 recent production lots of 100 items each were as follows:

3, 1, 0, 2, 24, 4, 1, 0, 5, 8, 6, 3, 10, 4, 2

a) Calculate the mean number of defectives per lot. (4.87)

b) Array the observations in ascending order. Obtain the median of this data set. Why does the median differ substantially from the mean here? Obtain the range and the interquartile range. (3, 24, 4)

c) Calculate the variance and the standard deviation of the data set. Which observation makes the largest contribution to the magnitude of the variance through the sum of squared deviations? Which observation makes the smallest contribution? What general conclusions are implied by these findings? (36.12, 6.01)

d) Calculate the coefficient of variation for the number of defectives per lot. (81)

e) Calculate the standardised values of the fifteen numbers of defective items. Verify that, except for rounding effects, the mean and variance of these standardised observations are 0 and 1 respectively. How many standard deviations away from the mean is the largest observation? The smallest?

15. The variables X and Y below represent the number of sick days taken by the males and females respectively of seven married couples working for a given firm. All couples have small children.

X	8	5	4	6	2	5	3
Y	1	3	6	3	7	2	5

Calculate the covariance and the correlation coefficient between these variables and suggest a possible explanation of the association between them. (-3.88, -0.895)

Chapter 2

Probability

2.1 Why Probability?

We have seen that statistical inference is a methodology through which we learn about the characteristics of a population by analyzing samples of elements drawn from that population. Suppose that a friend asks you to invest $10000 in a joint business venture. Although your friend's presentation of the potential for profit is convincing, you investigate and find that he has initiated three previous business ventures, all of which failed. Would you think that the current proposed venture would have more than a 50/50 chance of succeeding? In pondering this question you must wonder about the likelihood of observing three failures in a sample of three elements from the process by which your friend chooses and executes business ventures if, in fact, more than half the population of ventures emanating from that process will be successful. This line of thinking is an essential part of statistical inference because we are constantly asking ourselves, in one way or other, what the likelihood is of observing a particular sample if the population characteristics are what they are purported to be. Much of statistical inference involves making an hypothesis about the characteristics of a population (which we will later call the null hypothesis) and then seeing whether the sample has a low or high chance of occurring if that hypothesis is true.

Let us begin our study of probability by starting with a population whose characteristics are known to us and inquire about the likelihood or chances of observing various samples from that population.

2.2 Sample Spaces and Events

Suppose we toss a single coin and observe whether it comes up heads or tails. The relevant population here is the infinite sequence of tosses of a single coin. With each toss there is uncertainty about whether the result will be a head or a tail. This coin toss is an example of a *random trial* or *experiment*, which can be defined as an activity having two or more possible outcomes with uncertainty in advance as to which outcome will prevail. The different possible outcomes of the random trial are called the *basic outcomes*. The set of all basic outcomes for a random trial is called the *sample space* for the trial. The sample space for a single coin toss, which we denote by S, contains two basic outcomes, denoted as H (head) and T (tail). This represents a sample of one from the infinite population of single coin tosses. The set of basic outcomes can be written

$$S = \{H, T\} \tag{2.1}$$

These basic outcomes are also called *sample points* or *simple events*. They are *mutually exclusive*—that is, only one can occur—and *mutually exhaustive*—that is, at least one of them must occur.

Now suppose we toss two coins simultaneously and record whether they come up heads or tails. One might think that there would be three basic outcomes in this case—two heads, head and tail, and two tails. Actually, there are four simple events or sample points because the combination head and tail can occur in two ways—head first and then tail, and tail first followed by head. Thus, the sample space for this random trial or experiment will be

$$S = \{HH, HT, TH, TT\} \tag{2.2}$$

A subset of the set of sample points is called an *event*. For example, consider the event 'at least one head'. This would consist of the subspace

$$E_1 = \{HH, HT, TH\} \tag{2.3}$$

containing three of the four sample points. Another event would be 'both faces same'. This event, which we can call E_2, is the subset

$$E_2 = \{HH, TT\}. \tag{2.4}$$

The set of outcomes not contained in an event E_j is called the *complementary event* to the event E_j which we will denote by E_j^c. Thus, the complementary events to E_1 and E_2 are, respectively,

$$E_1^c = \{TT\} \tag{2.5}$$

and

$$E_2^c = \{HT, TH\}. \tag{2.6}$$

The set of sample points that belongs to both event E_i and event E_j is called the *intersection* of E_i and E_j. The intersection of E_1 and E_2^c turns out to be the event E_2^c because both sample points in E_2^c are also in E_1. We can write this as

$$E_1 \cap E_2^c = \{HT, TH\} = E_2^c. \tag{2.7}$$

The intersection of E_1^c and E_2^c contains no elements, that is

$$E_1^c \cap E_2^c = \phi \tag{2.8}$$

where ϕ means *nil* or nothing. An event containing no elements is called the *null set* or *null event*. When the intersection of two events is the null event, those two events are said to be *mutually exclusive*. It should be obvious that the intersection of an event and its complement is the null event.

The set of sample points that belong to at least one of the events E_i and E_j is called the *union* of E_i and E_j. For example, the union of E_1^c and E_2^c is

$$E_3 = E_1^c \cup E_2^c = \{HT, TH, TT\}, \tag{2.9}$$

the event 'no more than one head'. Each sample point is itself an event—one of the elementary events—and the union of all these elementary events is the sample space itself. An event that contains the entire sample space is called the *universal event*.

We can express the intersection and union of several events as, respectively,

$$E_1 \cap E_2 \cap E_3 \cap E_4 \cap \cdots\cdots$$

and

$$E_1 \cup E_2 \cup E_3 \cup E_4 \cup \cdots\cdots.$$

The set of all possible events that can occur in any random trial or experiment, including both the universal event and the null event, is called the *event space*.

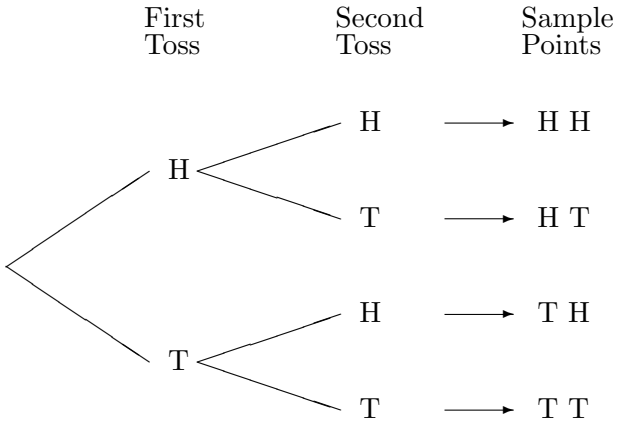
The above examples of random trials and sample spaces resulting therefrom represent perhaps the simplest cases one could imagine. More complex situations arise in experiments such as the daily change in the Dow Jones Industrial Average, the number of students of the College involved in accidents in a given week, the year-over-year rate of inflation in the United Kingdom, and so forth. Sample points, the sample space, events and the event space in these more complicated random trials have the same meanings and are defined in the same way as in the simple examples above.

2.3 Univariate, Bivariate and Multivariate Sample Spaces

The sample space resulting from a single coin toss is a univariate sample space—there is only one dimension to the random trial. When we toss two coins simultaneously, the sample space has two dimensions—the result of the first toss and the result of the second toss. It is often useful to portray bivariate sample spaces like this one in tabular form as follows:

	One	
Two	H	T
H	HH	TH
T	HT	TT

Each of the four cells of the table gives an outcome of the first toss followed by an outcome of the second toss. This sample space can also be laid out in tree form:

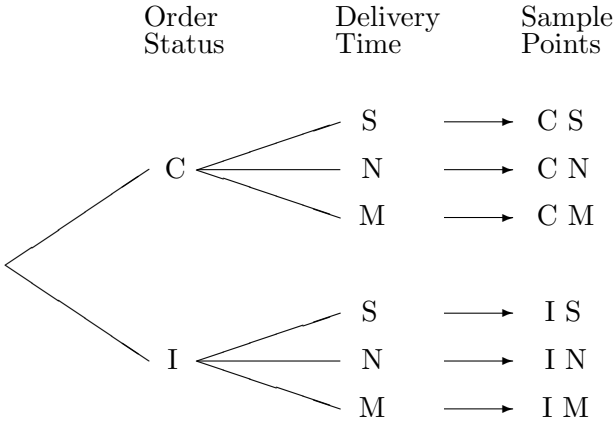


A more interesting example might be the parts delivery operation of a firm supplying parts for oil drilling rigs operating world wide. The relevant random trial is the delivery of a part. Two characteristics of the experiment are of interest—first, whether the correct part was delivered and second, the number of days it took to get the part to the drilling site. This is also a bivariate random trial the essence of which can be captured in the following table:

		Time of Delivery		
		S	N	M
Order	C	C S	C N	C M
Status	I	I S	I N	I M

The status of the order has two categories: 'correct part' (C) and 'incorrect part' (I). The time of delivery has three categories: 'same day' (S), 'next day' (N) and 'more than one day' (M). There are six sample points or basic outcomes. The top row in the table gives the event 'correct part' and the bottom row gives the event 'incorrect part'. Each of these events contain three sample points. The first column on the left in the main body of the table gives the event 'same day delivery', the middle column the event 'next day delivery' and the third column the event 'more than one day delivery'. These three events each contain two sample points or basic outcomes. The event 'correct part delivered in less than two days' would be the left-most two sample points in the first row, (C S) and (C N). The complement of that event, 'wrong part or more than one day delivery' would be the remaining outcomes (C M), (I S), (I N) and (I M).

Notice also that the basic outcome in each cell of the above table is the intersection of two events—(C S) is the intersection of the event C or 'correct part' and the event S 'same day delivery' and (I N) is the intersection of the event I, 'incorrect part', and the event N 'next day delivery'. The event 'correct part' is the union of three simple events, (C S) $\cup$ (C N) $\cup$ (C M). The parts delivery sample space can also be expressed in tree form as follows:



2.4 The Meaning of Probability

Although probability is a term that most of us used before we began to study statistics, a formal definition is essential. As we noted above, a random trial is an experiment whose outcome must be one of a set of sample points with uncertainty as to which one it will be. And events are collections of sample points, with an event occurring when one of the sample points or basic outcomes it contains occurs. *Probability* is a value attached to a sample point or event denoting the likelihood that it will be realized. These probability assignments to events in the sample space must follow certain rules.

1. The probability of any basic outcome or event consisting of a set of basic outcomes must be between zero and one. That is, for any outcome o_i or event E_i containing a set of outcomes we have

$$\begin{aligned} 0 \leq P(o_i) \leq 1 \\ 0 \leq P(E_i) \leq 1. \end{aligned} \tag{2.10}$$

If $P(o_i) = 1$ or $P(E_i) = 1$ the respective outcome or event is certain to occur; if $P(o_i) = 0$ or $P(E_i) = 0$ the outcome or event is certain not to occur. It follows that probability cannot be negative.

2. For any set of events of the sample space S (and of the event space E),

$$P(E_j) \quad = \quad \sum_{i=1}^{J} P(o_i). \tag{2.11}$$

where J is the number of basic events or sample points contained in the event E_j. In other words, the probability that an event will occur is the sum of the probabilities that the basic outcomes contained in that event will occur. This follows from the fact that an event is said to occur when one of the basic outcomes or sample points it contains occurs.

3. Since it is certain that at least one of the sample points or elementary events in the sample space will occur, $P(S) = 1$. And the null event cannot occur, so $P(\phi) = 0$ where ϕ is the null event. These results follow from the fact that $P(S)$ is the sum of the probabilities of all the simple or basic events.

A number of results follow from these postulates

- $P(E_i) \leq P(E_j)$ when E_i is a subset of (contained in) E_j.
- If E_i and E_j are mutually exclusive events of a sample space, then $P(E_i \cap E_j) = 0$. That is, both events cannot occur at the same time.

Probability can be expressed as an *odds ratio*. If the probability of an event E_j is a, then the odds of that event occurring are a to $(1 - a)$. If the probability that you will get into an accident on your way home from work tonight is .2, then the odds of you getting into an accident are .2 to .8 or 1 to 4. If the odds in favour of an event are a to b then the probability of the event occurring is

$$\frac{a}{a+b}$$

If the odds that your car will break down on the way home from work are 1 to 10, then the probability it will break down is $1/(10 + 1) = 1/11$.

2.5 Probability Assignment

How are probabilities established in any particular case? The short answer is that we have to assign them. The probability associated with a random trial or experiment can be thought of as a mass or "gob" of unit weight. We have to distribute that mass across the sample points or basic elements in the sample space. In the case of a single coin toss, this is pretty easy to do. Since a fair coin will come up heads half the time and tails half the time we will assign half of the unit weight to H and half to T, so that the probability of a head on any toss is .5 and the probability of a tail is .5. In the case where we flip two coins simultaneously our intuition tells us that each of the four sample points HH, HT, TH, and TT are equally likely, so we would assign a quarter of the mass, a probability of .25, to each of them. When it comes to determining the probability that I will be hit by a car on my way home from work tonight, I have to make a wild guess on the basis of information I might have on how frequently that type of accident occurs between 5 o'clock and 6 o'clock on weekday afternoons in my neighbourhood and how frequently I jay-walk. My subjective guess might be that there is about one chance in a thousand that the elementary event 'get hit by a car on my way home from work' will occur and nine-hundred and ninety-nine chances in a thousand that the mutually exclusive elementary event 'do not get hit by a car on my way home from work' will occur. So I assign a probability of .001 to the event 'get hit' and a probability of .999 to the

event 'not get hit'. Note that the implied odds of me getting hit are 1 to 999.

As you might have guessed from the above discussion the procedures for assigning probabilities fall into two categories—objective and subjective. In the case of coin tosses we have what amounts to a mathematical model of a fair coin that will come up heads fifty percent of the time. If the coin is known to be fair this leads to a purely objective assignment of probabilities—no personal judgement or guesswork is involved. Of course, the proposition that the coin is fair is an assumption, albeit a seemingly reasonable one. Before assigning the probabilities in a coin toss, we could toss the coin a million times and record the number of times it comes up heads. If it is a fair coin we would expect to count 500,000 heads. In fact, we will get a few more or less than 500,000 heads because the one million tosses is still only a sample, albeit a large one, of an infinite population. If we got only 200,000 heads in the 1,000,000 tosses we would doubt that the coin was a fair one. A theoretically correct assignment of probabilities would be one based on the frequencies in which the basic outcomes occur in an infinite sequence of experiments where the conditions of the experiment do not change. This uses a basic axiom of probability theory called the *law of large numbers*. The law states essentially that the relative frequency of occurrence of a sample point approaches the theoretical probability of the outcome as the experiment is repeated a larger and larger number of times and the frequencies are cumulated over the repeated experiments. An example is shown in Figure 2.1 where a computer generated single-coin toss is performed 1500 times. The fraction of tosses turning up heads is plotted against the cumulative number of tosses measured in hundreds.

In practice, the only purely objective method of assigning probabilities occurs when we know the mathematics of the data generating process—for example, the exact degree of 'fairness' of the coin in a coin toss. Any non-objective method of assigning probabilities is a subjective method, but subjective assignments can be based on greater or lesser amounts of information, according to the sample sizes used to estimate the frequency of occurrence of particular characteristics in a population. When relative frequencies are used to assign probabilities the only subjective component is the choice of the data set from which the relative frequencies are obtained. For this reason, the assignment of probabilities based on relative frequencies is often also regarded as objective. In fact, inferential statistics essentially involves the use of sample data to try to infer, as objectively as possible, the proximate probabilities of events in future repeated experiments or random draws from a population. Purely subjective assignments of probabilities are

those that use neither a model of the data-generating process nor data on relative frequencies.

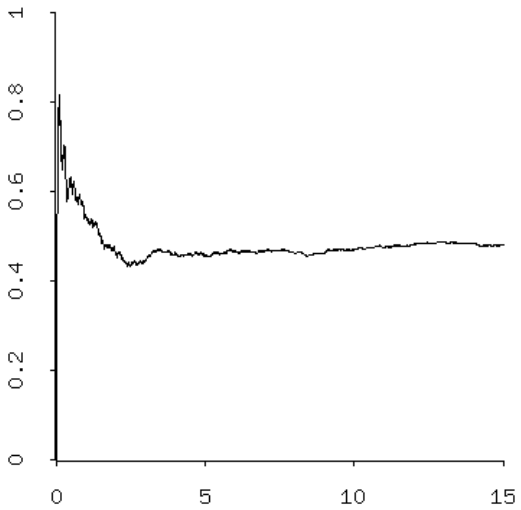


Figure 2.1: Illustration of the law of large numbers. Computer generated plot of the cumulative fraction of 1500 single coin-tosses turning up heads. The horizontal axis gives the number of tosses in hundreds and the vertical axis the fraction turning up heads.

Purely subjective probability measures tend to be useful in business situations where the person or organization that stands to lose from an incorrect assignment of probabilities is the one making the assignment. If I attach a probability of 0.1 that a recession will occur next year and govern my investment plans accordingly, I am the one who stands to gain or lose if the event 'recession' occurs and I will be the loser over the long run if my probability assignments tend to be out of line with the frequency with which the event occurs. Since I stand to gain or lose, my probability assessment is 'believable' to an outside observer—there is no strategic gain to me from 'rigging' it. On the other hand, if the issue in question is the amount my industry will lose from free trade, then a probability assignment I might make to the set of sample points comprising the whole range of losses that could be incurred should not be taken seriously by policy makers in deciding how much compensation, if any, to give to my industry. Moreover, outside ob-

servers' subjective probability assignments are also suspect because one does not know what their connection might happen to be to firms and industries affected by proposed policy actions.

2.6 Probability Assignment in Bivariate Sample Spaces

Probability assignment in bivariate sample spaces can be easily visualized using the following table, which further extends our previous example of world-wide parts delivery to oil drilling sites.

		Time of Delivery			
		S	N	M	Sum
Order	C	.600	.24	.120	.96
Status	I	.025	.01	.005	.04
	Sum	.625	.25	.125	1.00

Probabilities have been assigned to the six elementary events either purely subjectively or using frequency data. Those probabilities, represented by the numbers in the central enclosed rectangle must sum to unity because they cover the entire sample space—at least one of the sample points must occur. They are called *joint probabilities* because each is an intersection of two events—an 'order status' event (C or I) and a 'delivery time' event (S, N, or M). The probabilities in the right-most column and and along the bottom row are called *marginal probabilities*. Those in the right margin give the probabilities of the events 'correct' and 'incorrect'. They are the unions of the joint probabilities along the respective rows and they must sum to unity because the order delivered must be either correct or incorrect. The marginal probabilities along the bottom row are the probabilities of the events 'same day delivery' (S), 'next day delivery' (N) and 'more than one day to deliver' (M). They are the intersections of the joint probabilities in the respective columns and must also sum to unity because all orders are delivered eventually. You can read from the table that the probability of the correct order being delivered in less than two days is .60 + .24 = .84 and the probability of unsatisfactory performance (either incorrect order or two or more days to deliver) is (.12 + .025 + .01 + .005) = .16 = (1 - .84).

2.7 Conditional Probability

One might ask what the probability is of sending the correct order when the delivery is made on the same day. Note that this is different than the probability of both sending the correct order *and* delivering on the same day. It is the probability of getting the order correct *conditional upon* delivering on the same day and is thus called a *conditional probability*. There are two things that can happen when delivery is on the same day—the order sent can be correct, or the incorrect order can be sent. As you can see from the table a probability weight of .600 + .025 = .625 is assigned to same-day delivery. Of this probability weight, the fraction .600/.625 = .96 is assigned to the event 'correct order' and the fraction .25/.625 = .04 is assigned to the event 'incorrect order'. The probability of getting the order correct conditional upon same day delivery is thus .96 and we define the conditional probability as

$$P(C|S) = \frac{P(C \cap S)}{P(S)}. \tag{2.12}$$

where $P(C|S)$ is the probability of C occurring conditional upon the occurrence of S, $P(C \cap S)$ is the joint probability of C and S (the probability that both C and S will occur), and $P(S)$ is the marginal or unconditional probability of S (the probability that S will occur whether or not C occurs). The definition of conditional probability also implies, from manipulation of (2.12), that

$$P(C \cap S) = P(C|S)P(S). \tag{2.13}$$

Thus, if we know that the conditional probability of C given S is equal to .96 and that the marginal probability of C is .625 but are not given the joint probability of C and S, we can calculate that joint probability as the product of .625 and .96 —namely .600.

2.8 Statistical Independence

From application of (2.12) to the left-most column in the main body of the table we see that the conditional probability distribution of the event 'order status' given the event 'same day delivery' is

$$\begin{aligned} P(C|S) &\quad .96 \\ P(I|S) &\quad .04 \end{aligned}$$

which is the same as the marginal probability distribution of the event 'order status'. Further calculations using (2.12) reveal that the probability distributions of 'order status' conditional upon the events 'next day delivery' and 'more than one day delivery' are

$$\begin{aligned} P(C|N) &\quad .24/.25 = .96 \\ P(I|N) &\quad .01/.25 = .04 \end{aligned}$$

and

$$\begin{aligned} P(C|M) &\quad .120/.125 = .96 \\ P(I|M) &\quad .005/.125 = .04 \end{aligned}$$

which are the same as the marginal or unconditional probability distribution of 'order status'. Moreover, the probability distributions of 'time of delivery' conditional upon the events 'correct order' and 'incorrect order' are, respectively

$$\begin{aligned} P(S|C) &\quad .60/.96 = .625 \\ P(N|C) &\quad .24/.96 = .25 \\ P(M|C) &\quad .12/.96 = .125 \end{aligned}$$

and

$$\begin{aligned} P(S|I) &\quad .025/.04 = .625 \\ P(N|I) &\quad .010/.04 = .25 \\ P(M|I) &\quad .005/.04 = .125 \end{aligned}$$

which are the same as the marginal or unconditional probability distribution of 'time of delivery'. Since the conditional probability distributions are the same as the corresponding marginal probability distributions, the probability of getting the correct order is the same whether delivery is on the

same day or on a subsequent day—that is, independent of the day of delivery. And the probability of delivery on a particular day is independent of whether or not the order is correctly filled. Under these conditions the two events 'order status' and 'time of delivery' are said to be *statistically independent*. Statistical independence means that the marginal and conditional probabilities are the same, so that

$$P(C|S) = P(C). \tag{2.14}$$

The case where two events are not statistically independent can be illustrated using another example. Suppose that we are looking at the behaviour of two stocks listed on the New York Stock Exchange—Stock A and Stock B—to observe whether over a given interval the prices of the stocks increased, decreased or stayed the same. The sample space, together with the probabilities assigned to the sample points based on several years of data on the price movements of the two stocks can be presented in tabular form as follows:

Stock B		Stock A			
		Increase	No Change	Decrease	
		A_1	A_2	A_3	Sum
Increase	B_1	.20	.05	.05	.30
No Change	B_2	.15	.10	.15	.40
Decrease	B_3	.05	.05	.20	.30
	Sum	.40	.20	.40	1.00

The conditional probability that the price of stock A will increase, given that the price of stock B increases is

$$\begin{aligned} P(A_1|B_1) &= \frac{P(A_1 \cap B_1)}{P(B_1)} \\ &= \frac{.20}{.30} = .666 \end{aligned}$$

which is greater than the unconditional probability of an increase in the price of stock A, the total of the A_1 column, equal to .4. This says that the probability that the price of stock A will increase is greater if the price of stock B also increases. Now consider the probability that the price of stock A will fall, conditional on a fall in the price of stock B. This equals

$$\begin{aligned} P(A_3|B_3) &= \frac{P(A_3 \cap B_3)}{P(B_3)} \\ &= \frac{.20}{.30} = .666 \end{aligned}$$

which is greater than the 0.4 unconditional probability of a decline in the price of stock A given by the total at the bottom of the A_3 column. The probability that the price of stock A will decline conditional upon the price of stock B not declining is

$$\begin{aligned} \frac{P(A_3 \cap B_1) + P(A_3 \cap B_2)}{P(B_1) + P(B_2)} &= \frac{.05 + .15}{.30 + .40} \\ &= \frac{20}{70} = .286 \end{aligned}$$

which is smaller than the 0.4 unconditional probability of the price of stock A declining regardless of what happens to the price of stock B. The price of stock A is more likely to decline if the price of stock B declines and less likely to decline if the price of stock B does not decline. A comparison of these conditional probabilities with the relevant unconditional ones make it clear that the prices of stock A and stock B move together. They are not statistically independent.

There is an easy way to determine if the two variables in a bivariate sample space are statistically independent. From the definition of statistical independence (2.14) and the definition of conditional probability as portrayed in equation (2.13) we have

$$P(C \cap S) = P(C|S)P(S) = P(C)P(S). \tag{2.15}$$

This means that when there is statistical independence the joint probabilities in the tables above can be obtained by multiplying together the two relevant marginal probabilities. In the delivery case, for example, the joint probability of 'correct order' and 'next day' is equal to the product of the two marginal probabilities .96 and .25, which yields the entry .24. The variables 'order status' and 'time of delivery' are statistically independent. On the other hand, if we multiply the marginal probability of A_1 and the marginal probability of B_1 in the stock price change example we obtain $.30 \times .40 = .12$ which is less than .20, the actual entry in the joint probability distribution table. This indicates that the price changes of the two stocks are not statistically independent.

2.9 Bayes Theorem

Many times when we face a problem of statistical inference about a population from a sample, we already have some information prior to looking at the sample. Suppose, for example, that we already know that the probabilities that an offshore tract of a particular geological type contains no gas (A_1), a minor gas deposit (A_2) or a major gas deposit (A_3) are .7, .25 and .05 respectively. Suppose further that we know that a test well drilled in a tract like the one in question will yield no gas (B_1) if none is present and will yield gas (B_2) with probability .3 if a minor deposit is present and with probability .9 if a major deposit is present. A sensible way to proceed is to begin with the information contained in the probability distribution of gas being present in the tract and then upgrade that probability distribution on the basis of the results obtained from drilling a test well. Our procedure can be organized as follows:

	Prior Probability $P(A_i)$	$P(B_2\|A_i)$	Joint Probability $P(A_i \cap B_2)$	Posterior Probability $P(A_i\|B_2)$
No Gas (A_1)	0.70	0.00	0.000	0.000
Minor Deposit (A_2)	0.25	0.30	0.075	0.625
Major Deposit (A_3)	0.05	0.90	0.045	0.375
Total	1.00		0.120	1.000

Suppose that our test well yields gas (otherwise it's game over!). We begin with our prior probabilities $P(A_i)$ and then use the fact that the joint probability distribution $P(B_2 \cap A_i)$ equals the prior probabilities multiplied by the conditional probabilities $P(B_2|A_i)$ that gas will be obtained, given the respective A_i,

$$P(B_2 \cap A_i) = P(B_2|A_i)P(A_i).$$

These probabilities are entered in the second column from the right. Their sum gives the probability of finding gas, which equals .12 (the probability of finding gas and there being no gas (0.000) plus the probability of finding gas and there being a minor deposit (0.075) plus the probability of finding gas and there being a major deposit (0.045)). It then follows that the probability of there being no gas conditional upon gas being found in the test well is $0.000/.12 = 0.000$, the probability of there being a minor deposit conditional upon the test well yielding gas is $.075/.12 = .625$ and the probability of there being a major deposit conditional upon gas being found in the test well is

.045/.12 = .375. Since the test well yielded gas, these latter probabilities are the posterior (post-test or post-sample) probabilities of there being no gas, a minor deposit and a major deposit. They are entered in the column on the extreme right. When we are finished we can say that there is a .625 probability that the tract contains a minor gas deposit and a .375 probability that it contains a major deposit.

Notice what we have done here. We have taken advantage of the fact that the joint probability distribution $P(A_i \cap B_j)$ can be obtained in two ways:

$$P(A_i \cap B_j) = P(A_i|B_j)\, P(B_j)$$

and

$$P(A_i \cap B_j) = P(B_j|A_i)\, P(A_i).$$

Subtracting the second of these from the first, we obtain

$$P(A_i|B_j)\, P(B_j) \quad = \quad P(B_j|A_i)\, P(A_i)$$

which implies

$$P(A_i|B_j) \quad = \quad P(B_j|A_i)\, \frac{P(A_i)}{P(B_j)} \tag{2.16}$$

We can then use the fact that

$$P(B_j) \quad = \quad \sum_i P(B_j \cap A_i) \quad = \quad \sum_i [P(B_j|A_i)\, P(A_i)] \tag{2.17}$$

to express (2.16) as

$$P(A_i|B_j) \quad = \quad \frac{P(B_j|A_i)\, P(A_i)}{\sum_i [P(B_j|A_i)\, P(A_i)]} \tag{2.18}$$

This latter equation is called *Bayes Theorem*. Given the prior probability distribution $P(A_i)$ (the marginal or unconditional probabilities of gas being present) plus the conditional probability distribution $P(B_j|A_i)$ (the probabilities of finding gas conditional upon it being not present, present in a minor deposit or present in a major deposit), we can calculate the posterior probability distribution (probabilities of no deposit or a minor or major deposit being present conditional upon the information obtained from drilling a test hole).

The operation of Bayes Theorem can perhaps best be understood with reference to a tabular delineation of the sample space of the sort used in the parts delivery case.

Type of Deposit	Test Drill No Gas (B_1)	Test Drill Gas (B_2)	Prior Probability Distribution
(A_1)		0.000	0.70
(A_2)		0.075	0.25
(A_3)		0.045	0.05
Total		0.120	1.00

On the basis of our previous calculations we are able to fill in the right-most two columns. The column on the extreme right gives the prior probabilities and the second column from the right gives the joint probabilities obtained by multiplying together the prior probabilities and the probabilities of finding gas in a test well conditional upon its absence or minor or major presence in the tract. We can fill in the missing column by subtracting the second column from the right from the right-most column. This yields

Type of Deposit	Test Well No Gas (B_1)	Test Well Gas (B_2)	Prior Probability Distribution
(A_1)	0.700	0.000	0.70
(A_2)	0.175	0.075	0.25
(A_3)	0.005	0.045	0.05
Total	0.880	0.120	1.00

We can now see from the bottom row that the probability of not finding gas in a test well drilled in this type of tract is .88. The posterior probabilities conditional upon finding no gas or gas, respectively, in the test well can be calculated directly from the table by taking the ratios of the numbers in the two columns to the unconditional probabilities at the bottoms of those columns. The posterior probabilities are therefore

Type of Deposit	Posterior Probabilities No Gas ($A_i\|B_1$)	Posterior Probabilities Gas ($A_i\|B_2$)	Prior Probability Distribution
(A_1)	0.795	0.000	0.70
(A_2)	0.199	0.625	0.25
(A_3)	0.006	0.375	0.05
Total	1.000	1.000	1.00

Notice how the prior probabilities are revised as a consequence of the test results. The prior probability of no gas being present is .70. If the test well

yields no gas, that probability is adjusted upward to .795 and if the test well yields gas it is adjusted downward to zero. The prior probability that there is a minor deposit in the tract is .25. If the test well yields no gas this is adjusted downward to less than .2 while if gas is found in the test well this probability is adjusted upward to .625. Note that it is possible for gas to be present even if the test well yields no gas (gas could be present in another part of the tract) while if there is no gas present the test well will not find any. Finally, the prior probability of there being a major deposit present is adjusted upward from .05 to .375 if the test well yields gas and downward to .006 if the test well finds no gas.

2.10 The AIDS Test

Now consider another application of Bayes Theorem. You go to your doctor for a routine checkup and he tells you that you have just tested positive for HIV. He informs you that the test you have been given will correctly identify an AIDS carrier 90 percent of the time and will give a positive reading for a non-carrier of the virus only 1 percent of the time. He books you for a second more time consuming and costly but absolutely definitive test for Wednesday of next week.

The first question anyone would ask under these circumstances is "Does this mean that I have a 90 percent chance of being a carrier of HIV." On the way home from the doctor's office you stop at the library and rummage through some medical books. In one of them you find that only one person per thousand of the population in your age group is a carrier of aids.[1] You think "Am I so unfortunate to be one of these?" Then you remember about Bayes Theorem from your statistics class and decide to do a thorough analysis. You arrange the sample space as follows

An HIV Carrier?		Test Result Positive (T_1)	Test Result Negative (T_0)	Prior Probability Distribution
No	(A_0)	0.0099		0.999
Yes	(A_1)	0.0009		0.001
	Total	0.0108		1.000

and make special note that the test results give you some conditional probabilities. In particular, the probability of a positive result conditional upon

[1] These numbers, indeed the entire scenario, should not be taken seriously—I am making everything up as I go along!

you being a carrier is $P(T_1|A_1) = .90$ and the probability of a positive result conditional upon you not being a carrier is $P(T_1|A_0) = .01$. You obtain the joint probability of being a carrier and testing positive by multiplying $P(T_1|A_1)$ by $P(A_1)$ to obtain $.90 \times .001 = .0009$ and enter it into the appropriate cell of the above table. You then obtain the joint probability of testing positive and not being a carrier by multiplying $P(T_1|A_0)$ by $P(A_0)$. This yields $.01 \times .999 = .0099$ which you enter appropriately in the above table. You then sum the numbers in that column to obtain the unconditional probability of testing positive, which turns out to be .0108. You can now calculate the posterior probability—that is, the probability of being a carrier conditional on testing positive. This equals $.0009/.0108 = .08333$. The information from the test the doctor gave you has caused you to revise your prior probability of .001 upward to .0833. You can now fill in the rest of the table by subtracting the joint probabilities already there from the prior probabilities in the right margin.

An HIV Carrier?		Test Result Positive (T_1)	Test Result Negative (T_0)	Prior Probability Distribution
No	(A_0)	0.0099	.9891	0.999
Yes	(A_1)	0.0009	.0001	0.001
	Total	0.0108	.9892	1.000

Notice the importance to this problem of the 1% conditional probability of testing positive when you don't carry HIV. If that conditional probability were zero then the fact that the test will come up positive for a carrier 90% of the time is irrelevant. The joint probability of testing positive and not being a carrier is zero. A carrier of HIV will sometimes test negative but a non-carrier will never test positive. The above tabular representation of the bivariate sample space then becomes

An HIV Carrier?		Test Result Positive (T_1)	Test Result Negative (T_0)	Prior Probability Distribution
No	(A_0)	0.0000	.999	0.999
Yes	(A_1)	0.0009	.0001	0.001
	Total	0.0009	.9991	1.000

The probability that you carry HIV conditional upon testing positive is now $.0009/.0009 = 1.000$. You are a carrier.

2.11 Basic Probability Theorems

This chapter concludes with a statement of some basic probability theorems, most of which have already been motivated and developed and all of which will be used extensively in the chapters that follow. These theorems are best understood with reference to the Venn diagram presented in Figure 2.2. The area inside the square denotes the sample space with each point representing a sample point. The circular areas E_1, E_2 and E_3 represent events—the points inside these areas are those points belonging to the sample space contained in the respective events. The letters A, B, C and D denote collections of sample points inside the respective events. For example the event E_1 consists of $A + B$, event E_2 consists of $B + C$. And the area D represents event E_3. The probability theorems below apply to any two events of a sample space.

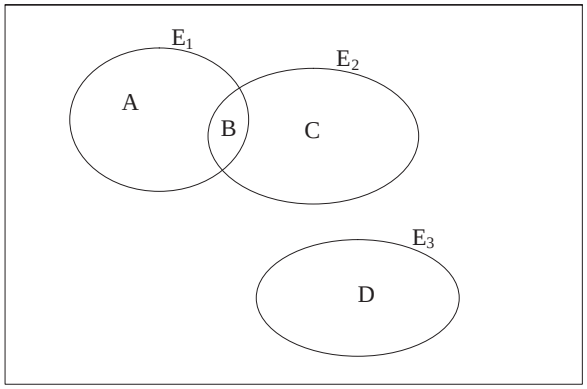


Figure 2.2: Venn diagram to illustrate basic probability theorems. The rectangle contains the sample space and the circular areas denote events E_1, E_2 and E_3.

1. **Addition**

$$P(E_1 \cup E_2) \quad = \quad P(E_1) + P(E_2) - P(E_1 \cap E_2) \tag{2.19}$$

The probabilities of the two events are added together and then the joint probability of the two events, given by the probability mass associated with the area B in Figure 2.2 is subtracted out to avoid double counting. If the events are mutually exclusive, as in the case of E_1 and E_3 the joint probability term will be zero,

$$P(E_1 \cup E_3) \quad = \quad P(E_1) + P(E_3). \tag{2.20}$$

2. **Complementation**

$$P(E_1) \quad = \quad 1 - P(E_1^c) \tag{2.21}$$

where E_1^c is the complementary event to E_1.

3. **Multiplication**

$$P(E_1 \cap E_2) \quad = \quad P(E_1)\, P(E_2|E_1). \tag{2.22}$$

This follows from the definition of conditional probability. In Figure 2.2, $P(E_2|E_1) = P(B)/(P(A) + P(B))$ (the proportion of the total weight in E_1 that also falls in E_2). And $P(E_1) = P(A) + P(B)$. So $P(E_1 \cap E_2) = [P(A) + P(B)][P(B)/(P(A) + P(B))] = P(B)$. If we know the joint probability and the marginal probability we can find the conditional probability. Similarly, if we know the conditional probability and the marginal probability we can find the joint probability.

2.12 Exercises

1. Suppose a random trial has three basic outcomes: o_1, o_2 and o_3. The probabilities of o_2 and o_3 are .5 and .4 respectively. Let E be the event consisting of basic outcomes o_2 and o_3. The probability of the complementary event to E is

a) .1

b) .9

c) .8

d) .2

e) none of the above.

2. Two marbles are drawn at random and without replacement from a box containing two blue marbles and three red marbles. Determine the probability of observing the following events.

a) Two blue marbles are drawn.

b) A red and a blue marble are drawn.

c) Two red marbles are drawn.

Hint: Organize the sample space according to a tree-diagram and then attach probabilities to the respective draws. Alternatively, you can organize the sample space in rectangular fashion with one draw represented as rows and the other as columns.

3. Three events, A, B, and C are defined over some sample space S. Events A and B are independent. Events A and C are mutually exclusive. Some relevant probabilities are $P(A) = .04$, $P(B) = .25$, $P(C) = .2$ and $P(B|C) = .15$. Compute the values of $P(A \cup B)$, $P(A \cup C)$, $P(A \cup B \cup C)$ and $P(C|B)$.

4. An experiment results in a sample space S containing five sample points and their associated probabilities of occurrence:

s_1	s_2	s_3	s_4	s_5
.22	.31	.15	.22	.10

The following events have been defined

- $E_1 = \{s_1, s_3\}$.
- $E_2 = \{s_2, s_3, s_4\}$.
- $E_3 = \{s_1, s_5\}$.

Find each of the following probabilities:

a) $P(E_1)$.

b) $P(E_2)$.

c) $P(E_1 \cap E_2)$.

d) $P(E_1|E_2)$.

e) $P(E_2 \cap E_3)$.

f) $P(E_3|E_2)$.

Consider each pair of events E_1 and E_2, E_1 and E_3 and E_2 and E_3. Are any of these events statistically independent? Why or why not? Hint: Are the joint probabilities equal to the products of the unconditional probabilities?

5. Roulette is a very popular game in Las Vegas. A ball spins on a circular wheel that is divided into 38 arcs of equal length, bearing the numbers 00, 0, 1, 2, ... , 35, 36. The number of the arc on which the ball stops after each spin of the wheel is the outcome of one play of the game. The numbers are also coloured as follows:

Red: 1,3,5,7,9,12,14,16,18,19,21,23,25,27,30,32,34,36
Black: 2,4,6,8,10,11,13,15,17,20,22,24,26,28,29,31,33,35
Green: 00,0

Players may place bets on the table in a variety of ways including bets on odd, even, red, black, high, low, etc. Define the following events:

- A: Outcome is an odd number (00 and 0 are considered neither even nor odd).
- B: Outcome is a black number.
- C: Outcome is a low number, defined as one of numbers 1–18 inclusive.

a) What is the sample space here?
b) Define the event $A \cap B$ as a specific set of sample points.
c) Define the event $A \cup B$ as a specific set of sample points.
d) Find $P(A)$, $P(B)$, $P(A \cup B)$, $P(A \cap B)$ and $P(C)$.
e) Define the event $A \cap B \cap C$ as a specific set of sample points.
f) Find $P(A \cup B)$.
g) Find $P(A \cap B \cap C)$.
h) Define the event $A \cup B \cup C$ as a specific set of sample points.

6. A bright young economics student at Moscow University in 1950 criticized the economic policies of the great leader Joseph Stalin. He was arrested and sentenced to banishment for life to a work camp in the east. In those days 70 percent of those banished were sent to Siberia and 30 percent were sent to Mongolia. It was widely known that a major difference between Siberia and Mongolia was that fifty percent of the men in Siberia wore fur hats, while only 10 percent of the people in Mongolia wore fur hats. The student was loaded on a railroad box car without windows and shipped east. After

many days the train stopped and he was let out at an unknown location. As the train pulled away he found himself alone on the prairie with a single man who would guide him to the work camp where he would spend the rest of his life. The man was wearing a fur hat. What is the probability he was in Siberia? In presenting your answer, calculate all joint and marginal probabilities. Hint: Portray the sample space in rectangular fashion with location represented along one dimension and whether or not a fur hat is worn along the other.

7. On the basis of a physical examination and symptoms, a physician assesses the probabilities that the patient has no tumour, a benign tumour, or a malignant tumour as 0.70, 0.20, and 0.10, respectively. A thermographic test is subsequently given to the patient. This test gives a negative result with probability 0.90 if there is no tumour, with probability 0.80 if there is a benign tumour, and with probability 0.20 if there is a malignant tumour.

a) What is the probability that a thermographic test will give a negative result for this patient?

b) Obtain the posterior probability distribution for the patient when the test result is negative?

c) Obtain the posterior probability distribution for the patient when the test result is positive?

d) How does the information provided by the test in the two cases change the physician's view as to whether the patient has a malignant tumour?

8. A small college has a five member economics department. There are two microeconomists, two macroeconomists and one econometrician. The World Economics Association is holding two conferences this year, one in Istanbul and one in Paris. The college will pay the expenses of one person from the department for each conference. The five faculty members have agreed to draw two names out of a hat containing all five names to determine who gets to go to the conferences. It is agreed that the person winning the trip to the first conference will not be eligible for the draw for the second one.

a) What is the probability that the econometrician will get to go to a conference?

b) What is the probability that macroeconomists will be the attendees at both conferences?

c) What is the probability that the attendees of the two conferences will be from different fields of economics?

d) The econometrician argued that a rule should be imposed specifying that both attendees could not be from the same field. She was outvoted. Would the provision have increased the probability that the econometrician would get to attend a conference?

Hint: Use a rectangular portrayal of the sample space with persons who can be chosen in the first draw along one axis and persons who can be chosen in the second draw along the other. Then blot out the diagonal on grounds that the same person cannot be chosen twice.

9. There is a 0.8 probability that the temperature will be below freezing on any winter's day in Toronto. Given that the temperature is below freezing my car fails to start 15 percent of the time. Given that the temperature is above freezing my car fails to start 5 percent of the time. Given that my car starts, what is the probability that the temperature is below freezing?

10. If a baseball player is hitting .250 (i.e., if averages one hit per four times at bat), how many times will he have to come up to bat to have a 90% chance of getting a hit? Hint: Ask yourself what the probability is of not getting a hit in n times at bat. Then take advantage of the fact that the event 'getting at least one hit in n times at bat' is the complementary event to the event of 'not getting a hit in n times at bat'.

11. A particular automatic sprinkler system for high-rise apartment buildings, office buildings, and hotels has two different types of activation devices on each sprinkler head. One type has a reliability of .91 (i.e., the probability that it will activate the sprinkler when it should is .91). The other type, which operates independently of the first type, has a reliability of .87. Suppose a serious fire starts near a particular sprinkler head.

a) What is the probability that the sprinkler head will be activated?

b) What is the probability that the sprinkler head will not be activated?

c) What is the probability that both activation devices will work properly?

d) What is the probability that only the device with reliability .91 will work properly?

Hint: Again use a rectangular portrayal of the sample space with the events 'type 1 activation (yes, no)' on one axis and 'type 2 activation (yes, no)' on the other.

12. At every one of the Toronto BlueJay's home games, little Johnny is there with his baseball mit. He wants to catch a ball hit into the stands. Years of study have suggested that the probability is .0001 that a person sitting in the type of seats Johnny and his dad sit in will have the opportunity to catch a ball during any game. Johnny is just turned six years old before the season started. If he goes to every one of the 81 home games from the start of the current season until he is 15 years old, what is the probability that he will have the opportunity to catch a ball.

13. A club has 100 members, 30 of whom are lawyers. Within the club, 25 members are liars and 55 members are neither lawyers nor liars. What proportion of the lawyers are liars?

14. The following is the probability distribution for an exam where students have to choose one of two questions. The pass mark is 3 points or more.

	5	4	3	2	1
Q1	.1	.1	.1	.2	0.0
Q2	0.0	.2	.1	.1	.1

a) Derive the marginal marks probability distribution.

b) What is the probability that a randomly selected student will pass? (.6)

c) Given that a randomly selected student got 4 marks, what is the probability that she did question 2?

15. Suppose you are on a game show and you are given the opportunity to open one of three doors and receive what ever is behind it. You are told that behind one of the doors is a brand new Rolls Royce automobile and behind the other two doors are goats. You pick a particular door—say door number 1—and before the host of the show, who knows what is behind each door, opens that door he opens one of the other doors—say door number 3—behind which is a goat. He then gives you the opportunity to stay with door number 1, which you originally chose, or switch your choice to door 2. Should you switch?

Answer:

This is a classic puzzle in statistics having a level of difficulty much greater than questions usually asked at the beginning level. Accordingly an effort is made here to present a detailed answer. One approach to answering this question is to examine the expected returns to "holding" (staying with the door originally picked) and "switching" to the other unopened door. Let us call the door you initially pick, which ever one it is, door A. Two mutually exclusive events are possible:

1) The car is behind door A —call this event AY.
2) The car is not behind door A —call this event AN.

If your initial guess is right (which it will be 1/3 of the time) you win the car by holding and lose it by switching. If your initial guess is wrong (which it will be 2/3 of the time) the host, by opening the door with the goat behind, reveals to you the door the car will be behind. You win by switching and lose by holding. If contestants in this game always switch they will win the car 2/3 of the time because their initial pick will be wrong 2/3 of the time. The expected payoff can be shown in tabular form. Let winning the car have a payoff of 1 and not winning it have a payoff of zero.

	Hold	Switch	Probability
AY	1	0	1/3
AN	0	1	2/3
Expected Payoff	$1/3 \times 1$ $+1/3 \times 0 = 1/3$	$1/3 \times 0$ $+2/3 \times 1 = 2/3$	

An alternative way to view the question is as a problem in Bayesian updating. Call the door you initially pick door A, the door the host opens door B, and the door you could switch to door C. On each play of the game the particular doors assigned the names A, B, and C will change as the doors picked by the contestant and opened by the host are revealed. The probabilities below are the probabilities that the car is behind the door in question.

Door	AY	AN	
	A	B	C
Prior Probability	1/3	1/3	1/3
Information From Host		$P(B\|AN) = 0$	$P(C\|AN) = 1$
Joint Probability		$P(B \cap AN) =$ $P(B\|AN)(P(AN)) = 0$	$P(C \cap AN) =$ $P(C\|AN)(P(AN)) = 2/3$
Posterior Probability	1/3	0	2/3

Keep in mind in looking at the above table that $P(AN) = P(B) + P(C) = 2/3$. The posterior probability of the car being behind the door the host leaves closed (i.e. the probability that it is behind door C conditional upon it not being behind door B) is $2/3$. The posterior probability of the car being behind door A (i.e., the probability of it being behind door A conditional upon it not being behind door B) is $1/3$, the same as the prior probability that it was behind door A. You should always switch!

Chapter 3

Some Common Probability Distributions

3.1 Random Variables

Most of the basic outcomes we have considered thus far have been non-numerical characteristics. A coin comes up either heads or tails; a delivery is on the same day with the correct order, the next day with the incorrect order, etc. We now explicitly consider random trials or experiments that relate to a quantitative characteristic, with a numerical value associated with each outcome. For example, patients admitted to a hospital for, say, X days where $X = 1, 2, 3, 4, \ldots$. Canada's GNP this year will be a specific number on the scale of numbers ranging upwards from zero. When the outcomes of an experiment are particular values on a natural numerical scale we refer to these values as a *random variable*. More specifically, a random variable is a variable whose numerical value is determined by the outcome of a random trial or experiment where a unique numerical value is assigned to each sample point.

Random variables may be *discrete* as in the length of hospital stay in days or *continuous* as in the case of next month's consumer price index or tomorrow's Dow Jones Industrial Average, the calculated values of which, though rounded to discrete units for reporting, fall along a continuum. The essential distinction between discrete and continuous random variables is that the sample points can be enumerated (or listed in quantitative order) in the case of a discrete random variable—for example, we can list the number of potential days of a hospital stay.[1] In the case of continuous random variables it

[1]Hospital stays could also be treated as a continuous variable if measured in fractions

is not possible to list the sample points in quantitative order—next month's consumer price index, for example, could be 120.38947 or 120.38948 or it could take any one of an infinity of values between 120.38947 and 120.38948. The number of sample points for a continuous random variable is always infinite. For a discrete random variable the number of sample points may or may not be infinite, but even an infinity of sample points could be listed or enumerated in quantitative order although it would take an infinite length of time to list them all. In the case of a continuous random variable any sample points we might put in a list cannot possibly be next to each other—between any two points we might choose there will be an infinity of additional points.

3.2 Probability Distributions of Random Variables

The *probability distribution* for a discrete random variable X associates with each of the distinct outcomes x_i, $(i = 1, 2, 3, \ldots, k)$ a probability $P(X = x_i)$. It is also called the *probability mass function* or the *probability function*. The probability distribution for the hospital stay example is shown in the top panel of Figure 3.1. The *cumulative probability distribution* or *cumulative probability function* for a discrete random variable X provides the probability that X will be at or below any given value—that is, $P(X \leq x_i)$ for all x_i. This is shown in the bottom panel of Figure 3.1. Note that X takes discrete values in both panels so that the lengths of the bars in the top panel give the probabilities that it will take the discrete values associated with those bars. In the bottom panel the length of each bar equals the sum of the lengths of all the bars in the top panel associated with values of X equal to or less than the value of X for that bar.

A continuous random variable assumes values on a continuum. Since there are an infinity of values between any two points on a continuum it is not meaningful to associate a probability value with a point on that continuum. Instead, we associate probability values with intervals on the continuum. The *probability density function* of a continuous random variable X is a mathematical function for which the area under the curve corresponding to any interval is equal to the probability that X will take on a value in that interval. The probability density function is denoted by $f(x)$, which gives the probability density at x. An example is given in the top panel of Figure 3.2 with the shaded area being the probability that X will take a value between 6 and 7. Note that $f(x)$ is always positive.

---

of hours or days. They are normally measure discretely in days, however, with patients being in hospital 'for the day' if not released during a given period in the morning.

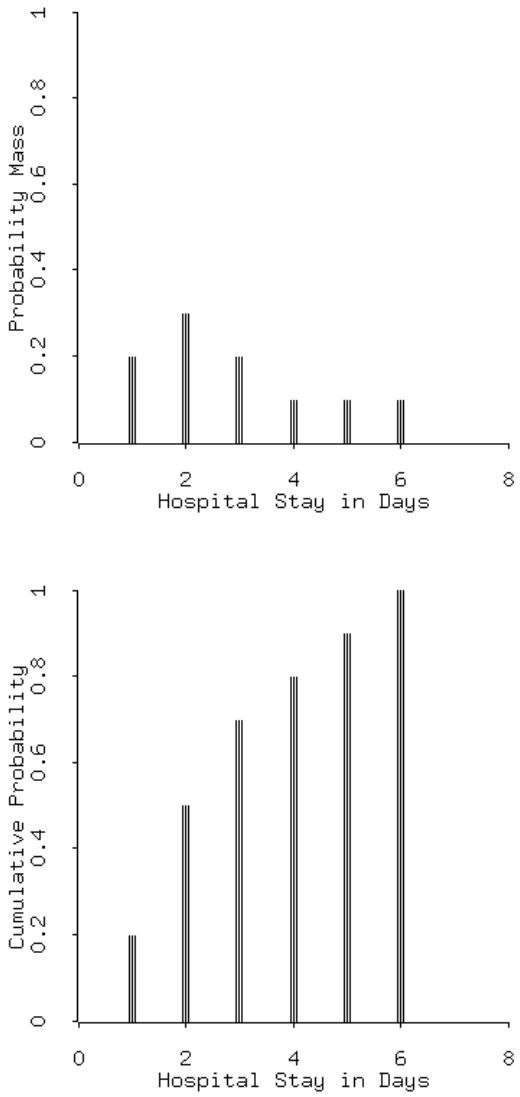


Figure 3.1: Probability mass function (top) and cumulative probability function (bottom) for the discrete random variable 'number of days of hospitalization'.

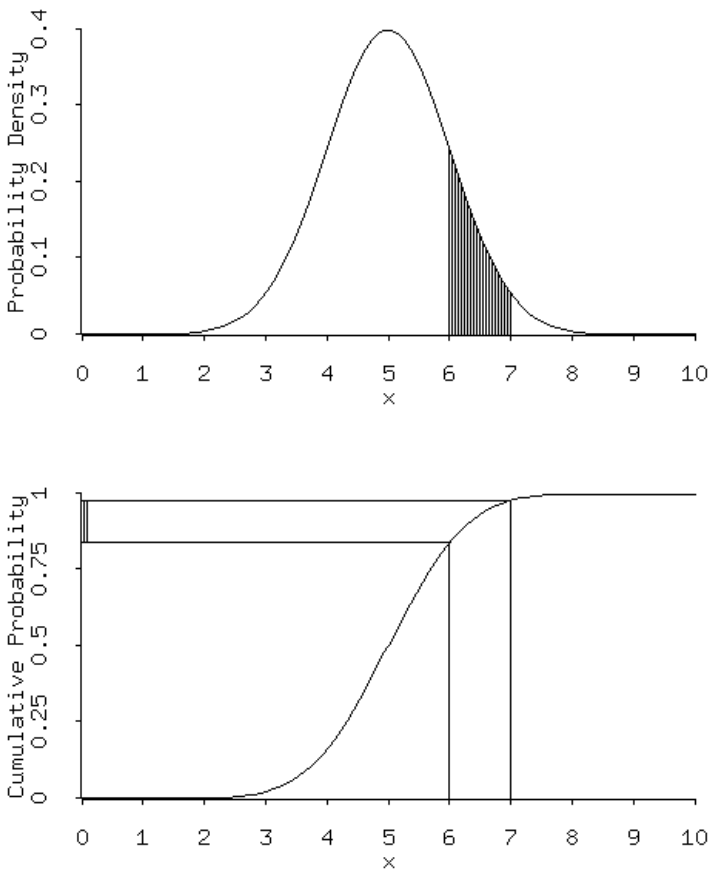


Figure 3.2: Probability density and cumulative probability functions for a continuouts random variable. The shaded area in the top panel equals the distance between the two vertical lines in the bottom panel.

The *cumulative probability function* of a continuous random variable X is denoted by $F(x)$ and is defined

$$F(x) = P(X \leq x) \tag{3.1}$$

where $-\infty \leq x \leq +\infty$. The cumulative probability function $F(x)$ gives the probability that the outcome of X in a random trial will be less than or equal to any specified value x. Thus, $F(x)$ corresponds to the area under

the probability density function to the left of x. This is shown in the bottom panel of Figure 3.2. In that panel, the distance between the two horizontal lines associated with the cumulative probabilities at $X \leq 6$ and $X \leq 7$ is equal to the shaded area in the top panel, and the distance of the lower of those two horizontal lines from the horizontal axis is equal to the area under the curve in the top panel to the left of $X = 6$. In mathematical terms we can express the probability function as

$$P(a \leq x \leq b) \quad = \quad \int_a^b f(x)\,dx \tag{3.2}$$

and the cumulative probability function as

$$P(X \leq x) \quad = \quad F(x) \quad = \quad \int_{-\infty}^{x} f(u)\,du \tag{3.3}$$

where u represents the variable of integration.

3.3 Expected Value and Variance

The mean value of a random variable in many trials is also known as its expected value. The *expected value of a discrete random variable* X is denoted by $E\{X\}$ and defined

$$E\{X\} \quad = \quad \sum_{i=1}^{k} x_i P(x_i) \tag{3.4}$$

where $P(x_i) = P(X = x_i)$. Since the process of obtaining the expected value involves the calculation denoted by $E\{\}$ above, $E\{\ \}$ is called the *expectation operator*.

Suppose that the probability distribution in the hospital stay example in Figure 3.1 above is

x:	1	2	3	4	5	6
$P(x)$:	.2	.3	.2	.1	.1	.1

The expected value of X is

$$\begin{aligned} E\{X\} \quad &= \quad (1)(.2) + (2)(.3) + (3)(.2) + (4)(.1) + (5)(.1) + (6)(.1) \\ &= \quad .2 + .6 + .6 + .4 + .5 + .6 \quad = \quad 2.9. \end{aligned}$$

Note that this result is the same as would result from taking the mean in the fashion outlined in Chapter 1. Let the probabilities be frequencies where

the total hospital visits is, say, 100. Then the total number of person-days spent in the hospital is

$$\begin{aligned} & (1)(20) + (2)(30) + (3)(20) + (4)(10) + (5)(10) + (6)(10) \\ = \; & 20 + 60 + 60 + 40 + 50 + 60 \;=\; 290 \end{aligned}$$

and the common mean is $290/100 = 2.9$. $E\{X\}$ is simply a weighted average of the possible outcomes with the probability values as weights. For this reason it is called the mean of the probability distribution of X. Note that the mean or expected value is a number that does not correspond to any particular outcome.

The *variance* of a discrete random variable X is denoted by $\sigma^2\{X\}$ and defined as

$$\sigma^2\{X\} \;=\; \sum_{i=1}^{k}(x_i - E\{X\})^2 P(x_i) \tag{3.5}$$

where $\sigma^2\{\,\}$ is called the *variance operator*. The calculation of the variance of the length of hospital stay can be organized in the table below:

x:	1	2	3	4	5	6
$P(x)$:	.20	.30	.20	.10	.10	.10
$x - E\{X\}$:	-1.90	-.90	.10	1.10	2.10	3.10
$(x - E\{X\})^2$:	3.61	.81	.01	1.21	4.41	9.61

from which

$$\begin{aligned} \sigma^2\{X\} \;&=\; (3.61)(.2) + (.81)(.3) + (.01)(.2) + (1.21)(.1) + (4.41)(.1) \\ &\quad +(9.61)(.1) \\ &=\; .722 + .243 + .002 + .121 + .441 + .961 \;=\; 2.49. \end{aligned}$$

The variance is a weighted average of the squared deviations of the outcomes of X from their expected value where the weights are the respective probabilities of occurrence. Thus $\sigma^2\{X\}$ measures the extent to which the outcomes of X depart from their expected value in the same way that the variance of the quantitative variables in the data sets examined in Chapter 1 measured the variability of the values about their mean. There is an important distinction, however, between what we are doing here and what we did in Chapter 1. In Chapter 1 we took an observed variable X and measured its observed variance. Here we are taking a *random variable* X and exploring the nature of its probability distribution.

Consider a random variable V for which $v_i = (x_i - E\{X\})^2$ in (3.5). Since each v_i has a corresponding x_i associated with it,

$$P(v_i) = P((x_i - E\{X\})^2) = P(x_i),$$

and (3.5) yields

$$\begin{aligned} \sigma^2\{X\} &= \sum_{i=1}^{k} v_i P(v_i) \\ &= E\{V\} = E\{(x_i - E\{X\})^2\}. \end{aligned} \tag{3.6}$$

The variance is simply the expectation of, or expected value of, the squared deviations of the values from their mean. The *standard deviation*, denoted by σ, is defined as the square root of the variance.

The discrete random variable X can be *standardised* or put in *standardised form* by applying the relationship

$$Z_i = \frac{X_i - E\{X\}}{\sigma\{X\}} \tag{3.7}$$

where the discrete random variable Z is the standardised form of the variable X. The variable Z is simply the variable X expressed in numbers of standard deviations from its mean. In the hospital stay example above the standardised values of the numbers of days of hospitalization are calculated as follows:

x:	1	2	3	4	5	6
$P(x)$:	.2	.3	.2	.1	.1	.1
$x - E\{X\}$:	-1.9	-.9	.1	1.1	2.1	3.1
$(x - E\{X\})^2$:	3.61	.81	.01	1.21	4.41	9.61
$(x - E\{X\})/\sigma\{X\}$:	-1.20	-.56	.06	.70	1.32	1.96

where $\sigma = \sqrt{2.49} = 1.58$.

The *expected value of a continuous random variable* is defined as

$$E\{X\} = \int_{-\infty}^{\infty} x f(x)\, dx. \tag{3.8}$$

This is not as different from the definition of the expected value of a discrete random variable in (3.4) as it might appear. The integral performs the same role for a continuous variable as the summation does for a discrete one. Equation (3.8) sums from minus infinity to plus infinity the variable x with

each little increment of x, given by dx, weighted by the probability $f(x)$ that the outcome of x will fall within that increment.[2] Similarly, the *variance of a continuous random variable* is defined as

$$\begin{aligned}\sigma^2\{X\} &= E\{(x - E\{X\})^2\} \\ &= \int_{-\infty}^{\infty} (x - E\{X\})^2 f(x)\, dx. \end{aligned} \tag{3.9}$$

In this equation the integral is taken over the probability weighted increments to $(x - E\{X\}^2)$ as compared to (3.8) where the integration is over the probability weighted increments to x.

Continuous random variables can be standardised in the same fashion as discrete random variables. The standardised form of the continuous random variable X is thus

$$Z = \frac{X - E\{X\}}{\sigma\{X\}}. \tag{3.10}$$

3.4 Covariance and Correlation

We noted in Chapter 1 that covariance and correlation are measures of the association between two variables. The variables in that case were simply quantitative data. Here we turn to an analysis of the covariance and correlation of two *random variables* as properties of their joint probability distribution. The *covariation* of the outcomes x_i and y_j of the discrete random variables X and Y is defined as

$$(x_i - E\{X\})(y_j - E\{Y\}).$$

The *covariance* of two random variables is the expected value of their covariation (i.e., their mean covariation after repeated trials). For two discrete random variables X and Y we thus have

$$\begin{aligned}\sigma\{X, Y\} &= E\{(x_i - E\{X\})(y_j - E\{Y\})\} \\ &= \sum_i \sum_j (x_i - E\{X\})(y_j - E\{Y\}) P(x_i, y_j) \end{aligned} \tag{3.11}$$

where $P(x_i, y_j)$ denotes $P(X = x_i \cap Y = y_j)$. We call $\sigma\{\ ,\ \}$ the *covariance operator*. Consider the following example:

[2] Notice that the definition of probability requires that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

	Y		
X	5	10	
2	.1	.4	.5
3	.3	.2	.5
	.4	.6	1.0

The two discrete random variables X and Y each take two values, 2 and 3 and 5 and 10 respectively. The four numbers in the enclosed square give the *joint probability distribution* of X and Y —that is, the probabilities

$$P(X = x_i \ \cap \ Y = y_j).$$

The numbers along the right and bottom margins are the marginal probabilities, which sum in each case to unity. On the basis of the earlier discussion it follows that

$$E\{X\} = (2)(.5) + (3)(.5) = 2.5$$

$$E\{Y\} = (5)(.4) + (10)(.6) = 8.0$$

$$\sigma^2\{X\} = (-.5^2)(.5) + (.5^2)(.5) = .25$$

and

$$\sigma^2\{Y\} = (-3^2)(.4) + (2^2)(.6) = 6.0$$

which renders $\sigma\{X\} = \sqrt{0.25} = .5$ and $\sigma\{Y\} = \sqrt{8} = 2.83$. The calculation of the covariance can be organized using the following table:

	$(X = 2 \cap Y = 5)$	$(X = 2 \cap Y = 10)$	$(X = 3 \cap Y = 5)$	$(X = 3 \cap Y = 10)$
$P(x_i, y_j)$	.1	.4	.3	.2
$(x_i - E\{X\})$	- .5	-.5	.5	.5
$(y_j - E\{Y\})$	- 3	2	- 3	2
$(x_i - E\{X\})(y_j - E\{Y\})$	1.5	-1	-1.5	1
$(x_i - E\{X\})(y_j - E\{Y\})P(x_i, y_j)$	.15	-.4	-.45	.2

The sum of the numbers in the bottom row gives

$$\sigma\{X, Y\} = \sum_i \sum_j (x_i - E\{X\})(y_j - E\{Y\})P(x_i, y_j) = -.5.$$

The *coefficient of correlation* of two random variables X and Y, denoted by $\rho\{X, Y\}$ is defined as

$$\rho\{X, Y\} = \frac{\sigma\{X, Y\}}{\sigma\{X\}\sigma\{Y\}}. \tag{3.12}$$

In the example above

$$\rho\{X, Y\} = -.5/((.5)(2.83)) = -.5/1.415 = -.35$$

which signifies a negative relationship between the two random variables. It is easy to show that the coefficient of correlation between X and Y is equivalent to the covariance between the standardised forms of those variables because the covariance of the standardised forms is the same as the covariance of the unstandardised variables and the standard deviations of the standardised forms are both unity. Thus, when the variables are standardised both the covariance and the correlation coefficient are unit free.

The *covariance of continuous random variables* X and Y is written

$$\begin{aligned} \sigma\{X, Y\} &= E\{(x - E\{X\})(y - E\{Y\})\} \\ &= \int\int (x - E\{X\})(y - E\{Y\}) f(x, y)\, dx\, dy \end{aligned} \tag{3.13}$$

where $f(x, y)$ is the *joint probability density function* of X and Y. The shape of a typical joint probability density function is portrayed graphically in Figure 3.3 (both variables are in standardised form). The *coefficient of correlation between continuous random variables* is defined by equation (3.12) with the numerator being (3.13) and the denominator the product of the standard deviations of X and Y obtained by taking the square roots of successive applications of (3.9).

When two variables are statistically independent both the covariance and correlation between them is zero. The opposite, however, does not follow. Zero covariance and correlation do not necessarily imply statistical independence because there may be a non-linear statistical relationship between two variables. An example is shown in Figure 3.4. The covariance and correlation between the two variables is zero, but they are obviously systematically related.

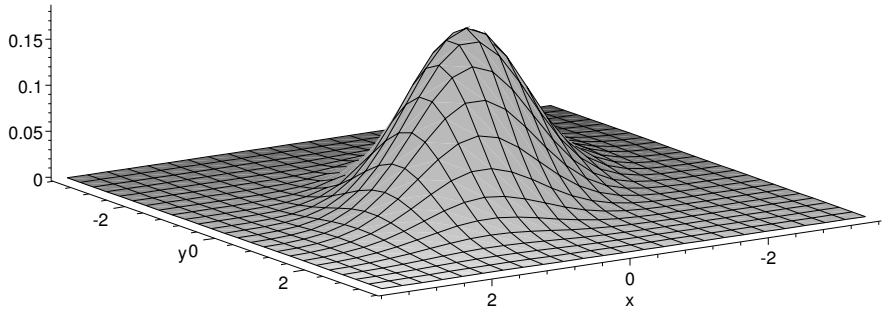


Figure 3.3: The joint probability density function of two continuous standardized random variables.

3.5 Linear Functions of Random Variables

Consider a linear function of the random variable X,

$$W = a + b\,X. \tag{3.14}$$

A number of relationships hold. First,

$$E\{W\} = E\{a + b\,X\} = E\{a\} + b\,E\{X\}, \tag{3.15}$$

which implies that

$$E\{a\} = a \tag{3.16}$$

and

$$E\{b\,X\} = b\,E\{X\}. \tag{3.17}$$

We can pass the expectation operator through a linear equation with the result that $E\{W\}$ is the same function of $E\{X\}$ as W is of X. Second,

$$\sigma^2\{W\} = \sigma^2\{a + b\,X\} = b^2\,\sigma^2\{X\} \tag{3.18}$$

which implies

$$\sigma^2\{a + X\} = \sigma^2\{X\}, \tag{3.19}$$

and

$$\sigma^2\{b\,X\} = b^2\,\sigma^2\{X\}. \tag{3.20}$$

This leads to the further result that

$$\sigma\{a + b\,X\} = |b|\,\sigma\{X\}. \tag{3.21}$$

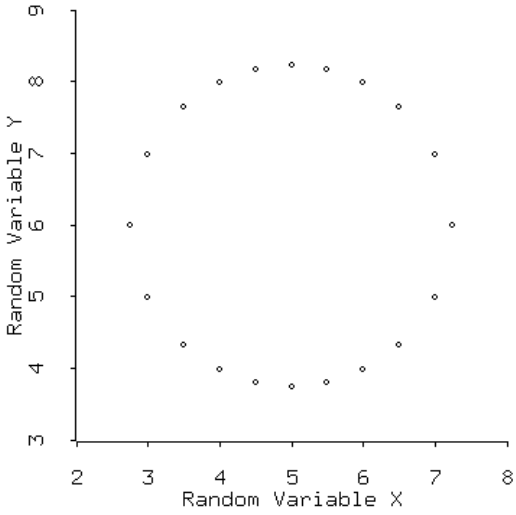


Figure 3.4: An example of two uncorrelated random variables that are not statistically independent.

3.6 Sums and Differences of Random Variables

If Z is the sum of two random variables X and Y, then the following two conditions hold:

$$E\{Z\} = E\{X+Y\} = E\{X\} + E\{Y\} \tag{3.22}$$

and

$$\sigma^2\{Z\} = \sigma^2\{X\} + \sigma^2\{Y\} + 2\,\sigma\{X,Y\}. \tag{3.23}$$

When Z is the difference between X and Y, these become

$$E\{Z\} = E\{X-Y\} = E\{X\} - E\{Y\} \tag{3.24}$$

and

$$\sigma^2\{Z\} = \sigma^2\{X\} + \sigma^2\{Y\} - 2\,\sigma\{X,Y\}. \tag{3.25}$$

To prove (3.23) and (3.25) we expand $\sigma^2\{Z\}$ using the definition of variance and the rules above:

$$\begin{aligned}
\sigma^2\{Z\} &= E\{(Z - E\{Z\})^2\} \\
&= E\{(X + Y - E\{X + Y\})^2\} \\
&= E\{((X - E\{X\}) + (Y - E\{Y\}))^2\} \\
&= E\{((X - E\{X\})^2 + 2\,(X - E\{X\})(Y - E\{Y\}) \\
&\qquad +(Y - E\{Y\})^2)\} \\
&= E\{(X - E\{X\})^2\} + 2\,E\{(X - E\{X\})(Y - E\{Y\})\} \\
&\qquad +E\{(Y - E\{Y\})^2\} \\
&= \sigma^2\{X\} + 2\,\sigma\{X, Y\} + \sigma^2\{Y\}. \qquad (3.26)
\end{aligned}$$

In the case where $Z = X - Y$ the sign of the covariance term changes but the variance of both terms remains positive because squaring a negative number yields a positive number.

When X and Y are statistically independent (and thus uncorrelated), $\sigma\{X, Y\} = 0$ and (3.23) and (3.25) both become

$$\sigma^2\{Z\} = \sigma^2\{X\} + \sigma^2\{Y\}.$$

More generally, if T is the sum of S *independent* random variables,

$$T = X_1 + X_2 + X_3 + \cdots + X_S,$$

where the X_i can take positive or negative values, then

$$E\{T\} = \sum_s^S E\{X_i\} \qquad (3.27)$$

and

$$\sigma^2\{T\} = \sum_s^S \sigma^2\{X_i\}. \qquad (3.28)$$

In concluding this section we can use the rules above to prove that the mean of a standardised variable is zero and its variance and standard deviation are unity. Let Z be the standardised value of X, that is

$$Z = \frac{X - E\{X\}}{\sigma\{X\}}.$$

Then

$$\begin{aligned} E\{Z\} &= E\left\{\frac{X - E\{X\}}{\sigma\{X\}}\right\} \\ &= \frac{1}{\sigma\{X\}} E\left\{X - E\{X\}\right\} \\ &= \frac{1}{\sigma\{X\}} \left(E\{X\} - E\{X\}\right) = 0 \end{aligned}$$

and

$$\begin{aligned} \sigma^2\{Z\} &= E\left\{\left(\frac{X - E\{X\}}{\sigma\{X\}} - 0\right)^2\right\} \\ &= E\left\{\left(\frac{X - E\{X\}}{\sigma\{X\}}\right)^2\right\} \\ &= \frac{1}{\sigma^2\{X\}} E\left\{(X - E\{X\})^2\right\} \\ &= \frac{\sigma^2\{X\}}{\sigma^2\{X\}} = 1. \end{aligned}$$

It immediately follows that $\sigma\{Z\}$ also is unity.

Finally, the correlation coefficient between two standardised random variables U and V will equal

$$\rho\{U, V\} = \frac{\sigma\{U, V\}}{\sigma\{U\}\,\sigma\{V\}} = \sigma\{U, V\}$$

since $\sigma\{U\}$ and $\sigma\{V\}$ are both unity.

3.7 Binomial Probability Distributions

We can think of many examples of random trials or experiments in which there are two basic outcomes of a qualitative nature—the coin comes up either heads or tails, the part coming off the assembly line is either defective or not defective, it either rains today or it doesn't, and so forth. These experiments are called *Bernoulli random trials*. To quantify these outcomes we arbitrarily assign one outcome the value 0 and the other the value 1. This random variable, $X_i = \{0, 1\}$ is called a *Bernoulli random variable*.

Usually we are interested in a whole sequence of random trials. In the process of checking the effectiveness of a process of manufacturing computer monitors, for example, we can let $X_i = 1$ if the ith monitor off the

line is defective and $X_i = 0$ if the ith monitor is not defective. The X_i, $(i = 1, 2, 3, \ldots, n)$ can then be viewed as a sequence of Bernoulli random variables. Such a sequence is called a *Bernoulli process*. Let $X_1, X_2, X_3, \ldots, X_n$ be a sequence of random variables associated with a Bernoulli process. The process is said to be *independent* if the X_i are statistically independent and *stationary* if every $X_i = \{0\,, 1\}$ has the same probability distribution. The first of these conditions means that whether or not, say, the 5th monitor off the assembly line is defective will have nothing to do with whether the 6th, 7th, 100th, 200th, or any other monitor is defective. The second condition means that the probability of, say, the 10th monitor off the line being defective is exactly the same as the probability that any other monitor will be defective—the X_i are *identically distributed*. The random variables in the sequence are thus *independently and identically distributed*.

In a sequence of Bernoulli random trials we are typically interested in the number of trials that have the outcome 1. The sum $X_1+X_2+X_3+\ldots+X_{300}$ would give the number of defective monitors in a sample of 300 off the line. The sum of n independent and identically distributed Bernoulli random variables, denoted by X,

$$X = X_1 + X_2 + X_3 + \ldots + X_n,$$

is called a *binomial random variable*. It can take $n+1$ values ranging from zero (when $X_i = 0$ for all i) to n (when $X_i = 1$ for all i). This random variable is distributed according to the *binomial probability distribution*.

The *binomial probability function*, which gives the probabilities that X will take values $(0, \ldots, n)$, is

$$P(x) \quad = \quad \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \tag{3.29}$$

where $P(x) = P(X = x)$, $x = 0, 1, 2, \ldots, n$, and $0 \leq p \leq 1$. The parameter p is the probability that $X_i = 1$. It is the same for all i because the Bernoulli random variables X_i are identically distributed. The term

$$\left(\begin{array}{c} n \\ x \end{array} \right)$$

represents a *binomial coefficient* which is defined as

$$\left(\begin{array}{c} n \\ x \end{array} \right) \quad = \quad \frac{n!}{x!(n-x)!}$$

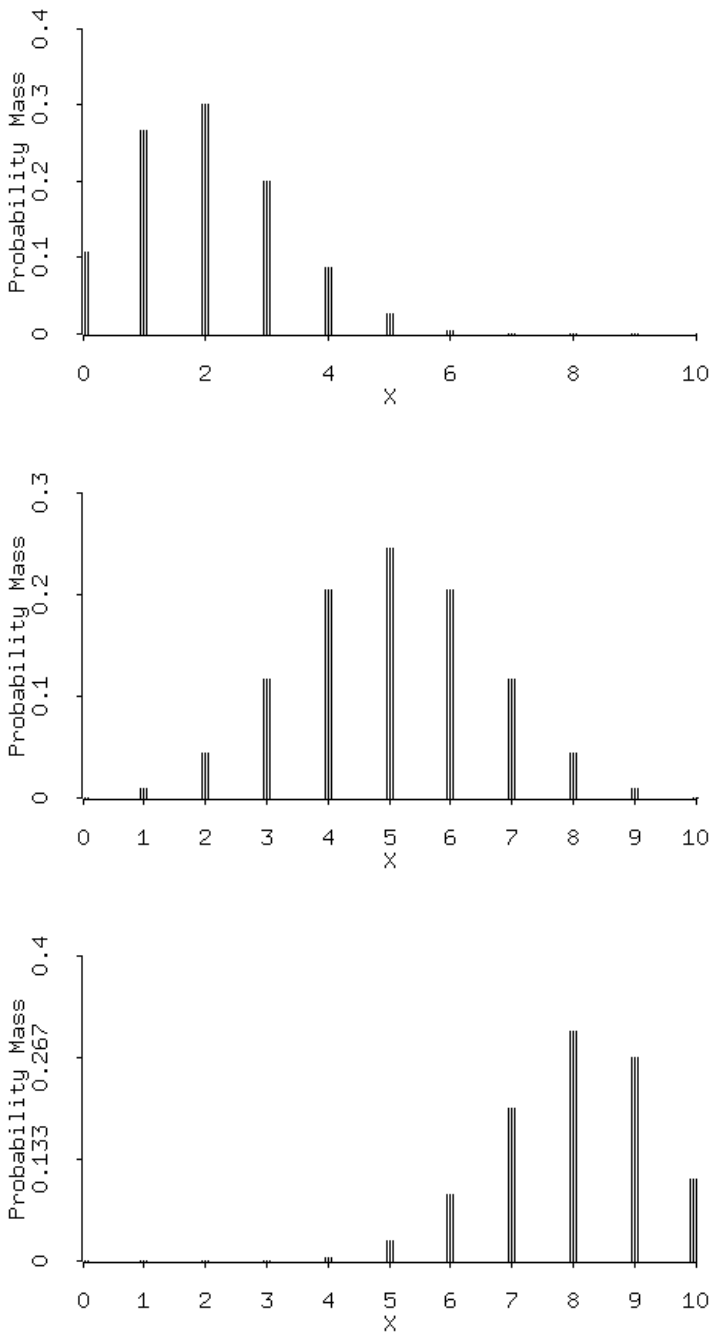


Figure 3.5: Binomial probability distributions with $n = 10$ and $p = .2$ (top), $p = .5$ (middle) and $p = .8$ (bottom.

where $a! = (a)(a-1)(a-2)(a-3)\ldots(1)$ and $0! = 1$.

The binomial probability distribution is a discrete probability distribution since X can only take the discrete values $0, 1, \ldots, n$. The parameters in the binomial probability distribution are p and n. Accordingly, there is a whole family of such distributions, one for each (p, n) combination. Figure 3.5 plots three examples—the distribution is skewed right if $p < .5$, skewed left if $p > .5$ and symmetrical if $p = .5$. The *mean of the binomial distribution* is

$$E\{X\} \quad = \quad np \tag{3.30}$$

and the variance is

$$\sigma^2\{X\} \quad = \quad np\,(1-p). \tag{3.31}$$

If we have two independent binomial random variables V and W with common probability parameter p and based on n_v and n_w trials, the sum $V + W$ is a binomial random variable with parameters p and $n = n_v + n_w$.

To more fully understand the workings of the binomial distribution consider the following problem. Four gauges are tested for accuracy. This involves four Bernoulli random trials $X_i = \{0, 1\}$ where 0 signifies that the gauge is accurate and 1 signifies that it is inaccurate. Whether or not any one of the four gauges is inaccurate has nothing to do with the accuracy of the remaining three so the X_i are statistically independent. The probability that each gauge is inaccurate is assumed to be .25. We thus have a binomial random variable X with $n = 4$ and $p = .25$. The sample space of X is

$$S = \{0, 1, 2, 3, 4\}.$$

Taking into account the fact that $n! = (4)(3)(2)(1) = 24$, the probability distribution can be calculated by applying equation (3.29) as follows:

x	$n!/(x!(n-x)!)$	p^x	$(1-p)^{n-x}$	$P(x)$
0	$24/(0!4!) = 1$	$.25^0 = 1.0000$	$.75^4 = .3164$	.3164
1	$24/(1!3!) = 4$	$.25^1 = .2500$	$.75^3 = .4219$	.4219
2	$24/(2!2!) = 6$	$.25^2 = .0625$	$.75^2 = .5625$	.2109
3	$24/(3!3!) = 4$	$.25^3 = .0156$	$.75^1 = .7500$	.0469
4	$24/(4!1!) = 1$	$.25^4 = .0039$	$.75^0 = 1.0000$	.0039
				1.0000

This probability distribution can be derived in a longer but more informative way by looking at the elementary events in the sample space and building up the probabilities from them. The four gauges are tested one after the other. There are 16 basic outcomes or sequences with probabilities attached to each sequence. The sequences are shown in Table 3.1. To see how the probabilities are attached to each sequence, consider sequence S_{12}. It consists of four outcomes of four independent and identically distributed Bernoulli random trials—0,1,0,0. The probability that 0 will occur on any trial is .75 and the probability that 1 will occur is .25. The probability of the four outcomes in the sequence observed is the product of the four probabilities. That is, the probability that first a 0 and then a 1 will occur is the probability of getting a 0 on the first trial times the probability of getting a 1 on the next trial. To obtain the probability that a sequence of 0,1,0 will occur we multiply the previously obtained figure by the probability of getting a zero. Then to get the probability of the sequence 0,1,0,0 we again multiply the previous figure by the probability of getting a zero. Thus the probability of the sequence S_{12} is

$$(.75)(.25)(.75)(.75) = (.25)^1(.75)^3 = .4219$$

which, it turns out, is the same as the probability of obtaining sequences S_8, S_{14} and S_{15}. Clearly, all sequences involving three zeros and a single one have the same probability regardless of the order in which the zeros and the one occur.

A frequency distribution of these sequences is presented in Table 3.2. There is one occurrence of no ones and four zeros, four occurrences of one and three zeros, six occurrences of two ones and two zeros, four occurrences of three ones and one zero, and one occurrence of four ones and no zeros. Thus, to find the probability that two ones and two zeros will occur we want the probability that any of the six sequences having that collection of ones and zeros will occur. That will be the probability of the union of the six elementary events, which will be the sum of the probabilities of the six sequences. Since all six sequences have the same probability of occurring the probability of two ones and two zeros is six times the probability associated with a single sequence containing two ones and two zeros.

Notice something else. Expand the expression $(x+y)^4$.

$$\begin{aligned}(x+y)^4 &= (x+y)(x+y)(x+y)(x+y)\\ &= (x+y)(x+y)(x^2+2\,x\,y+y^2)\\ &= (x+y)(x^3+3\,x^2\,y+3\,x\,y^2+y^3)\\ &= x^4+4\,x^3\,y+6\,x^2\,y^2+4\,x\,y^3+y^4.\end{aligned}$$

Table 3.1: Sequence of Outcomes in an Accuracy Test of Four Guages

X_1	X_2	X_3	X_4	Sequence	$X = \sum x_i$	Probability
			1	S_1	4	$(.25)^4(.75)^0$
		1				
			0	S_2	3	$(.25)^3(.75)^1$
	1					
			1	S_3	3	$(.25)^3(.75)^1$
		0				
			0	S_4	2	$(.25)^2(.75)^2$
1						
			1	S_5	3	$(.25)^3(.75)^1$
		1				
			0	S_6	2	$(.25)^2(.75)^2$
	0					
			1	S_7	2	$(.25)^2(.75)^2$
		0				
			0	S_8	1	$(.25)^1(.75)^3$
			1	S_9	3	$(.25)^3(.75)^1$
		1				
			0	S_{10}	2	$(.25)^2(.75)^2$
	1					
			1	S_{11}	2	$(.25)^2(.75)^2$
		0				
			0	S_{12}	1	$(.25)^1(.75)^3$
0						
			1	S_{13}	2	$(.25)^2(.75)^2$
		1				
			0	S_{14}	1	$(.25)^1(.75)^3$
	0					
			1	S_{15}	1	$(.25)^1(.75)^3$
		0				
			0	S_{16}	0	$(.25)^0(.75)^4$

Table 3.2: Frequency Distribution of Sequences in Table 3.1

x	Frequency	Probability of Sequence	$P(x)$
0	1	$(.25)^0(.75)^4$	$\times 1 = .3164$
1	4	$(.25)^1(.75)^3$	$\times 4 = .4219$
2	6	$(.25)^2(.75)^2$	$\times 6 = .2109$
3	4	$(.25)^3(.75)^1$	$\times 4 = .0469$
4	1	$(.25)^4(.75)^0$	$\times 1 = .0039$

It turns out that the coefficients of the four terms are exactly the frequencies of the occurrences in the frequency distribution above and the $x\,y$ terms become the sequence probabilities in the table when x is replaced by the probability of a one and y is replaced the probability of a zero and $n = 4$. The above expansion of $(x + y)^4$ is called the *binomial expansion*, whence the term binomial distribution. The easiest way to calculate the binomial coefficients for the simplest cases (where n is small) is through the use of *Pascal's Triangle*.

Pascal's Triangle

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

etc.. . . . . .

Additional rows can be added to the base by noting that each number that is not unity is the sum of the two numbers above it. The relevant binomial coefficients appear in the row with $n + 1$ entries.

Fortunately, all these complicated calculations need not be made every time we want to find a binomial probability. Probability tables have been calculated for all necessary values of n and p and appear at the end of every statistics textbook.[3]

3.8 Poisson Probability Distributions

The Poisson probability distribution applies to many random phenomena occurring during a period of time—for example, the number of inaccurate gauges coming off an assembly line in a day or week. It also applies to spatial phenomena such as, for example, the number of typographical errors on a page.

A *Poisson random variable* is a discrete variable that can take on any integer value from zero to infinity. The value gives the number of occurrences of the circumstance of interest during a particular period of time or within a particular spatial area. A unit probability mass is assigned to this sample space. Our concern is then with the probability that there will be, for example, zero, one, two, three, etc., calls to the police during a particular time period on a typical day, or that in typing this manuscript I will make zero, one, two, etc. errors on a particular page.

The *Poisson probability function* is

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{3.32}$$

where

$$P(x) = P(X = x)$$

with

$$x = 0, 1, 2, 3, 4, \ldots, \infty$$

and $0 < \lambda < \infty$. The parameter $e = 2.71828$ is a constant equal to the base of natural logarithms.[4] Note that, whereas the binomial distribution had two parameters, n and p, the Poisson distribution has only one parameter, λ, which is the average number of calls over the period.

[3] These probabilities can also be calculated, and the various distributions plotted, using XlispStat and other statistical software.

[4] The number e is equal to

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n.$$

Consider an example. Suppose that the number of calls to the 911 emergency number between 8:00 and 8:30 PM on Fridays is a Poisson random variable X with $\lambda = 3.5$. We can calculate a portion of the probability distribution as follows:

x	$P(X = x)$					$P(X \leq x)$
0	$[3.5^0 e^{-3.5}]/0!$	=	$[(1)(.03019738)]/1$	=	.0302	.0302
1	$[3.5^1 e^{-3.5}]/1!$	=	$[(3.5)(.03019738)]/1$	=	.1057	.1359
2	$[3.5^2 e^{-3.5}]/2!$	=	$[(12/250)(.03019738)]/2$	=	.1850	.3208
3	$[3.5^3 e^{-3.5}]/3!$	=	$[(42.875)(.03019738)]/6$	=	.2158	.5366
4	$[3.5^4 e^{-3.5}]/4!$	=	$[(150.0625)(.03019738)]/24$	=	.1888	.7254
5	$[3.5^5 e^{-3.5}]/5!$	=	$[(525.2188)(.03019738)]/120$	=	.1322	.8576
6	$[3.5^6 e^{-3.5}]/6!$	=	$[(1838.266)(.03019738)]/720$	=	.0771	.9347
7	$[3.5^7 e^{-3.5}]/7!$	=	$[(6433.903)(.03019738)]/5040$	=	.0385	.9732

The figures in the right-most column are the cumulative probabilities. The probably of receiving 3 calls is slightly over .2 and the probability of receiving 3 or less calls is just under .54. Note that over 97 percent of the probability mass is already accounted for by $x \leq 7$ even though x ranges to infinity.

As in the case of the binomial distribution, it is unnecessary to calculate these probabilities by hand—Poisson tables can be found at the back of any textbook in statistics.[5] The mean and variance of a Poisson probability distribution are

$$E\{X\} = \lambda$$

and

$$\sigma^2\{X\} = \lambda.$$

Plots of Poisson distributions are shown in Figure 3.6. The top panel shows a Poisson distribution with $\lambda = .5$, the middle panel shows one with $\lambda = 3$

[5] And, as in the case of other distributions, probabilities can be calculated using statistical software such as XlispStat.

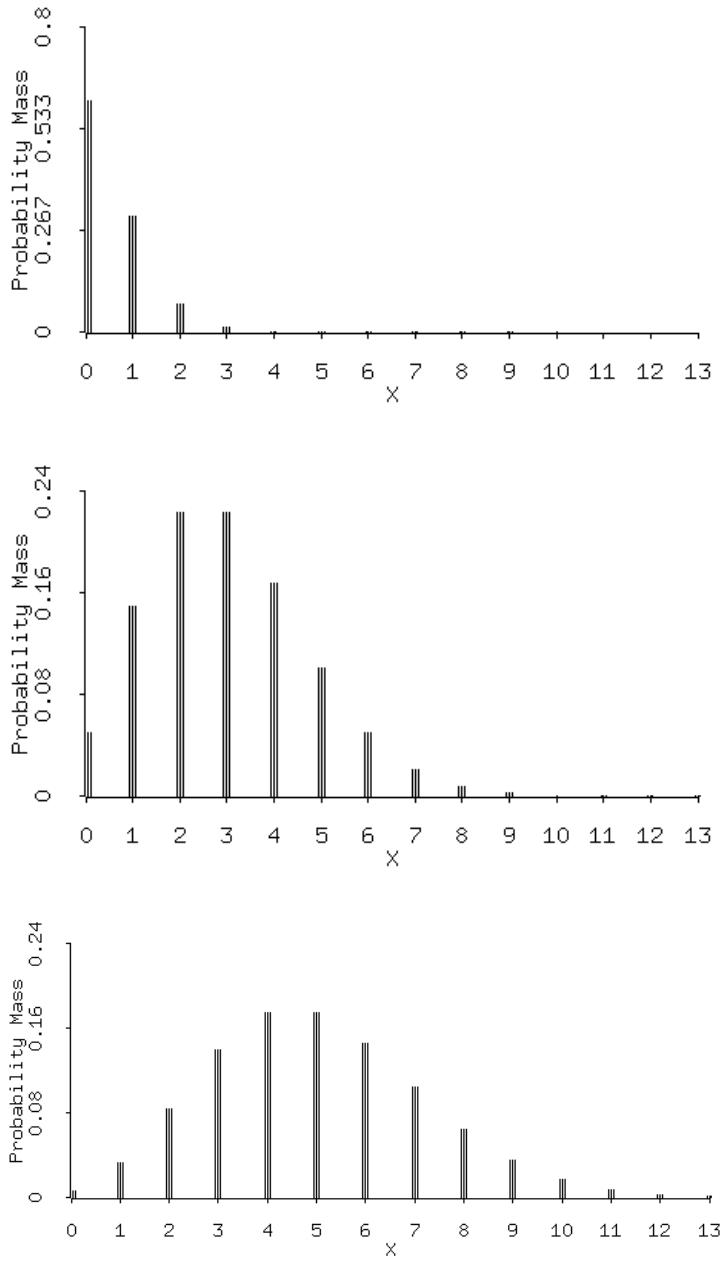


Figure 3.6: Poisson probability distributions with $\lambda = .5$ (top), $\lambda = 3$ (middle) and $\lambda = 5$ (bottom).

and the distribution plotted in the bottom panel has $\lambda = 5$. All Poisson probability distributions are skewed to the right although they become more symmetrical as λ gets larger.

Just as binomial distributions result from a Bernoulli process, Poisson distributions are the result of a *Poisson process*. A Poisson process is any process that generates occurrences randomly over a time or space continuum according to the following rules:

- The numbers of occurrences in non-overlapping time (space) intervals are statistically independent.
- The number of occurrences in a time (space) interval has the same probability distribution for all time (space) intervals.
- The probability of two or more occurrences in any interval $(t + \Delta t)$ is negligibly small relative to the probability of one occurrence in the interval.

When these postulates hold, the number of occurrences in a unit time (space) interval follows a Poisson probability distribution with parameter λ.

If V and W are two independent Poisson random variables with parameters λ_v and λ_w, respectively, the sum $V + W$ is a Poisson random variable with $\lambda = \lambda_v + \lambda_w$.

3.9 Uniform Probability Distributions

Uniform probability distributions result when the probability of all occurrences in the sample space are the same. These probability distributions may be either discrete or continuous.

Consider a computer random number generator that cranks out random numbers between 0 and 9. By construction of the computer program, the probability that any one of the 10 numbers will be turned up is 1/10 or 0.1. The probability distribution for this process is therefore

x:	0	1	2	3	4	5	6	7	8	9
$P(x)$:	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1

This random variable is called a *discrete uniform random variable* and its probability distribution is a *discrete uniform probability distribution*. The discrete probability function is

$$P(x) = \frac{1}{s} \tag{3.33}$$

where

$$P(x) = P(X = x),$$

$$x = a, a + 1, a + 2, \ldots, a + (s - 1).$$

The parameters a and s are integers with $s > 0$. Parameter a denotes the smallest outcome and parameter s denotes the number of distinct outcomes. In the above example, $a = 0$ and $s = 10$.

The mean and variance of a discrete uniform probability distribution are

$$E\{X\} = a + \frac{s - 1}{2}$$

and

$$\sigma^2 = \frac{s^2 - 1}{12}.$$

In the example above, $E\{X\} = 0 + 9/2 = 4.5$ and $\sigma^2 = 99/12 = 8.25$. A graph of a discrete probability distribution is shown in the top panel of Figure 3.7.

The *continuous uniform* or *rectangular* probability distribution is the continuous analog to the discrete uniform probability distribution. A *continuous uniform random variable* has uniform probability density over an interval. The *continuous uniform probability density function* is

$$f(x) \quad = \quad \frac{1}{b - a} \tag{3.34}$$

where the interval is $a \leq x \leq b$. Its mean and variance are

$$E\{X\} = \frac{b + a}{2}$$

and

$$\sigma^2\{X\} = \frac{(b - a)^2}{12}$$

and the *cumulative probability function* is

$$F(x) \quad = \quad P(X \leq x) \quad = \quad \frac{x - a}{b - a}. \tag{3.35}$$

Suppose, for example, that a team preparing a bid on an excavation project assesses that the lowest competitive bid is a continuous uniform random variable X with a =\$250.000 and b =\$300.000. With X measured in units of one thousand, the density function will be

$$f(x) = 1/50 = .02$$

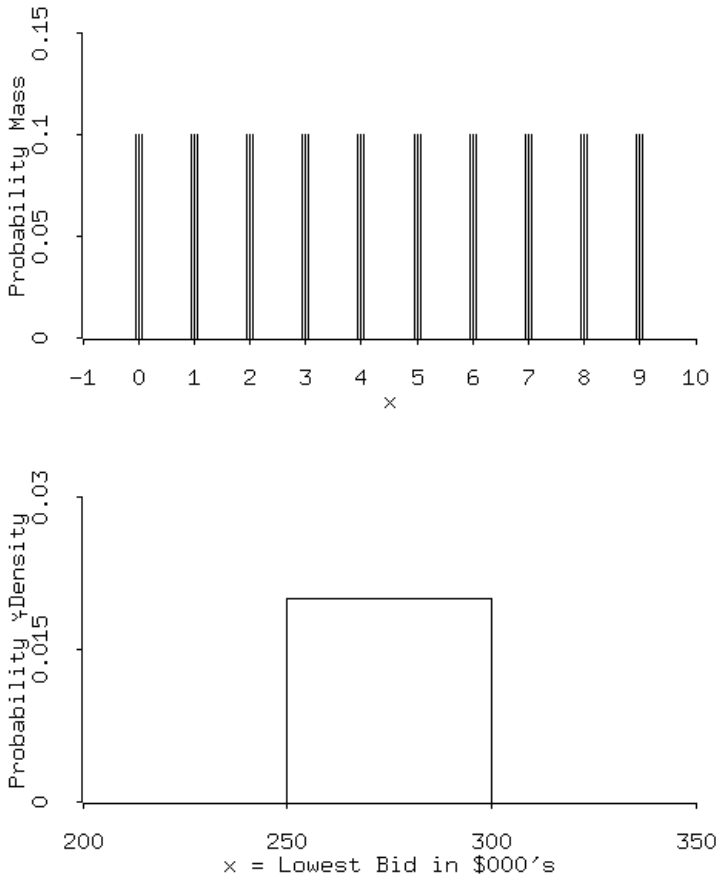


Figure 3.7: Discrete uniform probability distribution (top) and continuous uniform probability distribution (bottom).

where $250 \leq x \leq 300$. The graph of this distribution is shown in the bottom panel of Figure 3.7. The mean is 275 thousand and the variance is $50^2/12 = 2500/12 = 208.33$. The cumulative probability is the area to the left of x in the bottom panel of Figure 3.7. It is easy to eyeball the mean and the various percentiles of the distribution from the graph. The mean (and median) is the value of x that divides the rectangle in half, the lower quartile is the left-most quarter of the rectangle, and so forth. Keep in mind that, X being a continuous random variable, the probability that $X = x$ is zero.

3.10 Normal Probability Distributions

The family of normal probability distributions is the most important of all for our purposes. It is an excellent model for a wide variety of phenomena—for example, the heights of 10 year olds, the temperature in New York City at 12:01 on January 1, the IQs of individuals in standard IQ tests, etc. The *normal random variable* is a continuous one that may take any value from $-\infty$ to $+\infty$. Even though the normal random variable is not bounded, its probability distribution yields an excellent approximation to many phenomena.

The normal probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma]^2} \tag{3.36}$$

where $-\infty \leq x \leq +\infty$, $-\infty \leq \mu \leq +\infty$, $\sigma > 0$, $\pi = 3.14159$ and $e = 2.71828$.

The mean and variance of a normal probability distribution are

$$E\{X\} = \mu$$

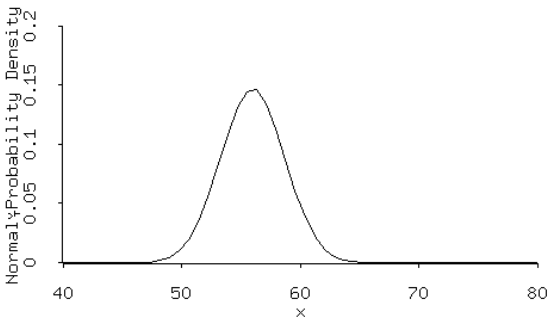
and

$$\sigma^2\{X\} = \sigma^2.$$

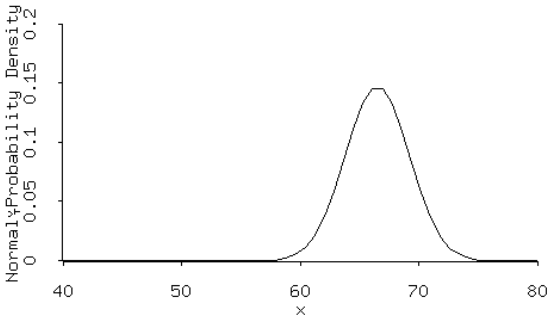
The distribution's two parameters, μ and σ, are its mean and standard deviation. Each parameter pair (μ, σ) corresponds to a different member of the family of normal probability distributions. Every normal distribution is bell shaped and symmetrical, each is centred at the value of μ and spread out according to the value of σ. Normal distributions are often referred to using the compact notation $N(\mu, \sigma^2)$. Three different members of the family of normal distributions are shown in Figure 3.8. In the top panel $\mu = 56$ and $\sigma = 2.7$ $[N(56, 7.29)]$ and in the middle panel $\mu = 66.5$ and $\sigma = 2.7$ $[N(66.5, 7.29)]$. In the bottom panel $\mu = 66.5$ and $\sigma = 4.1$ $[N(66.5, 16.81)]$.

The *standardised normal distribution* is the most important member of the family of normal probability distributions—the one with $\mu = 0$ and $\sigma = 1$. The normal random variable distributed according to the standard normal distribution is called the *standard normal variable* and is denoted by Z. It is expressed as

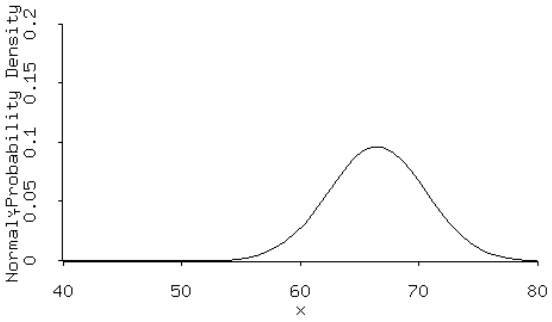
$$Z = \frac{X - \mu}{\sigma} \tag{3.37}$$



$\mu = 56\ ,\ \sigma = 2.7$



$\mu = 66.5\ ,\ \sigma = 2.7$



$\mu = 66.5\ ,\ \sigma = 4.1$

Figure 3.8: Three different members of the family of normal probability distributions.

A basic feature of normal distributions is that any linear function of a normal random variable is also a normal random variable. Thus

$$Z = -\frac{\mu}{\sigma} + \frac{1}{\sigma} X \tag{3.38}$$

and

$$X = \mu + \sigma Z \tag{3.39}$$

Figure 3.9 plots a normally distributed random variable in both its regular and standard form. It can be shown using (3.38) and (3.39) that $X = 67.715$ (i.e, 67.715 on the X scale) is equivalent to $Z = .45$ (i.e., .45 on the Z scale). This means that 67.715 is .45 standard deviations away from μ, which is 66.5. The probability that $X \geq 67.715$ is found by finding the corresponding value on the Z scale using (3.38) and looking up the relevant area to the left of that value in the table of standard normal values that can be found in the back of any textbook in statistics. To find the value of X corresponding to any cumulative probability value, we find the corresponding value of Z in the table of standard normal values and then convert that value of Z into X using (3.39). All calculations involving normal distributions, regardless of the values of μ and σ can thus be made using a single table of standard normal values.

If V and W are two independent normal random variables with means μ_v and μ_w and variances σ_v^2 and σ_w^2 respectively, the sum $V + W$ is a normal random variable with mean $\mu = \mu_v + \mu_w$ and variance $\sigma^2 = \sigma_v^2 + \sigma_w^2$. This extends, of course, to the sum of more than two random variables.

It is often useful to use the normal distribution as an approximation to the binomial distribution when the binomial sample space is large. This is appropriate when both np and $n(1-p)$ are greater than 5. To make a normal approximation we calculate the standard variate

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{n\,p\,(1-p)}}. \tag{3.40}$$

We can then look up a value of Z so obtained in the normal distribution table. Alternatively, if we are given a probability of X being, say, less than a particular value we can find the value of Z from the table consistent with that probability and then use (3.39) to find the corresponding value of X.

For example, suppose we were to flip a coin 1000 times and want to know the probability of getting more than 525 heads. That is, we want to find the probability that $X \geq 525$. It turns out that

$$n\,p = n\,(1-p) = 500 > 5$$

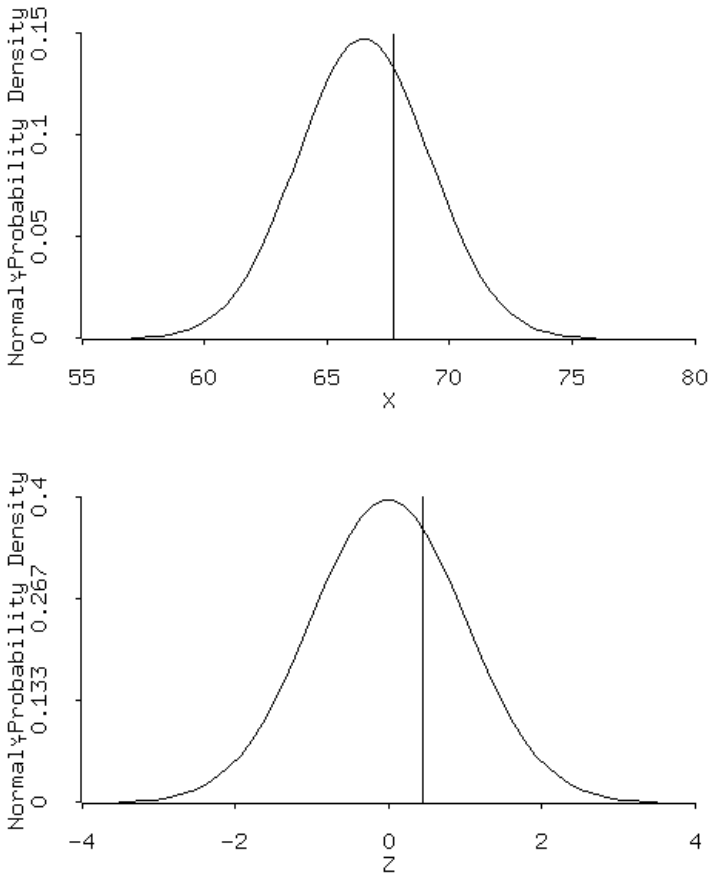


Figure 3.9: A normal distribution (top) and its standardized form (bottom). As marked by the vertical lines, 67.715 on the X scale in the top panel corresponds to .45 on the Z scale in the bottom panel.

so a normal approximation is appropriate. From (3.40) we have

$$Z = (525 - 500)/\sqrt{(1000)(.5)(.5)} = 25/\sqrt{250} = 1.58.$$

It can be seen from the probability tables for the normal distribution that

$$P(Z \leq 1.58) = .9429$$

which implies that

$$P(X \geq 525) = P(Z \geq 1.58) = 1 - .9429 = .0571.$$

There is almost a 6% chance of getting more than 525 heads.

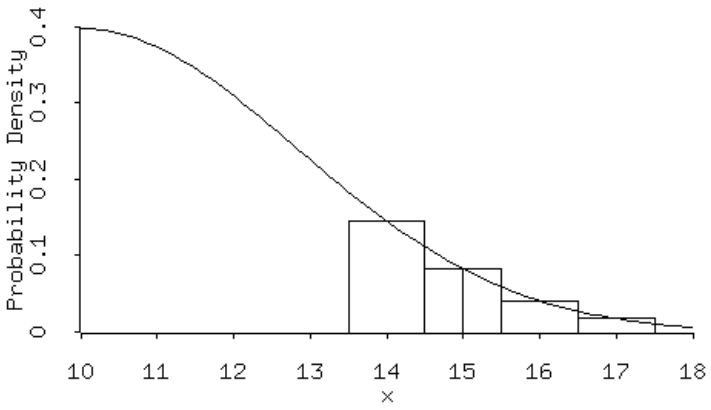


Figure 3.10: A normal approximation to a binomial distribution requiring correction for continuity.

Consider a second example of using a normal distribution to approximate a binomial one. Suppose the probability that a machine is in an unproductive state is $p = .2$. Let X denote the number of times the machine is in an unproductive state when it is observed at 50 random moments in time. It is permissible to use a normal approximation here because $(.2)(50) = 10$ and $(1-.2)(50) = 40$ and both these numbers exceed 5. The mean of distribution of X is $n\,p = 10$ and the standard deviation is

$$\sigma\{X\} = \sqrt{n\,p\,(1-p)} = \sqrt{(50)(.2)(.8)} = 2.83.$$

Now suppose we want to obtain the probability that $X = 15$. Since $n = 50$, X can be located at only 51 of the infinitely many points along the continuous line from 0 to 50. The probability that $X = 15$ on the continuum is zero. Since the underlying distribution is discrete, the probability that $X = 15$ is the area of the vertical strip under the probability density function between $X = 14.5$ and $X = 15.5$. This can be seen in Figure 3.10. So the probability that $X = 15$ becomes

$$\begin{aligned} P(X \leq 15.5) - P(X \leq 14.5) &= P(Z \leq (15.5-10)/2.83) \\ &\quad - P(Z \leq (14.5-10)/2.83) \\ &= P(Z \leq 1.94) - P(Z \leq 1.59) \\ &= .9738 - .9441 = .0297. \end{aligned}$$

Similarly, if we want to calculate the probability that $X > 15$ we must calculate

$$P(X \geq 15.5) = (P(Z \geq 1.59) = 1 - P(Z \leq 1.59) = 1 - .9739 = .0262.$$

We base the calculation on $X \geq 15.5$ rather than $X \geq 15$ to correct for the fact that we are using a continuous distribution to approximate a discrete one. This is called a *correction for continuity*. It can be seen from Figure 3.10 that if we were to base our calculation on $X \geq 15$ the number obtained would be too large.

3.11 Exponential Probability Distributions

The Poisson probability distribution applies to the number of occurrences in a time interval. The *exponential* probability distribution applies to the amount of time between occurrences. For this reason it is often called the *waiting-time distribution*. It is a continuous distribution because time is measured along a continuum. An *exponential random variable* X is the time between occurrences of a random event. The probability density function is

$$f(x) = \lambda e^{-\lambda x}, \qquad (x > 0). \tag{3.41}$$

It turns out that the probability that $X \geq x$ is

$$P(X \geq x) = e^{-\lambda x}. \tag{3.42}$$

The mean and variance of an exponential distribution are

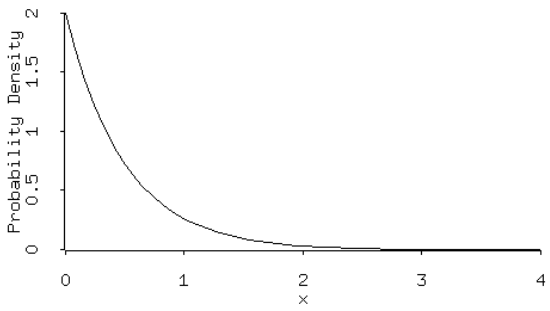
$$E\{X\} = \frac{1}{\lambda}$$

and

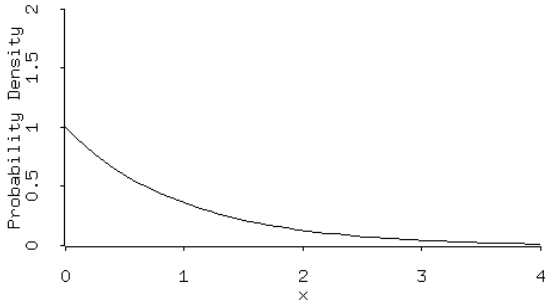
$$\sigma^2\{X\} = \frac{1}{\lambda^2}.$$

The shape of the exponential distribution is governed by the single parameter λ. As indicated in the plots of some exponential distributions in Figure 3.11, the exponential probability density function declines as x increases from zero, with the decline being sharper the greater the value of λ. The probability density function intersects the y-axis at λ.

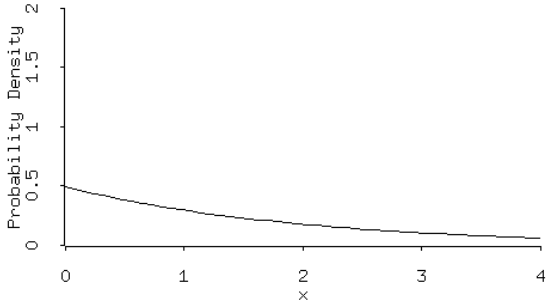
The area to the right of any value of x—that is, $P(X \geq x)$—can be looked up in the exponential distribution table at the back of any statistics textbook.



$\lambda = 2$



$\lambda = 1$



$\lambda = 0.5$

Figure 3.11: Three different members of the family of exponential probability distributions.

Consider an example. Suppose that the manufacturer of an electronic component has good reason to believe that its length of life in years follows an exponential probability distribution with $\lambda = 16$. He is considering giving a guarantee on the component and wants to know what fraction of the components he will have to replace if he makes the guarantee a five-year one. The mean time until the component breaks will be $1/\lambda = 1/16 = 6.25$ years. To find the fraction of components that will have to be replaced within 5 years we need $P(X \leq 5)$—that is, the area under the distribution to the left of $x = 5$. That area is equal to $(1-P(X \geq 5))$ which can be found by using either equation (3.42) or the exponential distribution table. The value obtained is .550671. This means that about 55% of the components will have to be replaced within five years.

There is a close relationship between the exponential and Poisson distributions. If occurrences are generated by a Poisson process with parameter λ then the number of occurrences in equal non-overlapping units are independent random variables having a Poisson distribution with parameter λ and the durations between successive occurrences are independent random variables having the exponential distribution with parameter λ.

3.12 Exercises

1. The random variable X has a normal probability distribution with $\mu = 12$ and $\sigma = 16$. Estimate the following probabilities:

a) $P(X \leq 14.4)$

b) $P(7.2 \leq X \leq 12.8)$ (.35)

c) $P((X - \mu) \leq 5.6)$

d) $P(X \geq 8.0)$

2. The number of coating blemishes in 10-square-meter rolls of customized wallpaper is a Poisson random variable X_1 with $\lambda_1 = 0.3$. The number of printing blemishes in these 10-square-meter rolls of customized wallpaper is a Poisson random variable X_2 with $\lambda_2 = 0.1$. Assume that X_1 and X_2 are independent and let $T = X_1 + X_2$.

a) According to what distribution is the random variable T distributed?

b) What is the most probable total number of blemishes in a roll? (0)

c) If rolls with a total of two or more blemishes are scrapped, what is the probability that a roll will be scrapped? (.062)

d) What are the mean and standard deviation of the probability distribution of T?

3. There are three surviving members of the Jones family: John, Sarah, and Beatrice. All live in different locations. The probability that each of these three family members will have a stay of some length in the hospital next year is 0.2.

a) What is the probability that none of the three of them will have a hospital stay next year? (.512)

b) What is the probability that all of them will have a hospital stay next year?

c) What is the probability that two members of the family will spend time in hospital next year? (.096)

d) What is the probability that either John or Sarah, but not both, will spend time in the hospital next year?

Hint: Portray the sample space as a tree.

4. Based on years of accumulated evidence, the distribution of hits per team per nine-innings in Major League Baseball has been found to be approximately normal with mean 8.72 and standard deviation 1.10. What percentage of 9-inning Major League Baseball games will result in fewer than 5 hits?

5. The *Statistical Abstract of the United States, 1995* reports that that 24% of households are composed of one person. If 1,000 randomly selected homes are to participate in a Nielson survey to determine television ratings, find the approximate probability that no more than 240 of these homes are one-person households.

6. Suppose the f-word is heard in the main hall in a Toronto high school every 3 minutes on average. Find the probability that as many as 5 minutes could elapse without us having to listen to that profanity. (.188)

7. A manufacturer produces electric toasters and can openers. Weekly sales of these two items are random variables which are shown to have positive covariance. Therefore, higher sales volumes of toasters:

a) are less likely to occur than smaller sales volumes of toasters.

b) tend to be associated with higher sales volumes of can openers.

c) tend to be associated with smaller sales volumes of can openers.

d) are unrelated to sales of can openers.

8. Which of the following could be quantified as a Bernoulli random variable?

a) number of persons in a hospital ward with terminal diagnoses.

b) weights of deliveries at a supermarket.

c) square foot areas of houses being built in a suburban tract development.

d) whether or not employees wear glasses.

e) none of the above.

9. Fifteen percent of the patients seen in a pediatric clinic have a respiratory complaint. In a Bernoulli process of 10 patients, what is the probability that at least three have a respiratory complaint?

a) .1298

b) .1798

c) .1960

d) .9453

e) none of the above.

10. Two random variables X and Y have the following properties: $\mu_x = 10$, $\sigma_x = 4$, $\mu_y = 8$, $\sigma_y = 5$, $\sigma_{x,y} = -12$.

a) Find the expected value and variance of $(3X - 4Y)$.

b) Find the expected value of X^2. (Hint: work from the definition of the variance of X.)

c) Find the correlation between X and $(X + Y)$.

d) Find the covariance between the standardised values of X and Y.

11. John Daly is among the best putters on the PGA golf tour. He sinks 10 percent of all puts that are of length 20 feet or more. In a typical round of golf, John will face puts of 20 feet or longer 9 times. What is the probability that John will sink 2 or more of these 9 puts? What is the probability that he will sink 2 or more, given that he sinks one? Hint: If we know he is going to sink at least one then the only remaining possibilities are that he will sink only that one or two or more. What fraction of the remaining probability weight (excluding the now impossible event that he sinks zero) falls on the event 'two or more'.

12. Let X and Y be two random variables. Derive formulae for $E\{X+Y\}$, $E\{X-Y\}$, $\sigma^2\{X+Y\}$, and $\sigma^2\{X-Y\}$. Under what conditions does $\sigma^2\{X+Y\} = \sigma^2\{X-Y\}$?

13. According to the Internal Revenue Service (IRS), the chances of your tax return being audited are about 6 in 1000 if your income is less than \$50,000, 10 in 1000 if your income is between \$50,000 and \$99,999, and 49 in 1000 if your income is \$100,000 or more (*Statistical Abstract of the United States: 1995*.

a) What is the probability that a taxpayer with income less than \$50,000 will be audited by the IRS? With income between \$50,000 and \$99,999? With income of \$100,000 or more?

b) If we randomly pick five taxpayers with incomes under \$50,000, what is the probability that one of them will be audited? That more than one will be audited? Hint: What are n and p here?

14. Let $X_i = 1$ with probability p and 0 with probability $1-p$ where X_i is an independent sequence. For

$$X = \sum_{i=1}^{n} X_i$$

show that

$$E\{X\} = np$$

and

$$\sigma^2\{X\} = np(1-p).$$

15. The number of goals scored during a game by the Toronto Maple Leafs is a normally distributed random variable X with $\mu_x = 3$ and $\sigma_x = 1.2$. The number of goals given up during a game when Curtis Joseph is the goaltender for the Maple Leafs is a normally distributed random variable Y with $\mu_y = 2.85$ and $\sigma_y = 0.9$. Assume that X and Y are independent.

a) What is the probability that the Maple Leafs will win a game in which Curtis Joseph is the goaltender? (The probability of a game ending in a tie is zero here.)

b) What is the probability that the Maple Leafs will lose a game by 2 or more goals when Curtis Joseph is the goaltender?

c) Let T denote the total number of goals scored by both the Maple Leafs and their opponent during a game in which Curtis Joseph is the Leafs' goaltender. What is the expected value and variance of T?

d) Given your answer to a) and assuming that the outcomes of consecutive games are independent, what is the expected number of wins for the Maple Leafs over 50 games in which Curtis Joseph is the goaltender? Hint: What kind of process is occurring here?

16. The elapsed time (in minutes) between the arrival of west-bound trains at the St. George subway station is an exponential random variable with a value of $\lambda = .2$.

a) What are the expected value and variance of X?

b) What is the probability that 10 or more minutes will elapse between consecutive west-bound trains?

c) What is the probability that 10 or more minutes will elapse between trains, given that at least 8 minutes have already passed since the previous train arrived? Hint: What proportion of the probability weight that remains, given that a waiting time of less than 8 minutes is no longer possible, lies in the interval 8 minutes to 10 minutes?

17. The number of houses sold each month by a top real estate agent is a Poisson random variable X with $\lambda = 4$.

a) What are the expected value and standard deviation of X?

b) What is the probability that the agent will sell more than 6 houses in a given month?

c) Given that the agent sells at least 2 houses in a month, what is the probability that she will sell 5 or more?

18. In the National Hockey League (NHL), games that are tied at the end of three periods are sent to "sudden death" overtime. In overtime, the team to score the first goal wins. An analysis of NHL overtime games played between 1970 and 1993 showed that the length of time elapsed before the winning goal is scored has an exponential distribution with mean 9.15 minutes (*Chance*, Winter 1995).

a) For a randomly selected overtime NHL game, find the probability that the winning goal is scored in three minutes or less.

b) In the NHL, each period (including overtime) lasts 20 minutes. If neither team scores a goal in one period of overtime, the game is considered a tie. What is the probability of an NHL game ending in a tie?

19. A taxi service based at an airport can be characterized as a transportation system with one source terminal and a fleet of vehicles. Each vehicle takes passengers from the terminal to different destinations and then returns after some random trip time to the terminal and makes another trip. To improve the vehicle-dispatching decisions involved in such a system, a study was conducted and published in the *European Journal of Operational Research* (Vol. 21, 1985). In modelling the system, the authors assumed travel times of successive trips to be independent exponential random variables with $\lambda = .05$.

a) What is the mean trip time for the taxi service?

b) What is the probability that a particular trip will take more than 30 minutes?

c) Two taxis have just been dispatched. What is the probability that both will be gone more than 30 minutes? That at least one of the taxis will return within 30 minutes?

20. The probability that an airplane engine will fail is denoted by π. Failures of engines on multi-engine planes are independent events. A two engine plane will crash only if both of its engines fail. A four engine plane can remain airborne with two or more engines in operation. If $\pi = 0$ or $\pi = 1$, a traveller will clearly be indifferent between planes with two or four engines. What are the values of π that make a two engine plane safer than a four engine plane? Hint: Set the sample space up in tree form.

Chapter 4

Statistical Sampling: Point and Interval Estimation

In the previous chapter we assumed that the probability distribution of a random variable in question was known to us and from this knowledge we were able to compute the mean and variance and the probabilities that the random variable would take various values (in the case of discrete distributions) or fall within a particular range (in the case of uniform distributions). In most practical applications of statistics we may have some reason to believe that a random variable is distributed according to a binomial, Poisson, normal, etc., distribution but have little knowledge of the relevant parameter values. For example, we might know what n is in the case of a binomial distribution but know nothing about the magnitude of p. Or we may suspect that a variable is normally distributed by have no idea of the values of the parameters μ and σ. The practical procedure for finding information about these parameters is to take a sample and try to infer their values from the characteristics of the sample.

4.1 Populations and Samples

Let us first review what we learned about populations and samples in Chapter 1. A population is the set of elements of interest. It may be finite or infinite. Processes, mechanisms that produce data, are infinite populations. In terms of the analysis of the previous chapter, populations are the complete set of outcomes of a random variable. And a process is a mechanism by which outcomes of a random variable are generated. The population of outcomes of a particular random variable is distributed according to some probability

distribution—possibly but not necessarily binomial, Poisson, normal, uniform, or exponential. The parameters of the population are the parameters of its probability distribution. As such, they are numerical descriptive measures of the population. A census is a listing of the characteristics of interest of every element in a population. A sample is a subset of the population chosen according to some set of rules. *Sample statistics* are numerical descriptive measures of the characteristics of the sample calculated from the observations in the sample. We use these sample statistics to make inferences about the unobserved population parameters. You should keep in mind that a *statistic* refers to a sample quantity while a *parameter* refers to a population quantity. The sample mean is an example of a sample statistic, while the population mean is an example of a population parameter.

A *sample* is thus a part of the population under study selected so that inferences can be drawn from it about the population. It is cheaper and quicker to use samples to obtain information about a population than to take a census. Furthermore, testing items sampled may destroy them so that tests cannot be conducted on the whole population.

A *probability sample* is one where the selection of the elements from the population that appear in the sample is made according to known probabilities. A *judgment sample* is one where judgment is used to select "representative" elements or to infer that a sample is "representative" of the population. In probability samples, no discretion is allowed about which population elements enter the sample.

The most common sampling procedure is to select a *simple random sample*. A simple random sample is one for which *each possible sample combination* in the population has an equal probability of being selected. Every element of the population has the same probability of occupying each position in the sample. The sampling is without replacement, so that no element of the population can appear in the sample twice.

Note that simple random sampling requires more than each element of the population having the same probability of being selected. Suppose that we select a sample of 10 students to interview about their career plans. It is not enough that every student in the population have an equal chance of being among the 10 selected. Each student must have the same chance of being the first selected, the second selected, the third selected, etc. For example, we could divide the population into males and females (suppose the population contains an equal number of each) and select 5 males and 5 females at random for the sample. Each student would have an equal chance of being in the sample, but the sample combinations that contain an unequal number of males and females would be ruled out. One might wish

to rule these combinations out, but then the sample would not be a simple random sample.

One way to ensure that each possible sample combination has an equal chance of being in the sample is to select the sample elements one at a time in such a way that each element of the population not already in the sample has an equal chance of being chosen. In the case of a finite population, select the first element by giving each of the N population elements an equal chance of being picked. Then select the second sample element by giving the remaining $N - 1$ elements of the population an equal chance of being chosen. Repeat this process until all n sample elements have been selected.

Suppose we have a population of 5000 students that we wish to sample. We could assign each student in the population a number between 0 and 4999 and chose 100 numbers at random from the set of integers in this interval, using the numbers so selected to pick the students to appear in the sample. To choose the numbers randomly we could get a computer to spit out 100 numbers between 0 and 4999 in such a way that each of the 5000 numbers had an equal chance of being selected first and each of the 5000 numbers not yet selected had an equal chance of being selected second, third, etc. Alternatively, we could use a table of random numbers. Such a table might list five-digit numbers in the following fashion:

13284 21244 99052 00199 40578 ..... etc.

The table is constructed so each digit between 0 and 9 has an equal chance of appearing in each of the five positions for each number. We could select our sample as follows from these numbers:

1328, 2122, skip, 0019, 4057, skip, ... etc.

Numbers for which the four digits on the left side yield a number larger than 4999 are simply skipped—they can be treated as not being in the table, so that numbers between 0 and 4999 have an equal chance of being selected and numbers over 4999 have a zero chance of being selected. Any number already selected would also be discarded because we want the probability that an element of the population will be selected more than once to be zero. If the size of the population is, say, 500000, requiring that we select the elements in the sample from 6 digit numbers, we merely take each succession of 6 digits in the table of random numbers as a separate number, so that the above line in the table of random numbers would yield

132842 124499 052001 994057 ..... etc.

The first three numbers would be used to select the corresponding population elements, the fourth number would be skipped, and so on. Random numbers can also be obtained from the table by reading down the columns rather than across the rows, and the selection process can begin anywhere in the table.

When the population is generated by a process, the process itself furnishes the sample observations. Take the case of pairs of shoes coming off an assembly line. To test the quality of the production process we could select a sample of 10 pairs by simply taking the next (or any) 10 pairs off the line. This will give us a simple random sample if two conditions are met: First, each item must have the same probability of being defective as any other item. Second, the probability that any one item is defective must be independent of whether any particular other item is defective. More formally, the n random variables $X_1, X_2, X_3, \ldots X_n$ generated by a process constitute a simple random sample from an infinite population if they are *independently and identically distributed*.

Once a sample has been selected and observations on the sample elements have been made, the observations constitute a data set and the usual summary measures can be made. If $X_1, X_2, X_3, \ldots X_n$ represent the values of the n sample observations, we have

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{4.1}$$

and

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \tag{4.2}$$

where $\bar{X}$ and s^2 are the sample mean and variance, and s is the sample standard deviation. These magnitudes are called sample statistics. The population mean, variance and standard deviation—that is, the population *parameters*—are denoted by μ, σ^2 and σ.

4.2 The Sampling Distribution of the Sample Mean

Consider an example of pairs of newly produced shoes coming off an assembly line. We want to verify their quality. The sample space consists of three sample points—neither shoe defective, one shoe defective, both shoes defective. Suppose that the process by which the shoes are manufactured generates the following *population* probability distribution for the three values that the random variable X can take:

x:	0	1	2
$P(x)$:	.81	.18	.01

Note that the population distribution is skewed to the right. Its mean is

$$E\{X\} = \mu = (0)(.81) + (1)(.18) + (2)(.01) = .2$$

and its variance is

$$\sigma^2\{X\} = (-.2)^2(.81) + (.8)^2(.18) + (1.8)^2(.01) = .18.$$

Now suppose that we do not observe the probability distribution for the population and do not know its parameters. We can attempt to make an inference about these parameters, and hence about the probability distribution of the population, by taking a sample. Suppose we take a sample of two and use the sample mean as an estimate of E{X}. There are nine potential samples of two that can be taken from the population. These potential samples and the corresponding sample means together with the probabilities of picking each sample are listed below:

Sample	$\bar{X}$	$P(\bar{X})$
0 0	0.0	$(.81)^2 = .6561$
0 1	0.5	$(.81)(.18) = .1458$
0 2	1.0	$(.81)(.01) = .0081$
1 0	0.5	$(.18)(.81) = .1458$
1 1	1.0	$(.18)^2 = .0324$
1 2	1.5	$(.18)(.01) = .0018$
2 0	1.0	$(.01)(.81) = .0081$
2 1	1.5	$(.01)(.18) = .0018$
2 2	2.0	$(.01)^2 = .0001$
		1.0000

The sum of the probabilities is unity because all possible samples of two that can be drawn from the population are listed. It turns out that the sample mean can take five values— 0, .5, 1, 1.5 and 2. The probabilities that it will take each of these values can be obtained by adding the probabilities associated with the occurrence of each possible sample value in the table above. For example, the probability that the sample mean will be .5 equals .1458 + .1458 = .2916. We thus have

$\bar{X}$:	0	.5	1	1.5	2
$P(\bar{X})$:	.6561	.2916	.0486	.0036	.0001

for which the probabilities sum to unity. This is the exact sampling distribution of $\bar{X}$. It says that there is a probability of .6561 that a sample of two will have mean 0, a probability of .2916 that it will have mean 0.5, and so forth. The mean of the sampling distribution of $\bar{X}$ is

$$E\{\bar{X}\} = (0)(.6561) + (.5)(.2916) + (1)(.0486) + (1.5)(.0036) + (2)(.0001) = .2$$

which is equal to the population mean. The variance of the sample mean is

$$\sigma^2\{\bar{X}\} = (-.2)^2(.6561) + (.3)^2(.2916) + (.8)^2(.0486)$$

$$+(1.3)^2(.0036) + (1.8)^2(.0001) = .09$$

which turns out to be half the population variance.

Now consider all possible samples of three that we could take. These are presented in Table 4.1. The sample mean can now take seven values— 0, 1/3, 2/3, 1, 4/3, 5/3, and 2. The exact sampling distribution of the sample mean (which is obtained by adding up in turn the probabilities associated with all samples that yield each possible mean) is now

$\bar{X}$:	0	1/3	2/3	1	4/3	5/3	2
$P(\bar{X})$:	.531441	.354294	.098415	.014580	.001215	.000054	.000001

The usual calculations yield a mean of the sample mean of $E\{\bar{X}\} = .2$ and a sample variance of $\sigma^2\{\bar{X}\} = .06$. The mean sample mean is again the same as the population mean and the variance of the sample mean is now one-third the population variance.

On the basis of an analysis of the exact sampling distributions of the sample mean for sample sizes of 2 and 3, we might conjecture that the expected value of the sample mean always equals the population mean and the variance of the sample mean always equals the variance of the population divided by the sample size. This conjecture is correct. For a sample of size n consisting of $X_1, X_2, X_3, \ldots, X_n$, the expectation of the sample mean will be

$$\begin{aligned} E\{\bar{X}\} &= E\left\{\frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n)\right\} \\ &= \frac{1}{n}(E\{X_1\} + E\{X_2\} + E\{X_3\} + \ldots + E\{X_n\}) \\ &= \frac{1}{n}(n\,\mu) \;=\; \mu \end{aligned} \tag{4.3}$$

Table 4.1: All possible samples of three for the shoe-testing problem.

	$\bar{X}$	$P(\bar{X})$
0 0 0	0	$(.81)^3 = .531441$
0 0 1	1/3	$(.81)(.81)(.18) = .118098$
0 0 2	2/3	$(.81)(.81)(.01) = .006561$
0 1 0	1/3	$(.81)(.18)(.81) = .118098$
0 1 1	2/3	$(.81)(.18)(.18) = .026244$
0 1 2	1	$(.81)(.18)(.01) = .001458$
0 2 0	2/3	$(.81)(.01)(.81) = .006561$
0 2 1	1	$(.81)(.18)(.01) = .001458$
0 2 2	4/3	$(.81)(.01)(.01) = .000081$
1 0 0	1/3	$(.18)(.81)(.81) = .118098$
1 0 1	2/3	$(.18)(.81)(.18) = .026244$
1 0 2	1	$(.18)(.81)(.01) = .001458$
1 1 0	2/3	$(.18)(.18)(.81) = .026244$
1 1 1	1	$(.18)^3 = .005832$
1 1 2	4/3	$(.18)(.18)(.01) = .000324$
1 2 0	1	$(.18)(.01)(.81) = .001458$
1 2 1	4/3	$(.18)(.01)(.18) = .000324$
1 2 2	5/3	$(.18)(.01)(.01) = .000018$
2 0 0	2/3	$(.01)(.81)(.81) = .006561$
2 0 1	1	$(.01)(.81)(.18) = .001458$
2 0 2	4/3	$(.01)(.81)(.01) = .000081$
2 1 0	1	$(.01)(.18)(.81) = .001458$
2 1 1	4/3	$(.01)(.18)(.18) = .000324$
2 1 2	5/3	$(.01)(.18)(.01) = .000018$
2 2 0	4/3	$(.01)(.01)(.81) = .000081$
2 2 1	5/3	$(.01)(.01)(.18) = .000018$
2 2 2	2	$(.01)^3 = .000001$
		1.000000

and the variance of the sample mean will be

$$\begin{aligned}
\sigma^2\{\bar{X}\} &= E\left\{\left[\frac{1}{n}(X_1+X_2+X_3+\ldots+X_n)-E\{\bar{X}\}\right]^2\right\} \\
&= E\left\{\left[\frac{1}{n}(X_1+X_2+X_3+\ldots+X_n)-\mu\right]^2\right\} \\
&= E\left\{\left[\frac{1}{n}(X_1+X_2+X_3+\ldots+X_n)-\frac{n\mu}{n}\right]^2\right\} \\
&= \frac{1}{n^2}E\left\{[(X_1+X_2+X_3+\ldots+X_n)-n\mu]^2\right\} \\
&= \frac{1}{n^2}E\left\{[((X_1-\mu)+(X_2-\mu)+(X_3-\mu)+\ldots+(X_n-\mu)]^2\right\} \\
&= \frac{1}{n^2}\left[\sigma^2\{X_1\}+\sigma^2\{X_2\}+\sigma^2\{X_3\}+\ldots+\sigma^2\{(X_n\}\right] \\
&= \frac{1}{n^2}\left[n\,\sigma^2\right] = \frac{\sigma^2}{n}. \qquad (4.4)
\end{aligned}$$

Note that in the second last line we took advantage of the fact that the sample items were chosen independently to rule out any covariance between X_i and X_j.

It should be emphasized that the above calculations of the mean and variance of the sampling distribution are the same regardless of the distribution of the population. For the population above, increasing the sample size from two to three reduced the probability weight at the right tail of the distribution and also at $\bar{X}=0$.

The question immediately arises as to what the distribution of the sample mean will look like if we increase the sample size further. It is not practical to obtain the exact distribution of the sample mean from the above population for sample sizes bigger than three. We have to infer the probability distribution of the sample mean by taking many samples of each size and plotting histograms of the resulting sample means.

4.3 The Central Limit Theorem

Figure 4.1 shows the distribution of the sample means obtained for the shoe-testing problem by taking 1000 samples of $n=2$ (top), $n=3$ (middle) and $n=10$ (bottom). Notice how the range of the sample mean narrows as the sample size increases. Also, with a sample size as large as 10 the modal value ceases to be zero. Figure 4.2 is a continuation of Figure 4.1, showing

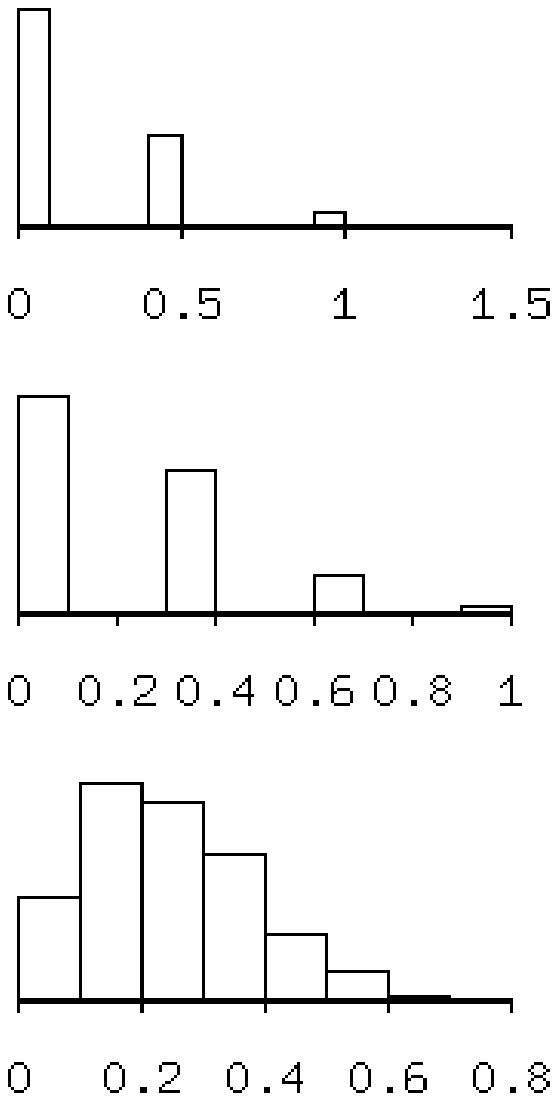


Figure 4.1: Distribution of the Sample Mean for 1000 samples of $n = 2$ (top), $n = 3$ (middle) and $n = 10$ (bottom).

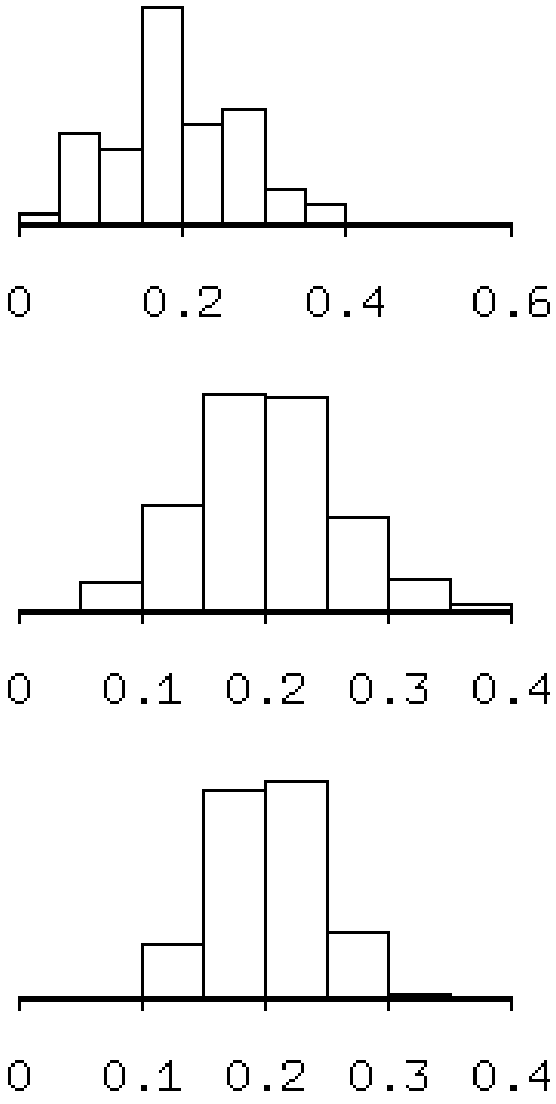


Figure 4.2: Distribution of the Sample Mean for 1000 samples of $n = 30$ (top), $n = 50$ (middle) and $n = 100$ (bottom).

the distribution of the sample means for 1000 samples when $n = 30$ (top), $n = 50$ (middle) and $n = 100$ (bottom). The range of the sample mean again narrows as the sample size increases and the distribution of the sample mean becomes more symmetrical around the population mean, $\mu = .2$.

Figure 4.3 is obtained by superimposing the relative frequencies of the sample means obtained from the 1000 samples of $n = 50$ in the middle panel of Figure 4.2 on a normal probability density function with $\mu = .2$ and $\sigma^2 = 1.8/50 = .0036$. Notice that the sampling distribution of the sample mean does not differ much from the normal distribution when we take account of the fact that the points representing the histogram are the center-points of the tops of its respective bars.

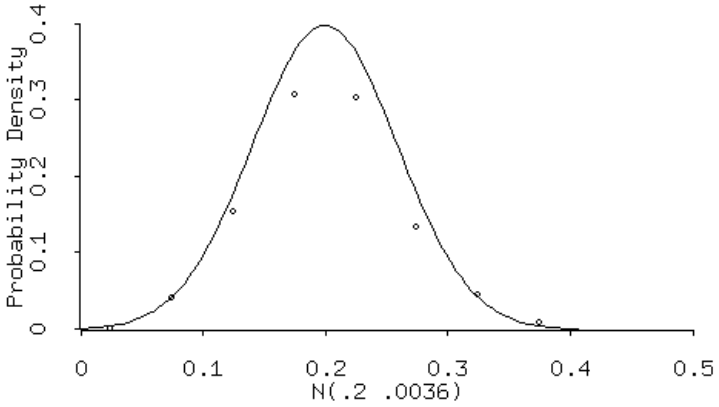


Figure 4.3: Relative frequencies of sample mean from 1000 samples of 50 plotted on normal density function with $\mu = .2$ and $\sigma^2_{\bar{X}} = .0036$.

It turns out that the similarity of the histograms to normal distributions as the sample size increases is not accidental. We have here a demonstration of the *Central Limit Theorem*. The Central Limit Theorem says that when the sample size is sufficiently large the sample mean $\bar{X}$ will become approximately normally distributed with mean equal to the population mean and variance equal to the population variance divided by the sample size. And the larger the sample size, the closer the approximation of the sampling distribution of $\bar{X}$ to a normal distribution. This holds true regardless of the distribution of the population provided it has a finite standard deviation.

The fact that the sample mean is normally distributed for large sample

sizes tells us that if the sample size is large enough the sample mean should lie within one standard deviation of the population mean 68% of the time and within two standard deviations of the population mean 95% of the time. The standard deviation referred to here is, of course, the standard deviation of the sample mean, not the standard deviation of the population.

The true standard deviation of the sample mean is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Since the population standard deviation is usually not known, we use

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

to provide an estimate of σ. The standard deviation of the sample mean is thus estimated as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

The Central Limit Theorem tells us the approximate nature of the sampling distribution of the sample mean when the sample is large and the distribution of the population is either unknown or the population is not normally distributed. If the population happens to be normally distributed the sampling distribution of the sample mean will turn out to be *exactly* normally distributed regardless of the sample size. This follows from two facts—first, that the mean of a sample from a normally distributed population is a linear function of the population elements in that sample, and second, that any linear function of normally distributed variables is normally distributed.

4.4 Point Estimation

The central purpose of statistical inference is to acquire information about characteristics of populations. An obvious source of information about a population mean is the mean of a random sample drawn from that population. When we use the sample mean to estimate the population mean the sample mean we obtain is called a *point estimate* of the population mean.

In general, suppose there is an unknown population characteristic or parameter that we will denote by θ. To estimate this parameter we select a simple random sample $X_1, X_2, X_3, \ldots, X_n$, from the population and then use some statistic S which is a function of these sample values as a point estimate of θ. For each possible sample we could take we will get a different set of sample values, $X_1, X_2, X_3, \ldots, X_n$, and hence a different S. The statistic S is thus a random variable that has a probability distribution which we

call the sampling distribution of S. We call S an *estimator* of θ. When we take our sample and calculate the value of S for that sample we obtain an *estimate* of θ.

Notice the difference between an estimate and an estimator. An *estimator* is a random variable used to estimate a population characteristic. An actual numerical value obtained for an estimator is an *estimate*.

Consider, for example, a trade association that needs to know the mean number of hourly paid employees per member firm, denoted by μ. To estimate this the association takes a random sample of $n = 225$ member firms (a tiny fraction of the total number of firms belonging to the association). The sample mean $\bar{X}$ is used as an *estimator* of μ. The *estimate* of μ is the particular value of $\bar{X}$ obtained from the sample, say, 8.31.

Note that the sample mean is only one possible estimator of the population mean. We could instead use the sample median or, perhaps, the average of largest and smallest values of X in the sample.

It should be evident from the discussion above that we are using

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

as an estimator of the population standard deviation σ. As an alternative we might think of using

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}.$$

Why should we use $\bar{X}$ rather than, say, the sample median, as an estimator of μ? And why should we use s rather than $\hat{s}$ as an estimator of σ?

4.5 Properties of Good Point Estimators

There are essentially three criteria which we use to select good estimators. The problem that arises, of course, is that a particular estimators may be better than another under one criterion but worse than that other estimator under another criterion.

4.5.1 Unbiasedness

An estimator is *unbiased* if the mean of its sampling distribution is equal to the population characteristic to be estimated. That is, S is an unbiased

estimator of θ if

$$E\{S\} = \theta.$$

If the estimate is biased, the bias equals

$$B = E\{S\} - \theta.$$

The median, for example, is a biased estimator of the population mean when the probability distribution of the population being sampled is skewed. The estimator

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

turns out to be a biased estimator of σ^2 while the estimator

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

is unbiased. This explains why we have been using s^2 rather than $\hat{s}^2$.

Unbiasedness in point estimators refers to the tendency of sampling errors to balance out over all possible samples. For any one sample, the sample estimate will almost surely differ from the population parameter. An estimator may still be desirable even if it is biased when the bias is not large because it may have other desirable properties.

4.5.2 Consistency

An estimator is a *consistent* estimator of a population characteristic θ if the larger the sample size the more likely it is that the estimate will be close to θ. For example in the shoe-pair testing example above, $\bar{X}$ is a consistent estimator of μ because its sampling distribution tightens around $\mu = .2$ as n increases. More formally, S is a consistent estimator of population characteristic θ if for any small positive value ϵ,

$$\lim_{n\to\infty}(P(|S - \theta| < \epsilon) = 1.$$

4.5.3 Efficiency

The *efficiency* of an *unbiased* estimator is measured by the variance of its sampling distribution. If two estimators based on the same sample size are both unbiased, the one with the smaller variance is said to have greater *relative efficiency* than the other. Thus, S_1 is relatively more efficient than S_2 in estimating θ if

$$\sigma^2\{S_1\} < \sigma^2\{S_2\} \qquad \text{and} \qquad E\{S_1\} = E\{S_2\} = \theta$$

For example, the sample mean and sample median are both unbiased estimators of the mean of a normally distributed population but the mean is a relatively more efficient estimator because at any given sample size its variance is smaller.

4.6 Confidence Intervals

Point estimates have the limitation that they do not provide information about the *precision* of the estimate—that is, about the error due to sampling. For example, a point estimate of 5 miles per gallon of fuel consumption obtained from a sample of 10 trucks out of a fleet of 400 would be of little value if the range of sampling error of the estimate is 4 miles per gallon—this would imply that the fuel consumption of the fleet could be anywhere between 1 and 9 miles per gallon. To provide an indication of the precision of a point estimate we combine it with an *interval estimate*. An interval estimate of the population mean μ would consist of two bounds within which μ is estimated to lie:

$$L \leq \mu \leq U$$

where L is the *lower bound* and U is the *upper bound*. This interval gives an indication of the degree of precision of the estimation process.

To obtain an estimate of how far the sample mean is likely to deviate from the population mean—i.e., how tightly it is distributed around the population mean—we use our estimate of the variance of the sample mean,

$$s_{\bar{x}}^2 = \frac{s^2}{n}.$$

This enables us to say that if the sample is large enough, $\bar{X}$ will lie within a distance of $\pm 2s$ of μ with probability .95.

Take, for example, the above-mentioned trade-association problem where a random sample of 225 firms was selected to estimate the mean number of hourly paid employees in member firms. Suppose the estimators $\bar{X}$ of μ and s of σ yield point estimates $\bar{X} = 8.31$ and $s = 4.80$. Since the sample size is quite large we can reasonably expect that in roughly 95 percent of such samples the sample mean will fall within $2s/\sqrt{n} = 9.60/15 = .64$ paid employees of μ in either direction. It would thus seem reasonable that by starting with the sample mean 8.31 and adding and subtracting .64 we should obtain an interval [7.67 — 8.95] which is likely to include μ.

If we take many large samples and calculate intervals extending two standard deviations of the sample mean on either side of that sample mean for each sample using the estimates of $\bar{X}$ and $s_{\bar{x}}$ obtained, about 95% of these intervals will bracket μ. The probability that any interval so obtained will bracket μ is roughly .95 (actually .9548).

More formally, consider an interval estimate $L \leq \mu \leq U$ with a specific probability $(1 - \alpha)$ of bracketing μ. The probability that a correct interval estimate (i.e., one that actually brackets μ) will be obtained is called a *confidence coefficient* and is denoted by $(1 - \alpha)$. The interval $L \leq \mu \leq U$ is called a *confidence interval* and the limits L and U are called the *lower* and *upper confidence limits*, respectively. The numerical confidence coefficient is often expressed as a percent, yielding the $100\,(1 - \alpha)\%$ confidence interval.

The confidence limits U and L for the population mean μ with approximate confidence coefficient $(1 - \alpha)$ when the random sample is reasonably large are

$$\bar{X} \pm z\,\frac{s}{\sqrt{n}}$$

where $z = z\,(1 - \alpha/2)$ is the $100\,(1 - \alpha/2)$ percentile of the standard normal distribution. The $100\,(1 - \alpha)$ percent confidence interval for μ is

$$\bar{X} - z\,\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z\,\frac{s}{\sqrt{n}}$$

Note that the confidence interval does *not* imply that there is a probability $(1 - \alpha)$ that μ will take a value between the upper and lower bounds. The parameter μ is not a variable—it is fixed where it is. Rather, there is a probability $(1 - \alpha)$ that the interval will bracket the *fixed* value of μ. The limits $-z\,(1 - \alpha/2)$ and $z\,(1 - \alpha/2)$ are given by the innermost edges of the shaded areas on the left and right sides of Figure 4.4. The shaded areas each contain a probability weight equal to $\alpha/2$. So for a 95% confidence interval these areas each represent the probability weight $(1 - .95)/2 = .05/2 = .025$ and the sum of these areas represents the probability weight .05. The area under the probability density function between the two shaded areas represents the probability weight .95. Note also that the probability $(1 - \alpha)$ is chosen in advance of taking the sample. The actual confidence interval calculated once the sample is taken may or may not bracket μ. If it does, the confidence interval is said to be correct.

What confidence coefficient should be chosen? This question hinges on how much risk of obtaining an incorrect interval one wishes to bear. In the trade-association problem above the 90, 95, and 99 percent confidence intervals are

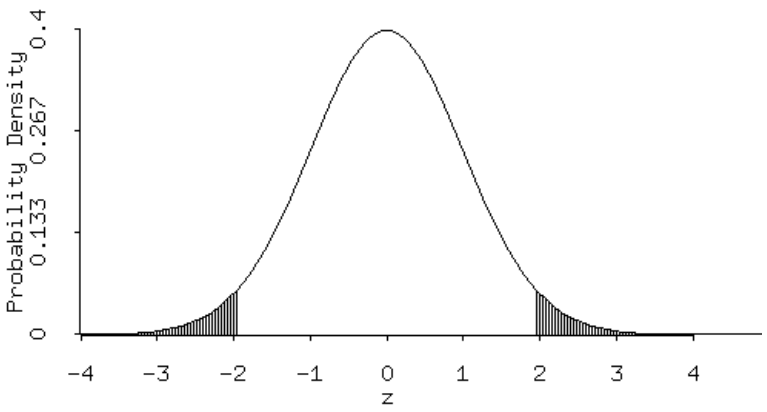


Figure 4.4: The areas $(1-\alpha)$ and $\alpha/2$ (shaded) for a standard normal probability distribution with $\alpha = .05$.

$1-\alpha$	$(1-\alpha/2)$	z	$s_{\bar{x}}$	$zs_{\bar{x}}$	$\bar{X}$	$\bar{X}+zs_{\bar{x}}$	$\bar{X}-zs_{\bar{x}}$
.90	.950	1.645	.32	.5264	8.31	8.84	7.78
.95	.975	1.960	.32	.6272	8.31	8.94	7.68
.99	.995	2.576	.32	.8243	8.31	9.13	7.48

Note that greater confidence in our results requires that the confidence interval be larger—as $(1-\alpha)$ gets bigger, $\alpha/2$ gets smaller and z must increase. We could, of course, narrow the confidence interval at every given level of confidence by increasing the sample size and thereby reducing $s/\sqrt{n}$.

4.7 Confidence Intervals With Small Samples

In making all the above calculations we standardised the sampling distribution of $\bar{X}$, obtaining

$$z = \frac{(\bar{X}-\mu)}{s/\sqrt{n}}$$

and then calculated limits for μ based on values for z in the table of standard normal probabilities. We used s as an estimator of σ. Had we known σ the standardised value would have been

$$z = \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} = -\frac{\mu}{\sigma/\sqrt{n}} + \frac{1}{\sigma/\sqrt{n}}\,\bar{X}.$$

Statistical theory tells us that when the population is normally distributed $\bar{X}$ is normally distributed because it is a linear function of the normally distributed X_i. Then the standardised value z is also normally distributed because it is a linear function of the normally distributed variable $\bar{X}$. But when we use s as an estimator of σ the above expression for z becomes

$$z = -\frac{\mu}{s/\sqrt{n}} + \frac{1}{s/\sqrt{n}}\,\bar{X}.$$

Whereas the divisor $\sigma/\sqrt{n}$ is a constant, $s/\sqrt{n}$ is a random variable. This immediately raises the question of the normality of z.

It turns out that the variable

$$\frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

is distributed according to the t-distribution, which approximates the normal distribution when the sample size is large. The t-distribution is symmetrical about zero like the standardised normal distribution but is flatter, being less peaked in the middle and extending out beyond the standard normal distribution in the tails. An example is presented in Figure 4.5. The t-distribution has one parameter, v, equal to the degrees of freedom, which equals the sample size minus unity in the case at hand. It has mean zero and variance $v/(v-2)$ with $v > 2$.

Because the t-distribution approximates the normal distribution when the sample size is large and because the Central Limit Theorem implies that $\bar{X}$ is approximately normally distributed for large samples, we could use $z = (\bar{X}-\mu)/s_{\bar{x}}$ to calculate our confidence intervals in the previous examples. When the sample size is small, however, we must recognize that $(\bar{X}-\mu)/s_{\bar{x}}$ is actually distributed according to the t-distribution with parameter $v = n-1$ for samples of size n drawn from a normal population. We calculate the confidence interval using the same procedure as in the large sample case except that we now set

$$t = \frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

and use the appropriate percentile from the t-distribution instead of from the normal distribution.

More formally, we can state that the confidence limits for μ with confidence coefficient $(1 - \alpha)$, when the sample is small and the population is normally distributed or the departure from normality is not too marked, are

$$\bar{X} \pm t\, s_{\bar{x}}$$

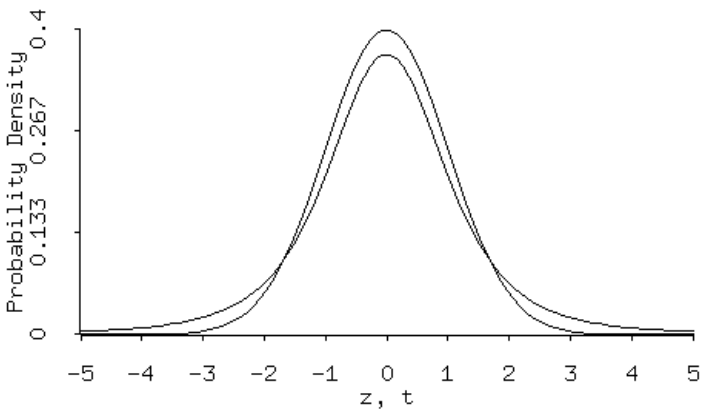


Figure 4.5: A t-distribution compared to the standard normal. The t-distribution is the flatter one with the longer tails.

where $t = t(1 - \alpha/2; n - 1)$. Expressing t in this way means that the value of t chosen will be the one with degrees of freedom $n - 1$ and percentile of the distribution $100(1 - \alpha/2)$.

Now consider an example. Suppose that the mean operating costs in cents per mile from a random sample of 9 vehicles (in a large fleet) turns out to be 26.8 and a value of s equal to 2.5966 is obtained. The standard deviation of the mean is thus $s/3 = .8655$. We want to estimate μ, the mean operating costs of the fleet. For a 90% confidence interval, $t(0.95; 8) = 1.860$. This implies a confidence interval of

$$26.80 \pm (1.8860)(.8655)$$

or

$$25.19 \leq \mu \leq 28.41.$$

Had the normal distribution been used, z would have been 1.645, yielding a confidence interval of

$$26.80 \pm 1.4237$$

or

$$25.38 \leq \mu \leq 28.22.$$

Inappropriate use of the normal distribution would give us a narrower interval and a degree of 'false confidence'.

Notice that the use of the t-distribution requires that the population be normal or nearly so. If the population is non-normal and n is large we can use z and the standard normal distribution. What do we do if the population is non-normal and the sample size is small? In this case we "cross our fingers" and use the t-distribution and allow that the confidence coefficient is now only approximately $1 - \alpha$. This assumes that the t-distribution is *robust*—i.e., applies approximately for many other populations besides normal ones. Essentially we are arguing, and there is disagreement among statisticians about this, that the distribution of $(\bar{X} - \mu)/s_{\bar{x}}$ is better approximated by the t-distribution than the normal distribution when the population is non-normal and the sample size is small.

4.8 One-Sided Confidence Intervals

Sometimes we are interested in an upper or lower bound to some population parameter. For example, we might be interested in the upper limit of fuel consumption of trucks in a fleet. One-sided confidence intervals are constructed the same as two-sided intervals except that all the risk that the interval will not bracket μ, given by α, is placed on one side. We would thus set a single lower confidence interval at $\bar{X} - z(1 - \alpha)s_{\bar{x}}$ instead of $\bar{X} - z(1 - \alpha/2)s_{\bar{x}}$. A single upper-confidence interval is set in similar fashion. Of course, for small samples we would use t instead of z.

4.9 Estimates of a Population Proportion

When the sample size is large the above methods apply directly to point and interval estimation of a population proportion. Suppose that we want to estimate the proportion of voters who will vote yes in the next referendum on whether Quebec should become independent from the rest of Canada. It is natural to take a large sample of voters to determine the sample proportion $\bar{p}$ that are in favour of independence. The Central Limit Theorem tells us that this sample proportion should be normally distributed around the population proportion p if the sample size is large enough. To construct a confidence interval we then need an estimate of the standard deviation of $\bar{p}$. Since the total number of people in the sample voting for independence, X, is distributed according to the binomial distribution with parameters n and p, its variance is $np\,(1 - p)$. The variance of the sample proportion $\bar{p}$ then

equals

$$\begin{aligned} Var\{\bar{p}\} &= Var\{\frac{X}{n}\} = \frac{1}{n^2} Var\{X\} \\ &= \frac{1}{n^2} np(1-p) = \frac{p\,(1-p)}{n}. \end{aligned} \tag{4.5}$$

It is natural to estimate the standard deviation of $\bar{p}$ as the square root of the above expression with $\bar{p}$ substituted for p. When we do so we divide by $n-1$ rather than n. This recognizes the fact that

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n-1}}$$

turns out to be an unbiased estimator of $\sigma^2_{\bar{p}}$ whereas

$$\tilde{s}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

is a biased estimator. The $100\,(1-\alpha)$ confidence interval for p therefore becomes

$$\bar{p} \pm z\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

where z is the value from the standard normal table that will produce the appropriate percentile $100\,(1-\alpha/2)$ for a two-sided confidence interval or $100\,(1-\alpha)$ for a one-sided confidence interval. Suppose that we took a random sample of 1000 voters and found that 350 of them would vote for making Quebec into a separate country. This yields $\bar{p} = .35$ as a point estimate of p. The standard deviation of $\bar{p}$ is estimated to be $\sqrt{(.35)(.65)/999} = .015083$. A two-sided 95% confidence interval for p, for which $z = z(1-\alpha/2) = z(.975) = 1.96$, thus becomes

$$[.35 - (1.96)(.015083)] \leq p \leq [.35 + (1.96)(.025083)]$$

$$.3204 \leq p \leq .3796.$$

4.10 The Planning of Sample Size

If we know the confidence we require in our results we can choose the sample size that will yield that confidence. Resources need not be wasted selecting an excessively large sample while at the same time the risk of choosing an uninformative sample can be avoided. We assume that the sample selected will be reasonably large in absolute value but a small fraction of the population. Let us call the distance between the sample mean and the upper (or lower) confidence limit the half-width (which is half the distance between the upper and lower limits) and denote it by h. The upper limit will then be

$$\bar{X} + h = \bar{X} + z\,\frac{\sigma}{\sqrt{n}}$$

where σ is a value of the population standard deviation picked for planning purposes, so that

$$h = z\,\frac{\sigma}{\sqrt{n}}.$$

Squaring both sides and then multiplying them by n yields

$$n\,h^2 = z^2\sigma^2$$

so that

$$n = \frac{z^2\sigma^2}{h^2}.$$

In formal terms we can thus state that the necessary random sample size to achieve the desired half-width h for the specified confidence coefficient $(1-\alpha)$ for a given planning value of the population standard deviation σ is

$$n = \frac{z^2\sigma^2}{h^2} \tag{4.6}$$

where $z = z(1-\alpha/2)$ and the half-width h represents the deviation of each interval from the sample mean. In the case of a one-sided confidence interval, h would equal the entire interval.

Consider an example. Suppose that a nationwide survey of physicians is to be undertaken to estimate μ, the mean number of prescriptions written per day. The desired margin of error is $\pm.75$ prescriptions, with a 99% confidence coefficient. A pilot study indicated that a reasonable value for the population standard deviation is 5. We therefore have $z = z(1-.01/2) = z(.995) = 2.575$, $h = .75$ and $\sigma = 5$. The proper sample size then equals

$$n = [(2.575)(5)]^2/(.75)^2 = (12.88)^2/.5625 = 165.89/.5625 = 295.$$

The same general principles apply to choosing the sample size required to estimate a population proportion to the desired degree of accuracy. Consider a poll to estimate the results of the next Quebec referendum. How big a sample will we need to estimate the proportion of the voters that will vote for separation to an accuracy of ±2 percentage points, 19 times out of 20? The ratio 19/20 = .95 provides us with $(1 - \alpha)$. We can obtain a planning value of $\sigma_{\bar{p}}$ by noting that $\sqrt{p(1-p)/n}$ will be a maximum when $p = .5$ and using this value of p to obtain the standard deviation of $\bar{p}$ for planning purposes.[1] Thus, a deviation of 2 percentage points or .02 from p must equal $z(1 - \alpha/2) = z(1 - .05/2) = z(.975)$, multiplied by $\sigma_{\bar{p}} = \sqrt{p(1-p)/n} = \sqrt{(.5)(.5)}/\sqrt{n} = .5/\sqrt{n}$. Letting U be the upper confidence limit, we thus have

$$U - \bar{p} = .02 = z\sqrt{\frac{p(1-p)}{n}} = \frac{(1.96)(.5)}{\sqrt{n}} = \frac{.98}{\sqrt{n}},$$

which implies that

$$\sqrt{n} = \frac{.98}{.02} = 49.$$

The appropriate sample size is therefore $(49)^2 = 2401$.

4.11 Prediction Intervals

Sometimes we want to use sample data to construct an interval estimate for a new observation. Consider the earlier problem of determining the operating costs for a vehicle fleet. Having established a confidence interval regarding the operating costs of vehicles in the fleet, we can use the same evidence to help determine whether a particular vehicle not in the sample meets standards.

Suppose that the vehicle in question is selected independently of our earlier random sample of 9 vehicles. Let the operating costs of this vehicle be X_{new}. And suppose that the population (i.e., the operating costs in cents per mile of all vehicles in the fleet) follows a normal distribution.

Now if we knew the values of μ and σ for the population the calculation of a prediction interval would be very simple. We simply obtain a value of z equal to the number of standard deviations from the mean of a normal distribution that would meet our desired level of confidence—that is,

[1] It can be easily seen that $(.4)(.6) = (.6)(.4) = .24 < (.5)(.5) = .25$ and that values of p less than .4 or greater than .6 yield even smaller values for $p(1 - p)$.

$z = z(1-\alpha/2)$, where $100\,(1-\alpha)$ is our desired level of confidence—and calculate $\mu \pm z\,\sigma$. We would predict that $100\,(1-\alpha)\%$ of the time X_{new} will fall in this interval. If X_{new} does not fall in this interval we can send the vehicle in for service on the grounds that the chance is no more than $100\,\alpha/2$ percent (looking at the upper tail) that its cost per mile is equal to or less than the mean for the fleet.

The problem is that we do not know μ and σ and have to use the sample statistics $\bar{X}$ and s as estimators. To calculate the prediction interval we have to know the standard deviation of X_{new}. The estimated variance of X_{new} is

$$\begin{aligned} s^2\{X_{new}\} &= E\{(X_{new}-\mu)^2\} = E\{[(X_{new}-\bar{X})+(\bar{X}-\mu)]^2\} \\ &= E\{(X_{new}-\bar{X})^2\} + E\{(\bar{X}-\mu)^2\} \\ &= s^2 + \frac{s^2}{n} = [1+\frac{1}{n}]\,s^2. \end{aligned}$$

The prediction interval for X_{new} then becomes

$$\bar{X} \pm t\,s\{X_{new}\}$$

where $t = t(1-\alpha/2; n-1)$ is the 'number of standard deviations' obtained from the t-distribution table for the probability weight $(1-\alpha/2)$ and degrees of freedom $(n-1)$. In the case of a vehicle selected from the fleet,

$$\bar{X} \pm t(.975;8)\,s\{X_{new}\} = 26.80 \pm (2.306)\sqrt{(1+1/9)}\,(2.5966)$$

$$= 26.80 \pm (2.306)(1.05409)(2.5966) = 26.80 \pm 6.31$$

which yields

$$20.49 \leq \mu \leq 33.11.$$

Notice that the prediction interval is much wider than the 95% confidence interval for $\bar{X}$ which would be

$$26.80 \pm (2.306)\frac{s}{\sqrt{n}} = 26.80 \pm (2.306)(.8655) = 26.80 \pm 3.1715$$

or

$$23.63 \leq 26.80 \leq 29.97.$$

This is the case because there are two sources of deviation of X_{new} from μ—the deviation from the sample mean, taken as a point estimate of μ, and the deviation of that sample mean from μ. The confidence interval for the sample mean only includes the second source of deviation.

4.12 Exercises

1. Find the following probabilities for the standard normal random variable z:

a) $P(-1 \leq z \leq 1)$

b) $P(-2 \leq z \leq 2)$

c) $P(-2.16 \leq z \leq .55)$

d) $P(-.42 < z < 1.96)$

e) $P(z \geq -2.33)$

f) $P(z > 2.33)$

2. Suppose that a random sample of n measurements is selected from a population with mean $\mu = 100$ and variance $\sigma^2 = 100$. For each of the following values of n, give the mean and standard deviation of the sampling distribution of the sample mean $\bar{X}$.

a) $n = 4$.

b) $n = 25$.

c) $n = 100$.

d) $n = 50$.

e) $n = 50$.

f) $n = 500$.

g) $n = 1000$.

3. A particular experiment generates a random variable X that has only two outcomes: $X = 1$ (success) with probability $p = 0.6$ and $X = 0$ (failure) with probability $(1 - p) = .4$. Consider a random sample consisting of $n = 3$ independent replications of this experiment. Find the exact sampling distribution of the sample mean.

4. Write down the Central Limit Theorem and explain what it means.

5. The mean and standard deviation of a random sample of n measurements are equal to 33.9 and 3.3 respectively.

a) Find a 95% confidence interval for μ if $n = 100$. (33.2532, 34.5468)

b) Find a 95% confidence interval for μ if $n = 400$.

c) What is the effect on the width of the confidence interval of quadrupling the sample size while holding the confidence coefficient fixed?

6. Health insurers and the federal government are both putting pressure on hospitals to shorten the average length of stay of their patients. In 1993 the average length of stay for men in the United States was 6.5 days and the average for women was 5.6 days (*Statistical Abstract of the United States: 1995*). A random sample of 20 hospitals in one state had a mean length of stay for women in 1996 of 3.6 days and a standard deviation of 1.2 days.

a) Use a 90% confidence interval to estimate the population mean length of stay for women in the state's hospitals in 1996.

b) Interpret the interval in terms of this application.

c) What is meant by the phrase '90% confidence interval'?

7. The population mean for a random variable X is $\mu = 40$. The population variance is $\sigma^2 = 81$. For a (large) random sample of size n drawn from this population, find the following:

a) The expected value and the variance of the sample mean $\bar{X}$ when $n = 36$.

b) The probability that $P(\bar{X} \geq 41)$ in the above case.

c) The probability $P(38.5 \leq \bar{X} \leq 40.5)$ when $n = 64$.

8. A number of years ago, Lucien Bouchard and John Charest were in a tough fight for the premiership of Quebec. How big a simple random sample would have been needed to estimate the proportion of voters that would vote for Bouchard to an accuracy of ± 1 percentage points, 19 times out of 20?

9. One of the continuing concerns of U.S. industry is the increasing cost of health insurance for its workers. In 1993 the average cost of health premiums

per employee was $2,851, up 10.5% from 1992 (*Nation's Business*, Feb. 1995). In 1997, a random sample of 23 U.S. companies had a mean health insurance premium per employee of $3,321 and a standard deviation of $255.

a) Use a 95% confidence interval to estimate the mean health insurance premium per employee for all U.S. companies.

b) What assumption is necessary to ensure the validity of the confidence interval?

c) Make an inference about whether the true mean health insurance premium per employee in 1997 exceeds $2,851, the 1993 mean.

10. The mean and the standard deviation of the annual snowfalls in a northern city for the past 20 years are 2.03 meters and 0.45 meters, respectively. Assume that annual snowfalls for this city are random observations from a normal population. Construct a 95 percent prediction interval for next year's snowfall. Interpret the prediction interval.

11. Accidental spillage and misguided disposal of petroleum wastes have resulted in extensive contamination of soils across the country. A common hazardous compound found in the contaminated soil is benzo(a)pyrene [B(a)p]. An experiment was conducted to determine the effectiveness of a treatment designed to remove B(a)p from the soil (*Journal of Hazardous Materials*, June 1995). Three soil specimens contaminated with a known amount of B(a)p were treated with a toxin that inhibits microbial growth. After 95 days of incubation, the percentage of B(a)p removed from each soil specimen was measured. The experiment produced the following summary statistics: $\bar{X} = 49.3$ and $s = 1.5$.

a) Use a 99% confidence interval to estimate the mean percentage of B(a)p removed from a soil specimen in which toxin was used.

b) Interpret the interval in terms of this application.

c) What assumption is necessary to ensure the validity of this confidence interval?

4.13 Appendix: Maximum Likelihood Estimators

The *Maximum Likelihood Method* is a general method of finding point estimators with desirable qualities.

Let us proceed by using an example. Suppose we know that the number of annual visits to a dentist by a child is a Poisson random variable X with unknown parameter λ. In a random sample of two children the numbers of visits to the dentist last year were $X_1 = 0$ and $X_2 = 3$.

The idea of maximum likelihood is to choose the value for λ for which it is most likely that we would observe the sample $\{X_1, X_2\}$. We do this by calculating the probability of observing the sample for various values of λ—say, 0, 1, 1.5, 2, 3, etc.—and picking the value of λ that maximizes this probability. The Poisson probability function, defined in equation (3.32), is

$$P(x) \quad = \quad \frac{\lambda^x \, e^{-\lambda}}{x!}.$$

Since the observations are independent of each other, the probability of observing the sample $\{X_1, X_2\}$ is $P(x = X_1)$ times $P(x = X_2)$. From the table of Poisson probabilities we obtain the following probabilities for various values of λ:

λ	$P(x = 0)$	$P(x = 3)$	$P(x = 0)P(x = 3)$
0.0	.0000	.0000	.0000
1.0	.3679	.0613	.0225
1.5	.2231	.1255	.0280
2.0	.1353	.1804	.0244
3.0	.0498	.2240	.0112

The value of λ that maximizes the likelihood of observing the sample in the above table is $\lambda = 1.5$.

We could calculate $P(x = 0)P(x = 3)$ for values of λ between the ones in the table above and plot them to obtain the smooth curve in Figure 4.6. This curve maps the probability density as a function of λ which is called the *likelihood function*. It confirms that 1.5 is the maximum likelihood estimate of λ.

Let us now approach the problem more formally and suppose that we have a set of sample observations X_i from which we want to estimate a parameter θ. There is some probability

$$P(X_1, X_2, X_3, \ldots, X_n; \theta)$$

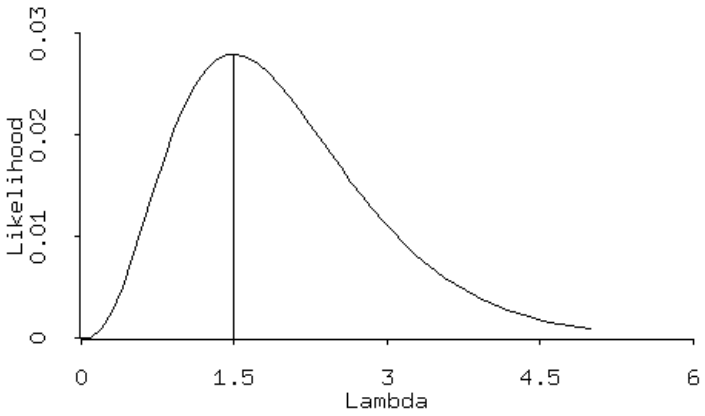


Figure 4.6: The likelihood function for the children-to-the dentist example.

of drawing a particular sample of observations, given the magnitude of the unknown parameter θ. Because the sample observations $X_1, X_2, X_3, \ldots, X_n$ are independent, this probability function equals

$$P(X_1, X_2, X_3, \ldots, X_n; \theta) = P(X_1; \theta)P(X_2; \theta)P(X_3; \theta) \ldots P(X_n; \theta).$$

This product of probabilities, when viewed as a function of θ for given $X_1, X_2, X_3, \ldots, X_n$ is called the *likelihood function*

$$L(\theta) \quad = \quad P(X_1; \theta)P(X_2; \theta)P(X_3; \theta) \ldots P(X_n; \theta). \tag{4.7}$$

We find the value of θ that maximizes $L(\theta)$ either by analytic methods or, when that approach is not feasible, by efficient numerical search procedures.

Consider a Poisson process with unknown parameter λ and select a random sample $X_1, X_2, X_3, \ldots, X_n$. Using the formula for the Poisson probability distribution, the likelihood function can be expressed

$$\begin{aligned} L(\theta) &= \left[\frac{\lambda^{X_1} e^{-\lambda}}{X_1!}\right]\left[\frac{\lambda^{X_2} e^{-\lambda}}{X_2!}\right] \ldots\ldots \left[\frac{\lambda^{X_n} e^{-\lambda}}{X_n!}\right] \\ &= \left[\frac{\lambda^{\sum X_i} e^{-n\lambda}}{X_1! X_2! \ldots X_n!}\right] = \left[\frac{\lambda^{n\bar{X}} e^{-n\lambda}}{X_1! X_2! \ldots X_n!}\right]. \end{aligned} \tag{4.8}$$

To maximize $L(\lambda)$ we differentiate it with respect to λ and find the value for λ for which this differential is zero. Differentiating (using the chain rule

whereby $dxy = xdy + ydx$) we have

$$\begin{aligned}
\frac{dL(\theta)}{d\theta} &= \frac{1}{X_1!X_2!\ldots X_n!}\left[\frac{d}{d\lambda}\left(\lambda^{n\bar{X}}\, e^{-n\lambda}\right)\right] \\
&= \frac{1}{X_1!X_2!\ldots X_n!}\left[\lambda^{n\bar{X}}\frac{d}{d\lambda}\left(e^{-n\lambda}\right) + e^{-n\lambda}\frac{d}{d\lambda}\left(\lambda^{n\bar{X}}\right)\right] \\
&= \frac{1}{X_1!X_2!\ldots X_n!}\left[-\lambda^{n\bar{X}}\, e^{-n\lambda}\, n + e^{-n\lambda}\, n\bar{X}\, \lambda^{n\bar{X}-1}\right] \\
&= \frac{1}{X_1!X_2!\ldots X_n!}\left[n\left(\frac{\bar{X}}{\lambda} - 1\right)\left(\lambda^{n\bar{X}}\, e^{-n\lambda}\right)\right] \qquad (4.9)
\end{aligned}$$

This expression equals zero—i.e., $L(\lambda)$ is a maximum—when

$$\left[\frac{\bar{X}}{\lambda} - 1\right] = 0,$$

which occurs when $\lambda = \bar{X}$. Thus, the sample mean is a maximum likelihood estimator of λ for a random sample from a Poisson distribution. In the children-to-dentist example above, the sample mean is $(0 + 3)/2 = 1.5$, the value of λ that produced the largest value for $L(\lambda)$ in Figure 4.6.

Chapter 5

Tests of Hypotheses

In the previous chapter we used sample statistics to make point and interval estimates of population parameters. Often, however, we already have some theory or hypothesis about what the population parameters are and we need to use our sample statistics to determine whether or not it is reasonable to conclude that the theory or hypothesis is correct. Statistical procedures used to do this are called *statistical tests*.

Consider, for example, the case of a firm that has developed a diagnostic product for use by physicians in private practice and has to decide whether or not to mount a promotional campaign for the product. Suppose that the firm knows that such a campaign would lead to higher profits only if the mean number of units ordered per physician is greater than 5. Office demonstrations are conducted with a random sample of physicians in the target market in order to decide whether or not to undertake the campaign. The campaign is very costly and the firm will incur substantial losses if it undertakes it only to find that the mean number of orders after the campaign is less than or equal to 5.

5.1 The Null and Alternative Hypotheses

We can think of two possibilities. The mean number of orders in the population of all physicians will exceed 5 or the mean will not exceed 5. Suppose the firm accepts the hypothesis that the mean number of orders in the population will be greater than 5 when it turns out to be less. A promotional campaign will be conducted at great loss. Had the guess that the mean number of orders will be greater than 5 been correct the firm would have earned a substantial profit. Alternatively, if the firm accepts the hypothesis

that the mean number of orders in the population will be less than 5 when it turns out to be greater, some profit will be foregone. Had the guess that the mean number of orders in the population will be less than 5 been correct, however, huge losses from the promotional campaign will have been avoided. It turns out that the cost of guessing that the mean number of orders will be greater than 5, mounting the promotional campaign, and being wrong is much greater than the cost of guessing that the mean number of orders will be less than or equal to 5, not mounting the promotional campaign, and being wrong.

We call the more serious of the two possible errors a *Type I error* and the least serious error a *Type II error*. We call the hypothesis which *if wrongly rejected* would lead to the more serious (Type I) error the *null hypothesis* and denote it by the symbol H_0. The other hypothesis, which if wrongly rejected would lead to the less serious (Type II) error, we call the *alternative hypothesis* and denote it by the symbol H_1.

In the problem we have been discussing, the most serious error will occur if the mean number of orders in the population of physicians will be less than 5 and the firm erroneously concludes that it will be greater than 5. Hence, the null hypothesis is

$$H_0 \colon \mu \leq 5$$

and the alternative hypothesis is

$$H_1 \colon \mu > 5.$$

Acceptance of either hypothesis on the basis of sample evidence involves a risk, since the hypothesis chosen might be the incorrect one. We denote the probability of making a Type I error (incorrectly rejecting the null hypothesis) an α-risk and the probability of making a Type II error (incorrectly rejecting the alternative hypothesis) a β-risk. It turns out that if the sample size is predetermined (i.e., beyond the firm's control) the firm has to choose which risk to control. Control of the α-risk at a lower level will imply a greater degree of β-risk and vice versa. Since by construction Type I errors are the most damaging, the firm will obviously want to control the α-risk.

Of course, the situation could have been different. The market for the type of diagnostic product that the firm has developed may be such that the first firm providing it could achieve quite an advantage. An erroneous conclusion by the firm that the mean number of orders will be less than 5, and the resulting decision not to promote the product, could lead to the loss of substantial future market opportunities. On the other hand, if the

cost of the promotion is small, an erroneous conclusion that the number of orders per physician in the population will equal or exceed 5 would perhaps lead to a minor loss. In this case we would define the null and alternative hypotheses as

$$H_0 \colon \mu \geq 5$$

and

$$H_1 \colon \mu < 5.$$

A Type I error will then result when the null hypothesis is incorrectly rejected—i.e., when we erroneously conclude that the mean order per physician in the population will be less than 5 when it turns out to be equal to or greater than 5. The probability of this happening will be the α-risk. A Type II error will result when the alternative hypothesis is incorrectly rejected—i.e., when the firm erroneously concludes that the mean order per physician will be greater than or equal to 5 when it turns out not to be. The probability of this happening will be the β-risk.

The hypotheses in the above problem were one-sided alternatives. The crucial question was whether the population parameter μ was above a particular value μ_0 $(= 5)$ or below it. We can also have two sided alternatives.

Suppose it is found that the mean duration of failed marriages was 8.1 years before the divorce law was changed and we want to determine whether the new legislation has affected the length of time unsuccessful marriages drag on. A sociologist has a random sample of divorce records accumulated since the law was changed upon which to make a decision. Erroneously concluding that the new legislation has changed people's behaviour when it has not is judged to be a more serious error than incorrectly concluding that behaviour has not changed as a result of the new law when it in fact has. Accordingly, the sociologist chooses the null hypothesis as

$$H_0 \colon \mu = 0$$

and the alternative hypothesis as

$$H_1 \colon \mu \neq 0.$$

A Type I error will arise if the sociologist concludes that behaviour has changed when it has not—i.e., incorrectly rejects the null hypothesis—and a Type II error will arise if she erroneously concludes that behaviour has not changed when it in fact has. The probability of a Type I error will again be the α-risk and the probability of a Type II error the β-risk.

5.2 Statistical Decision Rules

Take the case of the diagnostic product discussed above where H_0 is $\mu \leq 5$ and H_1 is $\mu > 5$. If upon conducting the office demonstrations the mean number of orders of physicians in the sample is less than 5, it would be reasonable to accept the null hypothesis that $\mu \leq 5$. If the sample mean is greater than 5, however, should we reject the null hypothesis? Clearly, the costs of a Type I error are greater than the costs of a Type II error, so we would not want to reject the null hypothesis if the sample mean is just a little bit above 5 because the sample mean could be greater than 5 entirely as a result of sampling error. On the other hand, if the sample mean is 20, it might seem reasonable to reject the null hypothesis. The question is: At what value of the sample mean should we reject the null hypothesis that $\mu \leq 5$. That value of the mean (or *test statistic*) at which we decide (ahead of time, before the sample is taken) to reject the null hypothesis is called the *action limit* or *critical value*. The choice of this critical value is called a *statistical decision rule*.

The general form of the statistical decision rule for one-sided and two-sided alternatives is given in Figure 5.1. Possible values of the sample mean are divided into two groups along the continuum of values the sample mean can take. The groups are separated by the critical value A in the case of one-sided tests shown in the top two panels, or by the critical values A_1 and A_2 in the case of a two-sided test shown in the bottom panel. The region between the critical value or values and μ_0, the level of μ at which the test is being conducted, is called the *acceptance region*. The region on the other side(s) of the critical value(s) from μ_0 is called the *critical region* or *rejection region*. If the sample mean falls in the rejection region, we reject the null hypothesis and accept the alternative hypothesis. If it falls in the acceptance region we accept the null hypothesis and reject the alternative hypothesis. Note that acceptance of the null hypothesis means only that we will act *as if* it were true—it does not mean that the null hypothesis is in fact true.

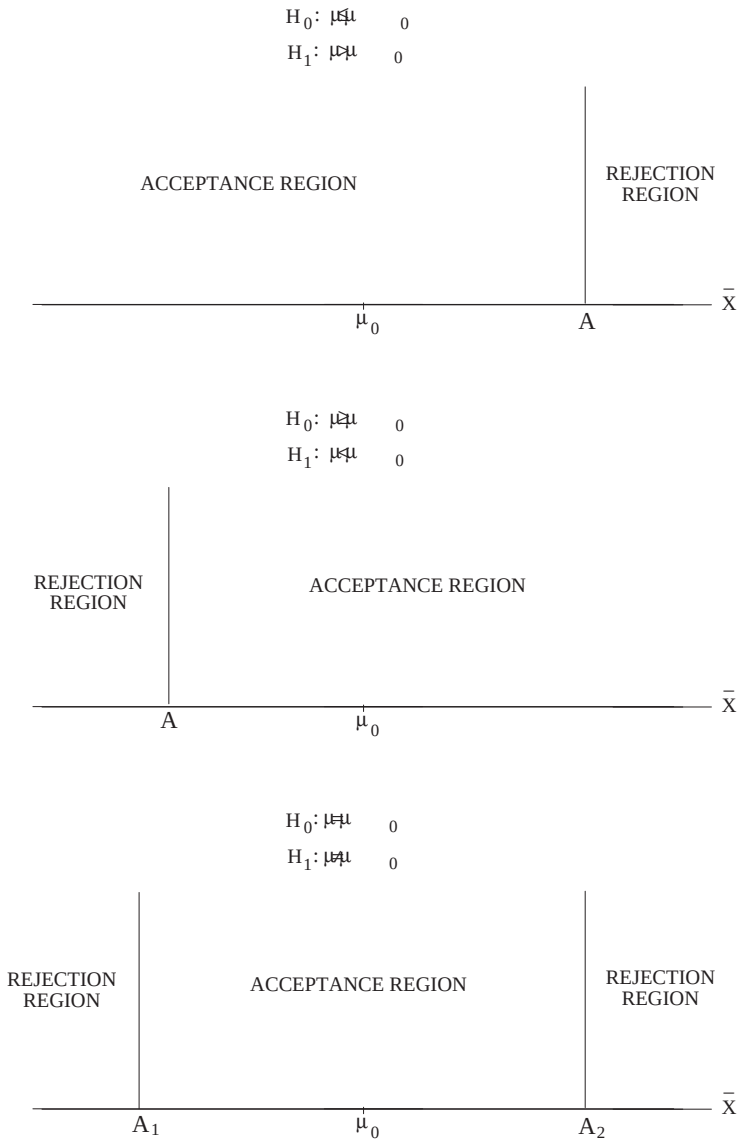


Figure 5.1: Ilustration of statistical decision rules for one-sided upper-tail (top), one-sided lower-tail (middle) and two-sided (bottom) alternatives concerning the population mean μ.

5.3 Application of Statistical Decision Rules

In order to actually perform the statistical test we must establish the degree of α-risk (risk of erroneously rejecting the null hypothesis) we are willing to bear. We must also make sure we are satisfied with the level of μ at which the α-risk is to be controlled—that is, with the level at which we set μ_0. In the example of the diagnostic product, we need not have set the level of μ at which the α-risk is to be controlled at 5. We could have been safer (in the case where the most costly error is to incorrectly conclude that the mean number of orders from the population of physicians is greater than 5 when it is in fact less than or equal to 5) to control the α-risk at $\mu_0 = 5.5$. At any given level of α-risk chosen this would have been a more stringent test than setting μ_0 at 5. We also have to establish the probability distribution of the standardised random variable $(\bar{X} - \mu)/s_{\bar{x}}$. If the sample size is large, the Central Limit Theorem tells us that it will be approximately normally distributed. If the sample size is small and the probability distribution of the population values X_i around μ is not too different from the normal distribution, $(\bar{X} - \mu)/s_{\bar{x}}$ will follow a t-distribution.

Suppose an airline takes a random sample of 100 days' reservation records which yields a mean number of no-shows on the daily flight to New York City of 1.5 and a value of s equal to 1.185. The resulting value of $s_{\bar{x}}$ is $1.185/\sqrt{100} = 1.185/10 = .1185$. The airline knows from extensive experience that the mean number of no-shows on other commuter flights is 1.32. The airline wants to test whether the mean number of no-shows on the 4 PM flight exceeds 1.32. We let H_0 be the null hypothesis that the mean number of no-shows is less than or equal to 1.320 and the alternative hypothesis H_1 be that the mean number of no shows exceeds 1.320. Notice that the hypothesis is about the number of no-shows in the whole population of reservations for the 4 PM flight to New York City. The airline wants to control the α-risk at .05 when $\mu = 1.320$. Since the sample is large

$$z = \frac{\bar{X} - \mu_0}{s_{\bar{x}}}$$

is approximately standard normal. The sample results in a value of z equal to

$$z^* = \frac{1.500 - 1.320}{.1185} = 1.519.$$

At an α-risk of .05 the critical value for z is 1.645 in a one-sided test. Thus, since z^* is less than the critical value we cannot reject the null hypothesis. We accept H_0 and reject H_1 since the standardised value of the sample mean

does not fall in the critical region. The probability of observing a sample mean of 1.50 when the population mean is 1.320 is more than .05. This is an example of a one-sided upper-tail test because the critical region lies in the upper tail of the distribution. For an example of a one-sided lower-tail test consider a situation where a customs department asks travellers returning from abroad to declare the value of the goods they are bringing into the country.

The authorities want to test whether the mean reporting error is negative—that is, whether travellers cheat by underreporting. They set the null hypothesis as H_0: $\mu \geq 0$ and the alternative hypothesis as H_1: $\mu < 0$. A random sample of 300 travellers yields $\bar{X} = -\$35.41$ and $s = \$45.94$. This implies $s_{\bar{x}} = 45.94/17.32 = 2.652$. The α-risk is to be controlled at $\mu_0 = 0$. The sample size is again so large that the test statistic is distributed approximately as the standardised normal distribution. The sample yields a value equal to

$$z^* = \frac{-35.41 - 0}{2.652} = -13.35.$$

The authorities want to control the α-risk at .001 so the critical value for z is -3.090. Since z^* is well within the critical region we can reject the null hypothesis H_0 that the mean reporting error is non-negative and accept the alternative hypothesis that it is negative. In fact, the observed sample mean is 13.35 standard deviations below the hypothesized population mean of zero while the critical value is only 3.090 standard deviations below zero. Note that the α-risk is only approximately .001 because z is only approximately normally distributed.

Now let us take an example of a two-sided test. Suppose that a random sample of 11 children out of a large group attending a particular camp are given a standard intelligence test. It is known that children of that age have mean scores of 100 on this particular test. The camp organizers want to know whether or not the children attending the camp are on average equal in intelligence to those in the population as a whole. Note that the relevant population here from which the sample is drawn is the entire group of children attending the camp. The sample mean score was $\bar{X} = 110$ and s was equal to 8.8, resulting in a value for $s_{\bar{x}}$ of $8.8/3.62 = 2.65$. Since the concern is about possible differences in intelligence in either direction the appropriate test is a two-tailed test of the null hypothesis H_0: $\mu = \mu_0 = 100$ against the alternative hypothesis H_1: $\mu \neq \mu_0 = 100$. With a small sample size, under the assumption that the distribution of the population is not too

far from normal,

$$\frac{\bar{X} - \mu}{s_{\bar{x}}}$$

will be distributed according to the t-distribution with 10 degrees of freedom. Suppose that the organizers of the camp want to control the α-risk at .05 at a value of $\mu_0 = 100$. Since the test is a two-tailed test the critical region has two parts, one at each end of the distribution, each containing probability weight $\alpha/2 = .025$ (the two together must have probability weight .05). This two-part region will contain those t-values greater than 2.228 and less than -2.228. The value of t that arises from the sample,

$$t^* = \frac{110 - 100}{2.65} = 3.77$$

clearly lies in the upper part of the critical region so that the null hypothesis that the intelligence level of the children in the camp is the same as that of those in the population as a whole must be rejected.

The decision rules for tests of μ can be shown in Figure 5.2. In the upper panel, which illustrates a one-sided upper-tail test, α is the probability that $\bar{X}$ will fall in the critical region if $\mu \leq \mu_0$. The area $1 - \alpha$ is the probability that $\bar{X}$ will fall in the acceptance region. If $\bar{X}$ in fact falls in the rejection region, the probability will be less than α of observing that value, given the sample size, if μ is really less than or equal to μ_0. The center panel does the same thing for a one-sided lower-tail test. Here, $\bar{X}$ must fall below A for the null hypothesis to be rejected. The bottom panel presents an illustration of a two-sided test. The null hypothesis is rejected if $\bar{X}$ falls either below A_1 or above A_2. The probability of rejecting the null hypothesis if $\mu = \mu_0$ is equal to $\alpha/2 + \alpha/2 = \alpha$. We reject the null hypothesis if the probability of observing a sample mean as extreme as the one we obtain conditional upon $\mu = \mu_0$ is less than α.

5.4 P–Values

In the statistical test involving the average intelligence of children at the camp the value of z that resulted from the sample was 3.77 whereas the critical value was ± 2.228. The probability of obtaining this sample from a population of children having mean intelligence of 100 is less than .05. An appropriate question is: What is the probability of observing a sample mean as extreme as the one observed if the mean intelligence of the population of children at the camp is 100? Or, to put it another way, what level of

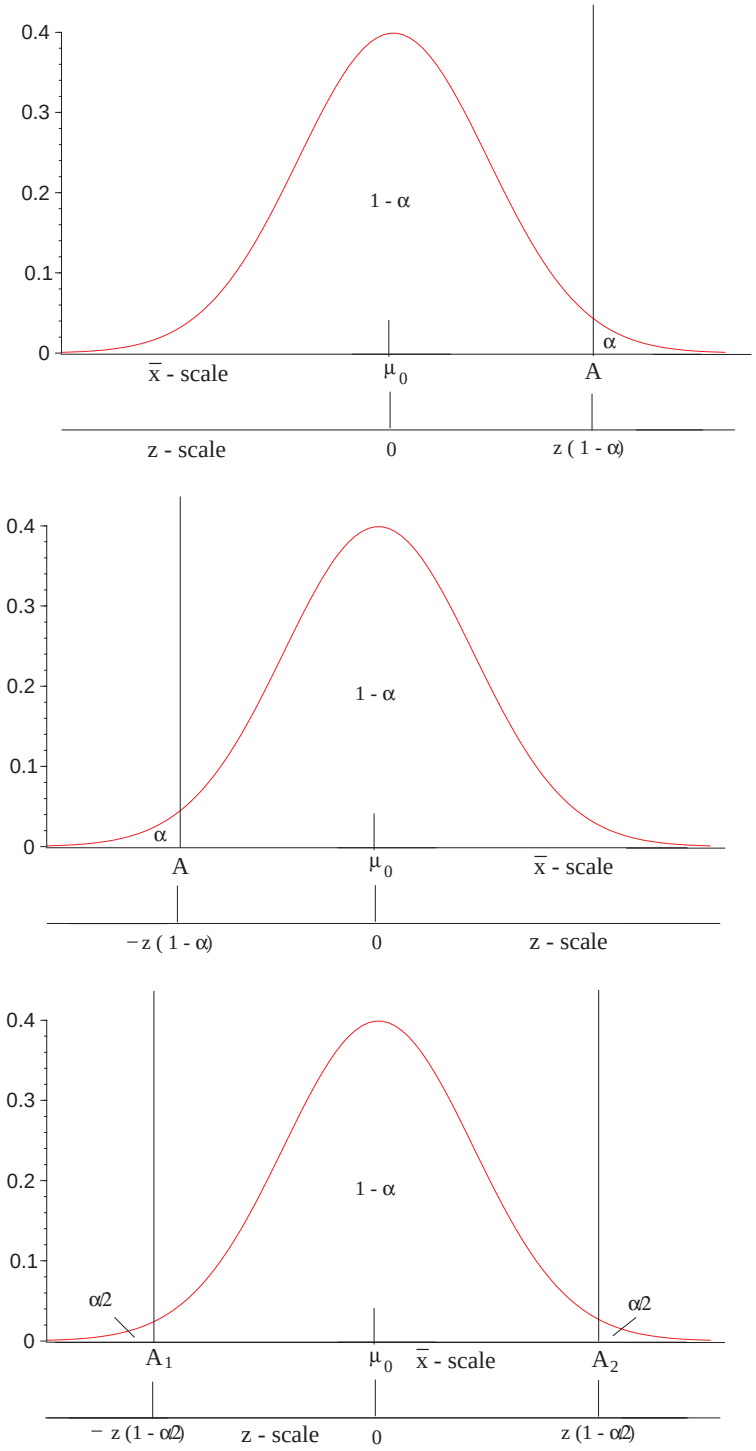


Figure 5.2: Ilustration of hypothesis tests—one-sided upper-tail (top), one-sided lower-tail (middle) and two-sided (bottom).

α-risk would have had to be selected for a borderline rejection of the null hypothesis? This probability is called the P–value. Formally, the P–value of a statistical test for μ is the probability that, if $\mu = \mu_0$, the standardised test statistic z might have been more extreme in the direction of the rejection region than was actually observed.

In the case of the children's intelligence, $\alpha/2$ would have had to be about .00275 for $t = 3.77$ to pass into the right rejection region of the t-distribution. Since the test is a two-sided one, the α-risk will be two times .00275 or .0055. The P–value is thus .0055 or somewhat more than half of one percent.

In the case of the customs department example, the value of z of roughly -13 is so far beyond the critical value of -2.28 that the α-risk required to get us to borderline reject the null hypothesis would be miniscule. Note that in this case there is only one critical region because the test is a one-tailed test, so we do not double the probability weight in that region to obtain the P–value.

The case of the no-shows on the commuter flight to New York City is more interesting because the value of z obtained from the sample is slightly less than the critical value of 1.645 when the α-risk is set at .05. The associated P–value equals

$$P(\bar{X} > 1.50|\mu = 1.32) = P(z > 1.519) = .0643.$$

There is a bit better than a 6 percent chance that we could have as many no-shows in a sample of 100 if the true mean number of no-shows on the 4 PM flight is 1.32, the mean number of no-shows on all flights.

In Figure 5.2 the P–Value would be the area to the right of our actual sample mean in the upper panel, the area to the left of our actual sample mean in the middle panel, and twice the smaller of the areas to the right or left of the actual sample mean in the lower panel.

5.5 Tests of Hypotheses about Population Proportions

When the population parameter of interest is a proportion p and the sample size is large enough to permit a normal approximation to the relevant binomial distribution, the above results go through with little modification apart from the calculation of the standard deviation of the sample proportion $\bar{p}$. It was shown in equation (4.5) of the previous chapter that the $\bar{p}$ has variance

$$Var\{\bar{p}\} = \frac{p\,(1-p)}{n},$$

and standard deviation

$$s_{\bar{p}} = \sqrt{\frac{p\,(1-p)}{n}}.$$

For example, consider a situation where the proportion of workers who are chronically ill in a particular region is known to be .11, and a random sample of 1000 workers in one of the many industries in that region yields a sample proportion of chronically ill equal to .153. We want to test whether the population of workers in that particular industry contains a higher proportion of chronically ill than the proportion of chronically ill in the entire region. Since the worst possible error would be to erroneously conclude that the proportion of chronically ill workers in the industry is bigger than the proportion in the region, we let the null hypothesis be H_0: $p \leq .11$ and the alternative hypothesis be H_1: $p > .11$. If the null hypothesis is true the standard deviation of $\bar{p}$ will equal $\sqrt{(.11)(1-.11)/1000} = .009894$. The value of the test statistic then becomes

$$z^* = \frac{\bar{p} - p}{s_{\bar{p}}} = \frac{.153 - .110}{.009894} = 4.35.$$

If we are willing to assume an α-risk of .01 in this one-sided upper-tail test the critical value of z would be 2.326. Since the sample statistic exceeds the critical value we reject the null hypothesis that the proportion of chronically ill workers in the industry is the same as or less than the proportion of chronically ill workers in the entire region.

5.6 Power of Test

Our decision rules for tests of μ have been set up to control the α-risk of the test when $\mu = \mu_0$. But we should not be indifferent about the β-risk—i.e., the risk of rejecting the alternative hypothesis when it is true. Tests that have a high risk of failing to accept the alternative hypothesis when it is true are said to have *low power*. So we now pose the question: How big is the β-risk?

Let us consider this question with in the framework of a practical problem. Suppose that the country-wide mean salary of members of a professional association is known to be \$55.5 thousand. A survey of 100 members of one of the provincial branches of the association found a mean salary in that province of $\bar{X}$ = \$62.1 thousand with s = \$24.9 thousand, yielding $s_{\bar{x}} = 24.9/10 = \$2.49$ thousand. We want to determine whether the mean salary of members in the province in question exceeds the known mean

salary of members country-wide. Let us set the α-risk at .05, controlled at $\mu_0 = 55.5$. The critical value of z is 1.645, yielding a value for A of

$$A = \mu + z(1-\alpha)s_{\bar{x}} = \mu + z(.95)(2.49) = 55.5 + (1.645)(2.49) = 59.5965.$$

The sample statistic is 62.1, well above the critical value. The standardised sample statistic is

$$z^* = \frac{62.1 - 55.5}{2.49} = \frac{6.6}{2.49} = 2.65$$

which is, of course, well above 1.645. The P–Value of the sample statistic is

$$P(\bar{X} \geq 62.1) = P(z^* \geq 2.65) = (1 - P(z^* < 2.65)) = 1 - .996 = .004.$$

While the α-risk is .05 controlled at $\mu_0 = 55.5$, the β-risk will depend on where μ actually is. Suppose that μ is actually an infinitesimal amount above 55. The null hypothesis is then false and the alternative hypothesis is true. Given our critical value A, however, there is almost a .05 probability that we will reject the null hypothesis and accept the alternative hypothesis. This means that the probability we will reject the alternative hypothesis when it is in fact true—the β-risk—is very close to .95.

Now suppose that μ is actually 57.1. The true distribution of $\bar{X}$ is then centered on $\mu = 57.1$ in the second panel from the top in Figure 5.3. About 16.1% of the distribution will now lie above the critical value A, so the probability that we will reject the null hypothesis is .16. This probability is called the *rejection probability* or the *power of test*. The probability that we will reject the alternative hypothesis is now 1 - .16 = .84. This probability—the probability of rejecting the alternative hypothesis when it is true—is the β-risk.

Suppose, instead, that μ is actually 59.6. As can be seen from the second panel from the bottom of Figure 5.3 this implies that the distribution of the test statistic is centered around the critical value A. The probability that we will reject the null hypothesis and accept the alternative hypothesis (i.e., the rejection probability or the power of test) is now .5. And the β-risk is also .5 (unity minus the rejection probability).

Finally, suppose that μ is actually 64.5. The distribution of the test statistic will now be centered around this value and, as can be seen from the bottom panel of Figure 5.3, .975 of that distribution now lies in the rejection region. The power of test is now .975 and the β-risk equals (1 - .975) = .025.

So the higher the actual value of μ the greater is the power of test and the lower is the β-risk. This can be seen from Figure 5.4. The curve in

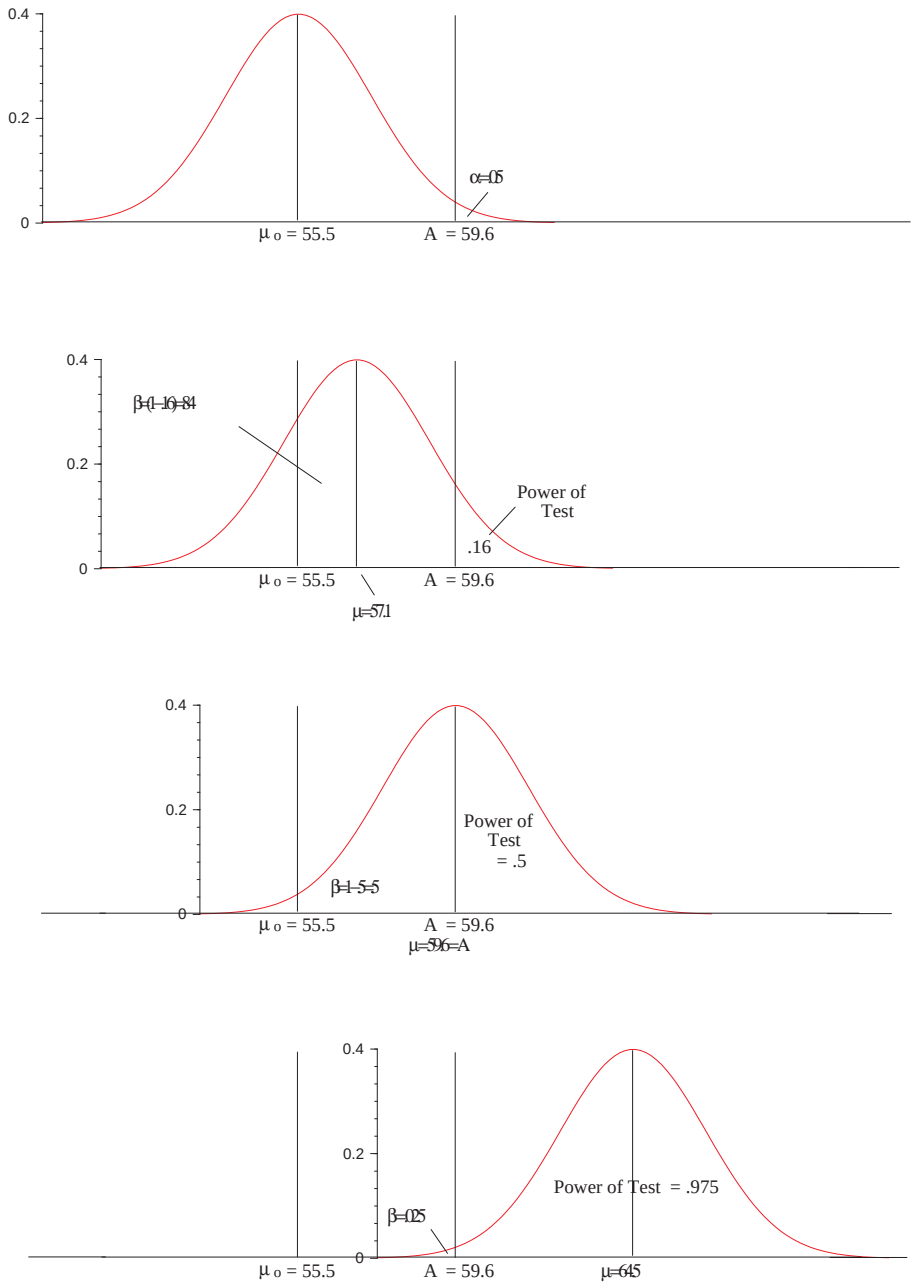


Figure 5.3: Power of test at different values of μ.

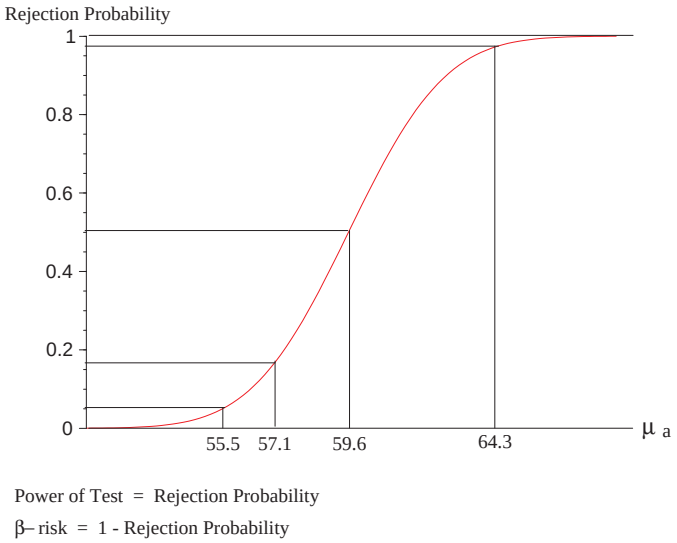


Figure 5.4: Rejection probabilities, β-risk and power of test.

that figure is called the *power curve*. The distance of that curve from the horizontal axis gives for each true value of μ the rejection probability or power of test. And the distance of the curve at each value of μ from the horizontal line at the top of the figure associated with a rejection probability of unity gives the β-risk.

The problem is, of course, that we do not know the actual value of μ (if we did, the test would be unnecessary). We thus have to choose the value of μ that we want to use to control for the β-risk. If we choose $\mu = 64.5$ as that value we can say that the power of test is .975 at μ equal to 64.5.

It can easily be seen from Figure 5.3 that the higher the value we set for the α-risk, the lower will be the β-risk at every value of μ we could set to control for the β-risk. A higher level of α will result in a critical value A closer to μ_0. The further to the left is the vertical line A, the bigger will be the power of test and the smaller will be the β-risk at every control value for μ.

The above illustration of the power of test is for one-sided upper-tail tests. For one-sided lower-tail tests the analysis is essentially the same except that A is now on the opposite side of μ_0. To portray the results graphically,

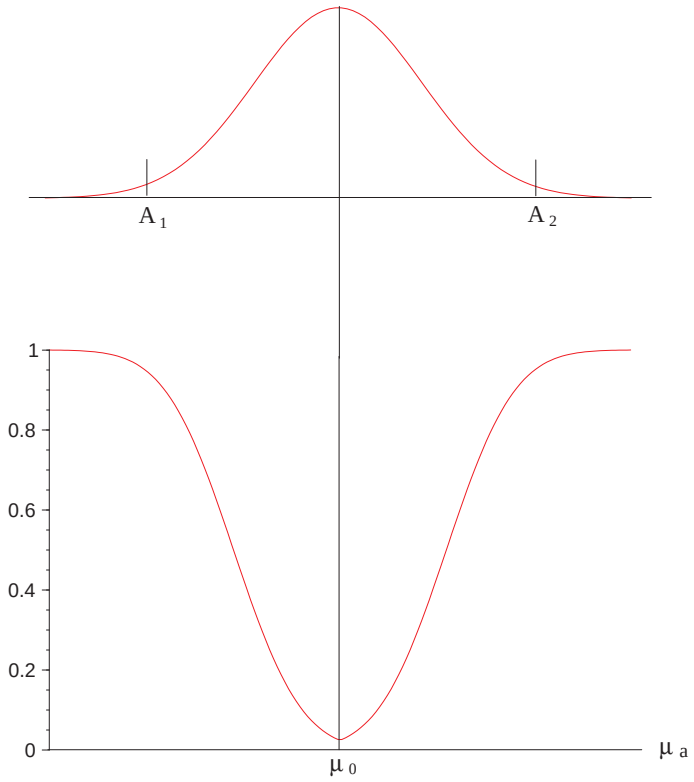


Figure 5.5: Two-sided Rejection probabilities.

simply use the mirror images of the panels in figures 5.3 and 5.4. In the case of two-sided tests the situation is a bit more complicated. The power curve now has two branches as can be seen in Figure 5.5. For any given level of μ selected to control for the β-risk the power of test will be the sum of the areas of the distribution of the sample statistic, now centered around that value of μ, to the left and right of the fixed critical levels A_1 and A_2 respectively. As the control value of μ deviates significantly from μ_0 in either direction, however, only the tail of the distribution in that direction remains relevant because the critical area on the other tail becomes miniscule. The power of test for hypotheses about population proportions is determined in exactly the same manner as above, except that the controls for the α-risk and β-risk are values of p instead of values of μ.

5.7 Planning the Sample Size to Control Both the α and β Risks

We have shown above that the lower the α-risk, the higher will be the β-risk at every level of μ at which the β-risk can be controlled. And the higher that control value of μ the greater will be the power of test.

To simultaneously control both the α-risk and the β-risk we have to choose an appropriate sample size. To choose the appropriate sample size (which must in any case be reasonably large and a small fraction of the size of the population) we must specify three things. We must specify the value μ_0 at which the α-risk is to be controlled, the value of μ, call it μ_a, at which the β-risk is to be controlled, and the planning value of σ.

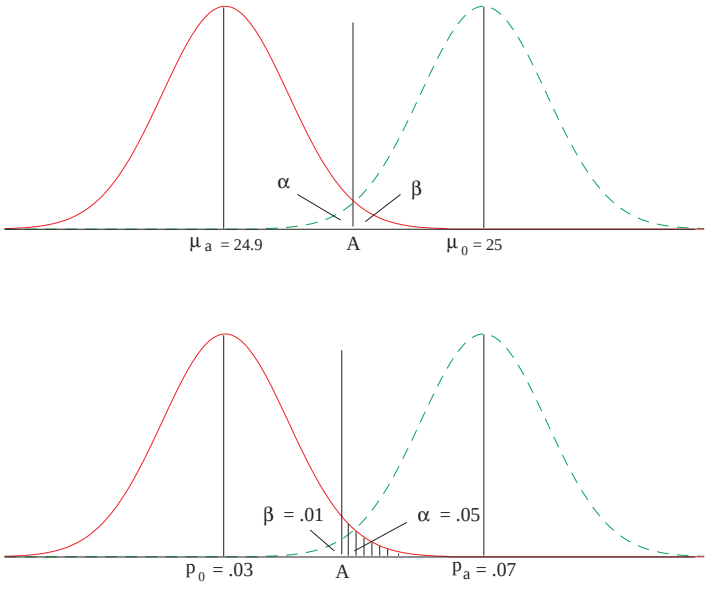


Figure 5.6: Selection of optimal sample size in butter purchase agreement problem (top) and tile shipment problem (bottom).

As a practical example, consider a purchase agreement between an aid agency and a group of producers of agricultural products. The agreement stipulates a price per 25-kilogram box of butter, but it is in the shipping

company's interest to make the boxes light. A sample is to be taken from each shipment by an independent inspection group to test whether the mean weight of butter per box is 25 kilograms. The seller does not want shipments rejected if the mean weight of butter per box is equal to or above 25 kilograms. The agreement thus specifies that null-hypothesis be H_0: $\mu \geq 25$, making the alternative hypothesis H_1: $\mu < 25$. The α-risk of rejecting a shipment when mean weight of the shipment is at least 25 kilograms is set at .05. At the same time, the company purchasing the butter is interested in the boxes not being too underweight. So the agreement also stipulates that there be no more than a five percent chance that a shipment will be accepted if it contains less than 24.9 kilograms per box of butter. This controls the β-risk of erroneously rejecting the alternative hypothesis at .05 for a value of $\mu = \mu_a = 24.9$. The problem then is to choose a sample size such that the examination process will control the α-risk at .05 when $\mu = 25$ and the β-risk at .05 when $\mu = 24.9$. The buyer and seller agree to adopt a planning value of σ equal to .2. The analysis can be illustrated with reference to the upper panel of Figure 5.6. Let the as yet to be determined critical value for rejection of a shipment be A. The standardised difference between μ_0 and A must equal

$$z_0 = \frac{\mu_0 - A}{\sigma/\sqrt{n}} = \frac{25 - A}{.2/\sqrt{n}} = 1.645$$

and the standardised difference between A and μ_a must equal

$$z_1 = \frac{A - \mu_a}{\sigma/\sqrt{n}} = \frac{A - 24.9}{.2/\sqrt{n}} = 1.645.$$

These expressions can be rewritten as

$$25 - A = (1.645)(.2/\sqrt{n})$$

and

$$A - 24.9 = (1.645)(.2/\sqrt{n}).$$

Adding them together yields

$$25 - 24.9 = .1 = (2)(1.645)(.2/\sqrt{n}) = (3.29)(.2)/\sqrt{n})$$

which implies that

$$n = (\sqrt{n})^2 = \left(\frac{(.2)(3.29)}{.1}\right)^2 = 43.3.$$

A sample size of 44 will do the trick. The critical value A will equal

$$25 - 1.645\frac{.2}{\sqrt{44}} = 25 - (1.645)(.0301) = 25 - .05 = 24.95.$$

Consider another example where a purchaser of a large shipment of tiles wishes to control the β-risk of accepting the shipment at .01 when the proportion of tiles that are damaged is $p = .07$ while the vendor wishes to control the α-risk of having the shipment rejected at .025 when the proportion of damaged tiles is .03. A random sample of tiles will be selected from the shipment by the purchaser on the basis of which a decision will be made to accept (H_0: $p \leq .03$) or reject (H_1: $p > .03$) the shipment. We need to find the sample size sufficient to meet the requirements of both the purchaser and vendor. The analysis can be conducted with reference to the bottom panel of Figure 5.6. The standardised distance between the as yet to be determined critical value A and $p = .03$ must be

$$z_0 = \frac{A - .03}{\sigma_{\bar{p}_0}} = \frac{A - .03}{\sqrt{(.03)(.97)/n}} = 1.96$$

and the standardised difference between .07 and A must be

$$z_1 = \frac{.07 - A}{\sigma_{\bar{p}_1}} = \frac{.07 - A}{\sqrt{(.07)(.93)/n}} = 2.326.$$

Note that we use the values of p at which the α-risk and β-risk are being controlled to obtain the relevant values of $\sigma_{\bar{p}}$ for standardizing their differences from the critical value A. Multiplying both of the above equations by $\sqrt{n}$ and then adding them, we obtain

$$(.04)\sqrt{n} = (1.96)\sqrt{(.03)(.97)} + (2.326)\sqrt{(.07)(.93)}$$

which yields

$$n = \left(\frac{(1.96)\sqrt{(.03)(.97)} + (2.326)\sqrt{(.07)(.93)}}{.04}\right)^2 = 538.$$

The critical value A will then equal

$$A = .03 + z_0\,\sigma_{\bar{p}_0} = .03 + 1.96\sqrt{(.03)(.97)/538} = .0444.$$

5.8 Exercises

1. It is desired to test H_0: $\mu \geq 50$ against H_1: $\mu < 50$ with a significance level $\alpha = .05$. The population in question is normally distributed with known standard deviation $\sigma = 12$. A random sample of $n = 16$ is drawn from the population.

a) Describe the sampling distribution of $\bar{X}$, given that $\mu = 50$.

b) If μ is actually equal to 47, what is the probability that the hypothesis test will lead to a Type II error. (.74059)

c) What is the power of this test for detecting the alternative hypothesis H_a: $\mu = 44$? (.5213)

2. A sales analyst in a firm producing auto parts laboriously determined, from a study of all sales invoices for the previous fiscal year, that the mean profit contribution per invoice was \$16.50. For the current fiscal year, the analyst selected a random sample of 25 sales invoices to test whether the mean profit contribution this year had changed from \$16.50 (H_1) or not (H_0). The sample of 25 invoices yielded the following results for the invoice profit contributions: $\bar{X} = \$17.14$, $s = \$18.80$. The α risk is to be controlled at 0.05 when $\mu = 16.50$.

a) Conduct the test. State the alternatives, the decision rule, the value of the standardised test statistic, and the conclusion.

b) What constitute Type I and Type II errors here? Given the conclusion above, is it possible that a Type I error has been made in this test? Is a Type II error possible here? Explain.

3. In a tasting session, a random sample of 100 subjects from a target consumer population tasted a food item, and each subject individually gave it a rating from 1 (very poor) to 10 (very good). It is desired to test H_0: $\mu \leq 6.0$ vs. H_1: $\mu > 6.0$, where μ denotes the mean rating for the food item in the target population. A computer analysis of the sample results showed that the one-sided P–value of the test is .0068.

a) Does the sample mean lie above or below $\mu_0 = 6.0$?

b) What must be the value of value of z generated by the sample? (2.47)

c) The sample standard deviation is $s = 2.16$. What must be the sample mean $\bar{X}$? (6.5332)

d) Does the magnitude of the P–value indicate that the sample results are inconsistent with conclusion H_0? Explain.

4. The developer of a decision-support software package wishes to test whether users consider a colour graphics enhancement to be beneficial, on balance, given its list price of $800. A random sample of 100 users of the package will be invited to try out the enhancement and rate it on a scale ranging from -5 (completely useless) to 5 (very beneficial). The test alternatives are H_0: $\mu \leq 0$, H_1: $\mu > 0$, where μ denotes the mean rating of users. The α risk of the test is to be controlled at 0.01 when $\mu = 0$. The standard deviation of users' ratings is $\sigma = 1.3$.

a) Show the decision rule for $\bar{X}$ relevant for this test.

b) Calculate the rejection probabilities at $\mu = 0, 0.5, 1.0$ and 1.5 for the decision rule above.

c) Sketch the rejection probability curve for the decision rule you selected above.

d) What is the incorrect conclusion when $\mu = 0.60$? What is the probability that the above decision rule will lead to the incorrect conclusion when $\mu = .60$? Is the probability an α or β risk?

5. "Take the Pepsi Challenge" was a marketing campaign used by the Pepsi-Cola Company. Coca Cola drinkers participated in a blind taste test where they were asked to taste unmarked cups of Pepsi and Coke and were asked to select their favourite. In one Pepsi television commercial the announcer states that "in recent blind taste tests more than half the Diet Coke drinkers surveyed said they preferred the taste of Pepsi." (*Consumer's Research*, May 1993). Suppose that 100 Coke drinkers took the Pepsi challenge and 56 preferred the taste of Diet Pepsi. Does this indicate that more than half of all Coke drinkers prefer the taste of Pepsi?

6. A salary survey conducted on behalf of the Institute of Management Accountants and the publication *Management Accounting* revealed that the average salary for all members of the Institute was $56,391. A random

sample of 122 members from New York were questioned and found to have a mean salary of \$62,770 and a standard deviation of $s = \$28972$ (*Management Accounting*, June 1995).

a) Assume that the national mean is known with certainty. Do the sample data provide sufficient evidence to conclude that the true mean salary of Institute members in New York is higher than the National Average?

b) Suppose the true mean salary for all New York members is \$66,391. What is the power of your test above to detect this \$10,000 difference?

7. One of the most pressing problems in high-technology industries is computer-security. Computer security is typically achieved by a *password*—a collection of symbols (usually letters and numbers) that must be supplied by the user before the computer system permits access to the account. The problem is that persistent hackers can create programs that enter millions of combinations of symbols into a target system until the correct password is found. The newest systems solve this problem by requiring authorized users to identify themselves by unique body characteristics. For example, system developed by Palmguard, Inc. tests the hypothesis

H_0: The proposed user is authorized

versus

H_1: The proposed user is unauthorized.

by checking characteristics of the proposed user's palm against those stored in the authorized users' data bank (*Omni, 1984*).

a) Define a Type I error and a Type II error for this test. Which is the more serious error? Why?

b) Palmguard reports that the Type I error rate for its system is less than 1% where as the Type II error rate is .00025%. Interpret these error rates.

c) Another successful security system, the EyeDentifyer, "spots authorized computer users by reading the one-of-a-kind patterns formed by the network of minute blood vessels across the retina at the back of the eye." The EyeDentifier reports Type I and Type II error rates of .01% (1 in 10,000) and .005% (5 in 100,000), respectively. Interpret these rates.

8. Under what circumstances should one use the t-distribution in testing an hypothesis about a population mean? For each of the following rejection regions, sketch the sampling distribution of t, and indicate the location of the rejection region on your sketch:

a) $t > 1.440$ where $v = 6$.

b) $t < -1.782$ where $v = 12$.

c) $t < -2.060$ or $t > 2.060$ where $v = 25$.

9. Periodic assessment of stress in paved highways is important to maintaining safe roads. The Mississippi Department of Transportation recently collected data on number of cracks (called *crack intensity*) in an undivided two-lane highway using van-mounted state-of-the-art video technology (*Journal of Infrastructure Systems*, March 1995). The mean number of cracks found in a sample of eight 5-meter sections of the highway was $\bar{X} = .210$, with a variance of $s^2 = .011$. Suppose that the American Association of State Highway and Transportation Officials (AASHTO) recommends a maximum mean crack intensity of .100 for safety purposes. Test the hypothesis that the true mean crack intensity of the Mississippi highway exceeds the AASHTO recommended maximum. Use $\alpha = .01$.

10. Organochlorine pesticides (OCP's) and polychlorinated biphenyls, the familiar PCB's, are highly toxic organic compounds that are often found in fish. By law, the levels of OCP's and PCB's in fish are constantly monitored, so it is important to be able to accurately measure the amounts of these compounds in fish specimens. A new technique called matrix solid-phase dispersion (MSPD) has been developed for chemically extracting trace organic compounds from solids (*Chromatographia*, March 1995). The MSPD method was tested as follows. Uncontaminated fish fillets were injected with a known amount of OCP or PCB. The MSPD method was then used to extract the contaminant and the percentage of the toxic compound uncovered was measured. The recovery percentages for $n = 5$ fish fillets injected with the OCP Aldrin are listed below:

99 102 94 99 95

Do the data provide sufficient evidence to indicate that the mean recovery percentage of Aldrin exceeds 85% using the new MSPD method? Set the α-risk at .05.

Chapter 6

Inferences Based on Two Samples

Frequently we want to use statistical techniques to compare two populations. For example, one might wish to compare the proportions of families with incomes below the poverty line in two regions of the country. Or we might want to determine whether electrical consumption in a community has increased during the past decade.

6.1 Comparison of Two Population Means

Take two populations with means μ_1 and μ_2. The central limit theorem tells us that sample means from these populations will be approximately normally distributed for large samples.

Suppose we select independent random samples of n_1 and n_2, both reasonably large, from the respective populations. We want to make inferences about the difference $\mu_2 - \mu_1$ on the basis of the two samples.

From the statistical theory developed in Chapter 3 (section 3.6) we know that

$$E\{\bar{Y} - \bar{X}\} = E\{\bar{Y}\} - E\{\bar{X}\} = \mu_2 - \mu_1$$

and, since the samples are independent,

$$\sigma^2\{\bar{Y} - \bar{X}\} = \sigma^2\{\bar{Y}\} + \sigma^2\{\bar{X}\}.$$

And it is natural to use

$$s^2\{\bar{Y} - \bar{X}\} = s^2\{\bar{Y}\} + s^2\{\bar{X}\}$$

as an unbiased point estimator of $\sigma^2\{\bar{Y} - \bar{X}\}$.

We can proceed in the usual fashion to construct confidence intervals and statistical tests. Suppose, for example, that a random sample of 200 households from a large community was selected to estimate the mean electricity use per household during February of last year and another simple random sample of 250 households was selected, independently of the first, to estimate mean electricity use during February of this year. The sample results, expressed in kilowatt hours, were

Last Year	This Year
$n_1 = 200$	$n_2 = 250$
$\bar{X} = 1252$	$\bar{Y} = 1320$
$s_1 = 157$	$s_2 = 151$

We want to construct a 99 percent confidence interval for $\mu_2 - \mu_1$.

An unbiased point estimate of $\mu_2 - \mu_1$ is

$$\bar{Y} - \bar{X} = 1320 - 1252 = 68.$$

The standard error of the difference between the sample means is

$$s\{\bar{Y} - \bar{X}\} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{157^2}{200} + \frac{151^2}{250}}$$

$$= \sqrt{123.45 + 91.20} = 14.64.$$

The 99 percent confidence interval will thus be

$$68 \pm z(1 - .01/2)(14.64) = 68 \pm (2.576)(14.64)$$

or

$$30.29 \leq \mu_2 - \mu_1 \leq 105.71.$$

The fact that the above confidence interval does not include zero makes it evident that a statistical test of the null hypothesis that $\mu_2 - \mu_1 \leq 0$ is likely to result in rejection of that null. To test whether the mean household use of electricity increased from February of last year to February of this year, controlling the α-risk at .005 when $\mu_2 = \mu_1$, we set

$$H_0\colon\ \mu_2 - \mu_1 \leq 0$$

and

$$H_1\colon\ \mu_2 - \mu_1 > 0.$$

The critical value of z is $z(.995) = 2.576$. From the sample,

$$z^* = \frac{68}{14.64} = 4.645,$$

which is substantially above the critical value. The P-value is

$$P(z > 4.645) = 0000017004.$$

We conclude that per-household electricity consumption has increased over the year.

6.2 Small Samples: Normal Populations With the Same Variance

The above approach is appropriate only for large samples. In some cases where samples are small (and also where they are large) it is reasonable to assume that the two populations are normally distributed with the same variance. In this case

$$E\{\bar{Y} - \bar{X}\} = E\{\bar{Y}\} - E\{\bar{X}\} = \mu_2 - \mu_1$$

as before but now

$$\sigma^2\{\bar{Y} - \bar{X}\} = \sigma^2\{\bar{Y}\} + \sigma^2\{\bar{X}\}.$$

$$= \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right].$$

To calculate confidence intervals we need an estimator for σ_2. It turns out that the *pooled* or *combined estimator*

$$\begin{aligned} s_c^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \end{aligned} \tag{6.1}$$

is an unbiased estimator of σ^2. We can thus use

$$s^2\{\bar{Y} - \bar{X}\} = s_c^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

as an unbiased estimator of $\sigma^2\{\bar{Y} - \bar{X}\}$.

We proceed as usual in setting the confidence intervals except that, given the small samples, the test statistic

$$\frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{s\{\bar{Y}-\bar{X}\}}$$

is distributed as $t(n_1+n_2-2)$—that is, as a t-distribution with $v = n_1+n_2-2$ degrees of freedom.

By making the assumptions of normality and equal variance we can use small samples whereas in the general case of the previous section the sample sizes had to be large enough to justify approximate normality according to the Central Limit Theorem.

Now consider an example. Suppose we wish to estimate the difference in mean tread life for a certain make of automobile tire when it is inflated to the standard pressure as compared to a higher-than-standard pressure to improve gas mileage. Two independent random samples of 15 tires were selected from the production line. The tires in sample 1 were inflated to the standard pressure and the tires in sample 2 were inflated to the higher pressure. Tread-life tests were conducted for all tires with the following results, expressed in thousands of miles of tread life.

Standard Pressure	Higher Pressure
$n_1 = 14$	$n_2 = 15$
$\bar{X} = 43$	$\bar{Y} = 40.7$
$s_1 = 1.1$	$s_2 = 1.3$

Because one tire in sample 1 turned out to be defective it was dropped from that sample, reducing the sample size to 14.

Note that the respective populations here are the infinite populations of tread lives of non-defective tires of the make tested when inflated to the standard and higher pressures respectively. We suppose that on the basis of other evidence it is reasonable to assume that both populations are normal with the same variance.

So we have

$$\bar{Y}-\bar{X} = 40.7 - 43.0 = -2.3$$

as an unbiased point estimate of $\mu_2-\mu_1$. In addition, we have

$$s_c^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{(13)(1.1)^2+(14)(1.3)^2}{14+15-2} = 1.45899,$$

so that

$$s^2\{\bar{Y}-\bar{X}\} = 1.45899\left[\frac{1}{14}+\frac{1}{15}\right] = (1.45899)(.0714+.0667) = .2015,$$

which implies that $s\{\bar{Y} - \bar{X}\} = .4488$. The 95 percent confidence interval is thus

$$-2.3 \pm t(1 - .05/2; 14 + 15 - 2)(.4488) = -2.3 \pm t(.975; 27)(.4488)$$

$$= -2.3 \pm (2.052)(.4488) = -2.3 \pm .9245.$$

Hence,

$$-3.2245 \leq \mu_2 - \mu_1 \leq -1.3755.$$

The mean life of non-defective tires inflated to the higher pressure is between 1.38 and 3.22 thousand miles less than that of non-defective tires inflated to the standard pressure, with 95 percent confidence.

The result of a test of the null hypothesis that $\mu_2 - \mu_1 > 0$ is obvious from the confidence interval above if we control the α-risk at .025 when $\mu_2 = \mu_1$. The critical value for t is -2.060 while

$$t^* = \frac{-2.3}{.4488} = -5.125.$$

The P-value is

$$P(t(27) < -5.125) = 0.00000050119.$$

We conclude that the tread life of tires inflated to the higher pressure is less than that for tires inflated to the standard pressure.

6.3 Paired Difference Experiments

Suppose that we want to find the weight loss in a shipment of bananas during transit. The procedures used above would suggest that we select and weigh a random sample of banana bunches before loading and then select and weigh another independent random sample of banana bunches after shipment and unloading. The differences in the mean weights before and after could then be used to make inferences about the weight loss during shipment. But there is a better way of handling this problem.

The better way would be to select and weigh a random sample of banana bunches before loading and then weigh the same bunch again after shipment and unloading. We could use the mean difference between the 'before' and 'after' weights of the sample of banana bunches to make inferences about the weight loss during shipping. It is important here that the sample of banana bunches be treated in the same way during transit as the rest of the shipment. To ensure that this is the case we would have to mark the

selected bunches of bananas in a way that would identify them to us after shipment but not to the people handling the shipping process. The shipping company would therefore not be able to cover up weaknesses in its handling of the shipment by giving the sample of banana bunches special treatment. In this case we are using *matched samples* and making the inference on the basis of *paired differences*.

By using paired differences we can take advantage of the fact that the 'before' and 'after' means are positively correlated—banana bunches which were heavier than average before shipment will also tend to be heavier than average after shipment. The covariance between the 'before' and 'after' weights is therefore positive so that the variance of the difference between the 'before' and 'after' mean weights will be less than the variance of the difference between the mean weights of independently selected random samples before and after shipment. That is,

$$\sigma^2\{\bar{Y} - \bar{X}\} = \sigma^2\{\bar{Y}\} + \sigma^2\{\bar{X}\} - 2\sigma\{\bar{Y}\bar{X}\} < \sigma^2\{\bar{Y}\} + \sigma^2\{\bar{X}\}.$$

It is thus more efficient to work directly with the paired differences in weights than with differences of mean weights. Indeed, if we select matched samples it is inappropriate to use the procedures of the previous sections because the matched samples are not independent of each other as required by those procedures.

So we can set

$$D_i = Y_i - X_i$$

where Y_i is the weight of the ith bunch before shipment and X_i is the weight of that same bunch after shipment. We can then calculate

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

and

$$s_D^2 = \sum_{i=1}^{n} \frac{(D_i - \bar{D})^2}{n-1}$$

from whence

$$s_{\bar{D}} = \sqrt{\frac{s_D^2}{n}}.$$

Consider another example. Suppose that a municipality requires that each residential property seized for non-payment of taxes be appraised independently by two licensed appraisers before it is sold. In the past 24

months, appraisers Smith and Jones independently appraised 50 such properties. The difference in appraised values $D_i = Y_i - X_i$ was calculated for each sample property, where X_i and Y_i denote Smith's and Jones' respective appraised values. The mean and standard deviation of the 50 differences were (in thousands of dollars)

$$\bar{D} = 1.21$$

and

$$s_D = 2.61$$

respectively. It thus follows that

$$s_{\bar{D}} = \frac{2.61}{\sqrt{50}} = \frac{2.61}{7.07} = .3692.$$

The 95 percent confidence interval for the mean difference in appraised values for these two appraisers is

$$1.21 \pm z(.975)(.3692) = 1.21 \pm (1.96)(.3692) = 1.21 \pm .724$$

which implies

$$.486 \leq \mu_D \leq 1.934.$$

The confidence interval applies to the hypothetical population of differences in appraised values given independently by Jones and Smith to properties of a type represented by those in the sample, namely, properties seized for non-payment of taxes.

Suppose that an observer who has not seen the sample suspects that Jones' appraised values tend to be higher on average than Smith's. To test whether this suspicion is true, setting the α-risk at .025 when $\mu_D = 0$, we set the null hypothesis

$$H_0 \colon \mu_D \leq 0$$

and the alternative hypothesis

$$H_1 \colon \mu_D > 0.$$

The critical value of z is 1.96. The value of z given by the sample is

$$z^* = \frac{1.21}{.3692} = 3.277.$$

We conclude that Jones' appraised values are on average higher than Smith's. The result of this hypothesis test is obvious from the fact that the confidence interval calculated above did not embrace zero.

Note that in the above example we used the normal approximation because the sample size of 50 was quite large. Had the sample size been small, say 8, we would have used the t-distribution, setting the critical value and confidence limits according to $t(.975; 7)$.

6.4 Comparison of Two Population Proportions

Inferences about two population proportions based on large samples can be made in straight-forward fashion using the relationships

$$E\{\bar{p}_2 - \bar{p}_1\} = p_2 - p_1$$

and

$$\sigma^2\{\bar{p}_2 - \bar{p}_1\} = \sigma^2\{\bar{p}_2\} + \sigma^2\{\bar{p}_1\}$$

and approximating the latter using

$$s^2\{\bar{p}_2 - \bar{p}_1\} = s^2\{\bar{p}_2\} + s^2\{\bar{p}_1\},$$

where

$$s^2\{\bar{p}_i\} = \frac{\bar{p}_i(1 - \bar{p}_i)}{n_i - 1}.$$

We use $(n_i - 1)$ in the denominator of the above expression for the same reason that $(n - 1)$ appears in the denominator of

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n - 1}.$$

Now consider an example. A manufacturer of consumer products obtains data on breakdowns of two makes of microwave ovens. In a sample of $n_1 = 197$ ovens of make 1 it is found that 53 broke down within 5 years of manufacture, whereas in a sample of $n_2 = 290$ ovens of make 2, only 38 ovens broke down within 5 years of manufacture. Assume that the samples are independent random samples from their respective populations. We want a 99 percent confidence interval for $p_2 - p_1$. We have

$$\bar{p}_2 - \bar{p}_1 = \frac{38}{290} - \frac{53}{197} = .13103 - .26904 = -.1380$$

$$s^2\{\bar{p}_1\} = \frac{(.26904)(.73096)}{196} = .00100335$$

$$s^2\{\bar{p}_2\} = \frac{(.13103)(.86897)}{289} = .00039439$$

$$s\{\bar{p}_2 - \bar{p}_1\} = \sqrt{.00100335 + .00039439} = .0374.$$

The 99 percent confidence interval is

$$-.1380 \pm z(.995)(.0374) = -.1380 \pm (2.576)(.0374) = -.1380 \pm .096$$

or

$$-.234 \leq p_2 - p_1 \leq -.042.$$

The percentage of units of make 1 that break down within 5 years of manufacture is between 4.2 and 23.4 percentage points more than that of make 2, with 99 percent confidence.

Now we want to test whether the proportion breaking down within one year for make 1 is larger than the proportion for make 2, controlling the α-risk at .005 when $p_2 = p_1$. We set

$$H_0 \colon\ p_2 - p_1 \geq 0$$

and

$$H_1 \colon\ p_2 - p_1 < 0.$$

The critical value of z is -2.576. To calculate z^* we need an estimate of $\sigma\{\bar{p}_2 - \bar{p}_1\}$ when $p_2 = p_1 = p$. The appropriate procedure is to use a *pooled estimator* of p to calculate an estimate of $\bar{p}$. We simply take a weighted average of $\bar{p}_1$ and $\bar{p}_2$ using the sample sizes as weights:

$$\bar{p}\,' = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}.$$

We thus have

$$\bar{p}\,' = \frac{(197)(.26904) + (290)(.13103)}{197 + 290} = .185.$$

An appropriate estimator of $\sigma^2\{\bar{p}_2 - \bar{p}_1\}$ is thus

$$s^2\{\bar{p}_2 - \bar{p}_1\} = \bar{p}\,'(1 - \bar{p}\,')\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

which yields

$$s\{\bar{p}_2 - \bar{p}_1\} = \sqrt{(.185)(.815)\left[\frac{1}{197} + \frac{1}{290}\right]} = .0378.$$

The resulting value of z^* is thus

$$z^* = \frac{-.1380 - 0}{.0378} = -3.65.$$

We conclude that the proportion of microwave ovens of make 1 breaking down within 5 years of manufacture is greater than the proportion of microwave ovens of make 2 breaking down within 5 years of manufacture. The P-value is

$$P(z < -3.65) = .00013112.$$

6.5 Exercises

1. Two random samples are independently drawn from two populations. A two-tailed test is used to evaluate H_0: $\mu_x = \mu_y$.

	X	Y
Sample size (n)	3	5
Mean	7.0	3.0
Variance	1.0	2.5

Find the lowest value of α at which the researcher will reject the null hypothesis. (.015) What assumptions did the researcher have to make about the populations to do this test?

2. The following describe the results of independent samples drawn from different populations.

Sample 1	Sample 2
$n_1 = 159$	$n_2 = 138$
$\bar{X}_1 = 7.4$	$\bar{X}_2 = 9.3$
$s_1 = 6.3$	$s_2 = 7.1$

a) Conduct a test of the hypothesis H_0: $\mu_1 - \mu_2 \geq 0$ against the alternative H_1: $\mu_1 - \mu_2 < 0$ with a significance level $\alpha = 0.10$.

b) Determine the P-value for the test statistic of a) above.

3. A pharmaceutical company wishes to test whether a new drug that it is developing is an effective treatment for acne (a facial skin disorder that is particularly prevalent among teenagers). The company randomly selects 100 teenagers who are suffering from acne and randomly divides them into two groups of 50 each. Members of Group 1 receive the drug each day while members of Group 2 receive no medication. At the end of three months,

members of both groups are examined and it is found that 27 of the teenagers in Group 1 no longer have acne as compared with 19 of the teenagers in Group 2 who no longer have acne. Using a significance level of $\alpha = 0.01$, set up and conduct a test of whether the drug is effective or not. Determine the P-value for your test statistic. (.40675)

4. A public opinion research institute took independent samples of 500 males and 500 females in a particular U.S. state, asking whether the respondents favoured a particular constitutional amendment. It was found that 335 of the males and 384 of the females were in favour of the amendment. Construct a 90% confidence interval for difference between the proportions of males and females favouring the amendment and test the hypothesis that the proportions are the same.

5. A manufacturer of automobile shock absorbers was interested in comparing the durability of his shocks with that of the shocks of his biggest competitor. To make the comparison, one of the manufacturer's and one of the competitor's shocks were randomly selected and installed on the rear wheels of each of six cars. After the cars had been driven 20,000 miles, the strength of each test shock was measured, coded and recorded. The results were as follows

Car	Manufacturer's Shock	Competitor's Shock
1	8.8	8.4
2	10.5	10.1
3	12.5	12.0
4	9.7	9.3
5	9.6	9.0
6	13.2	13.0

a) Do the data present sufficient evidence to conclude that there is a difference in the mean strength of the two types of shocks after 20,000 miles of use?

b) Find the approximate observed significance level for the test and interpret its value?

c) What assumptions did you make in reaching these conclusions.

6. A sociologist is researching male attitudes toward women. For her study, random samples of male students from the City of Toronto are interviewed and their results are tabulated. Sample One was conducted in 1988 and consisted of $n_1 = 100$ boys aged 6 to 8. From this group, $x_1 = 90$ subjects indicated in their responses that "girls are ugly, girls have cooties, girls eat worms and that all girls should just go away." The researcher concluded that a large proportion of young boys just don't like girls. A second sample conducted in 1998 consisting of $n_2 = 225$ boys also aged 6 to 8. From this group $x_2 = 180$ subjects exhibited beliefs similar to those 90 boys in the first sample. Using both samples, develop an hypothesis test to evaluate whether the proportion of boys who don't like girls has changed significantly over the 10 year period. When required, manage the α-risk at 5%. Provide a P-value for the test. What does it say regarding attitudes?

7. You know from earlier studies that about 7% of all persons are left-handed. You suspect that left-handedness is more prevalent among men than among women and wish to use independent random samples to measure the difference between the proportions of men and women who are left-handed. You would like an 80% confidence interval for this difference to be accurate within ± 0.01.

a) How many persons should be included in your sample? (91)

b) Will the sample size determined in a) above be large enough to permit a 95% confidence interval for the proportion of men who are left-handed to be accurate to within ± 0.01?

8. In an economics class of 100 students the term mark, T, for each student is compared with his/her marks on two term texts, X_1 and X_2, with $T = X_1 + X_2$. The summary statistics for the entire class were:

	Mean Mark	Standard Deviation
First Term Test	32.0	8.0
Second Term Test	36.0	6.0
Term Mark	68.0	12.0

a) Determine the values for the covariance and the correlation between X_1 and X_2.

b) Calculate the mean and standard deviation of the paired difference in the marks between the first and second term tests.

c) Conduct a test as to whether students performed better on the first term test than the second.

Chapter 7

Inferences About Population Variances and Tests of Goodness of Fit and Independence

In the last chapter we made inferences about whether two population means or proportions differed based on samples from those populations. Integral in all those tests and in the inferences in the previous chapters about population means and population proportions was our use of the statistic

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \tag{7.1}$$

as an unbiased point estimate of the population variance σ^2. A natural next step is to make inferences—set confidence intervals and test hypotheses—about σ^2 on the basis of the sample statistic s^2.

7.1 Inferences About a Population Variance

To proceed we must know the sampling distribution of s^2. This involves the *chi-square (χ^2) distribution*, the basis of which must now be explained. Suppose we have a set of independent random draws from a variable

$$X_1, X_2, X_3, \ldots\ldots$$

which is normally distributed with population mean μ and variance σ^2. Consider this sequence in standardised form

$$Z_1, Z_2, Z_3, \ldots\ldots$$

where, of course,

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

Now square the Z_i to obtain

$$Z_i^2 = \frac{(X_i - \mu)^2}{\sigma^2}.$$

It turns out that the sum of n of these squared standardised independent normal variates,

$$Z_1^2 + Z_2^2 + Z_3^3 + \ldots + Z_n^2,$$

is distributed as a chi-square distribution. We can thus write

$$\sum_{i=1}^{n} Z_i^2 \quad = \quad \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \quad = \quad \chi^2(n) \tag{7.2}$$

where $\chi^2(n)$ is a chi-square random variable—that is, a random variable distributed according to the chi-square distribution—with parameter n, which equals the number of independent normal variates summed. This parameter is typically referred to as the degrees of freedom. Notice now that

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

differs from the expression above in that $\bar{X}$ replaces μ in the numerator. This expression is also distributed as χ^2—indeed

$$\sum_{i=1}^{n} Z_i^2 \quad = \quad \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \quad = \quad \chi^2(n-1) \tag{7.3}$$

where the parameter, the degrees of freedom, is now $n - 1$.

At this point is worth while to pay further attention to what we mean by *degrees of freedom*. The degrees of freedom is the number of independent pieces of information used in calculating a statistic. In the expression immediately above, the n deviations of the X_i from their sample mean contain only $n - 1$ independent pieces of information. The sample mean is constructed from the n sample values of X_i by summing the X_i and dividing by

n. Accordingly, the sum of the deviations around this mean must be zero. Hence, if we know any $n-1$ of the n deviations around the mean we can calculate the remaining deviation as simply the negative of the sum of the $n-1$ deviations. Hence, only $n-1$ of the deviations are freely determined in the sample. This is the basis of the term 'degrees of freedom'. Even though there are n deviations, only $n-1$ of them produce independent sum of squared deviations from the sample mean. This is in contrast to the sum of squared deviations about the *true* mean μ, which contains n independent pieces of information because μ is independent of all the sample observations. Information is not used up in calculating the population mean as it is in calculating $\bar{X}$. This is why the standardised sum of squared deviations of the sample values about the true mean is distributed as $\chi^2(n)$ whereas the sum of squared deviations of the sample values from the sample mean, standardised by the true variance σ^2, is distributed as $\chi^2(n-1)$.

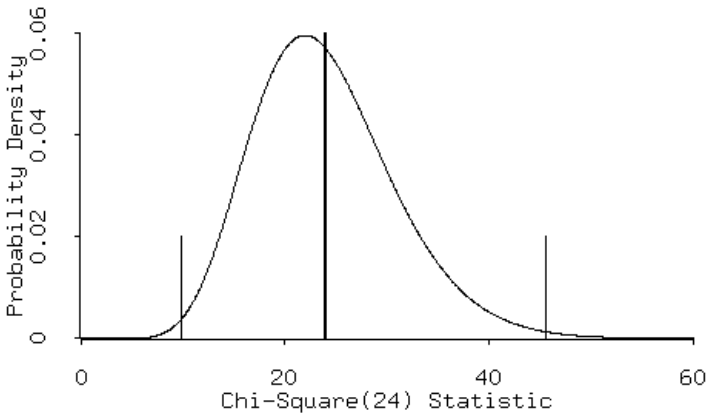


Figure 7.1: A chi-square distribution with 24 degrees of freedom. The thick vertical line shows the mean and the thin vertical lines the critical values for $\alpha = .99$.

Notice now that the expression for s^2, given in equation (7.1) above, can be rewritten

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad = \quad (n-1)\, s^2. \tag{7.4}$$

Substituting this into (7.3), we obtain

$$\frac{(n-1)\,s^2}{\sigma^2} \quad = \quad \chi^2(n-1). \tag{7.5}$$

The sampling distribution for this statistic is skewed to the right, with the skew being smaller the greater the degrees of freedom. Figure 7.1 shows a χ^2 distribution with 24 degrees of freedom. The thick vertical line gives the mean and the thin vertical lines the critical values for $\alpha = .99$. The mean of the χ^2 distribution is the number of degrees of freedom, usually denoted by v which in the above examples equals either n or $n-1$ or in Figure 7.1, 24. Its variance is $2\,v$ or twice the number of degrees of freedom. The percentiles of the χ^2 distribution (i.e., the fractions of the probability weight below given values of χ^2) for the family of chi-square distributions can be obtained from the chi-square tables at the back of any standard textbook in statistics.[1]

Now let us look at an example. Suppose a sample of 25 mature trout whose lengths have a standard deviation of 4.35 is taken from a commercial fish hatchery. We want a confidence interval for the true population variance σ^2, based on the two statistics $s^2 = 18.9225$ and $n = 25$. From a standard chi-square table we obtain the values of the χ^2 distribution with 24 degrees of freedom below which and above which the probability weight is .005,

$$\chi^2(\alpha/2; n-1) = \chi^2(.005; 24) = 9.89$$

and

$$\chi^2(1-\alpha/2; n-1) = \chi^2(.995; 24) = 45.56.$$

These are indicated by the thin vertical lines in Figure 7.1. Substituting these values into (7.5) after rearranging that expression to put σ^2 on the right-hand-side, we obtain

$$L = \frac{24s^2}{\chi^2(.995; 24)} = \frac{(24)(18.9225)}{45.56} = 9.968$$

and

$$U = \frac{24s^2}{\chi^2(.005; 24)} = \frac{(24)(18.9225)}{9.89} = 45.919$$

so that

$$9.968 \leq \sigma^2 \leq 45.919.$$

[1] Or calculated using XlispStat or another statistical computer program.

Now suppose we want to test whether the population variance of the lengths of trout in this hatchery differs from $\sigma_0^2 = 16.32$, an industry standard, controlling the α-risk at .01 when $\sigma = 16.32$. The null and alternative hypothesis then are

$$H_0 \colon \sigma^2 = 16.32$$

and

$$H_1 \colon \sigma^2 \neq 16.32.$$

From (7.5) the test statistic is

$$X = \frac{(n-1)s^2}{\sigma_0^2}$$

which we have shown to be distributed as $\chi^2(n-1)$. Its value is

$$X = \frac{(24)(18.9225)}{16.32} = 27.82$$

which can be compared the critical values 9.89 and 45.56 beyond which we would reject the null hypothesis of no difference between the variance of the lengths of trout in this hatchery and the industry standard. Clearly, the test statistic falls in the acceptance region so that we cannot reject the null hypothesis.

7.2 Comparisons of Two Population Variances

We are often interested in comparing the variability of two populations. For example, consider a situation where two technicians have made measurements of impurity levels in specimens from a standard solution. One technician measured 11 specimens and the other measured 9 specimens. Our problem is to test whether or not the measurements of impurity levels have the same variance for both technicians.

Suppose that we can assume that the technicians' sets of measurements are independent random samples from normal populations. The sample results are $s_1 = 38.6$ on the basis of the sample $n_1 = 11$ for technician number 1, and $s_2 = 21.7$ on the basis of the sample $n_2 = 9$ for technician number 2.

To proceed further we need a statistic based on the two values of s_i and n_i that is distributed according to an analytically tractable distribution. It turns out that the ratio of two chi-square variables, each divided by their

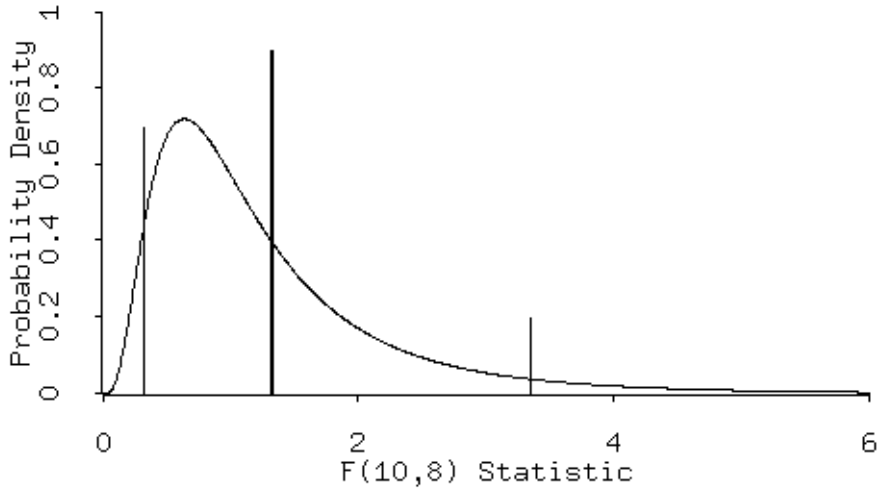


Figure 7.2: An F-distribution with 10 degrees of freedom in the numerator and 8 degrees of freedom in the denominator. The thick vertical line shows the mean and the thin vertical lines the critical values for $\alpha = .90$.

respective degrees of freedom, is distributed according to the *F-distribution.* In particular

$$\frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2} \quad = \quad F(v_1, v_2) \tag{7.6}$$

is distributed according to the F-distribution with parameters v_1 and v_2, which are the degrees of freedom of the respective chi-square distributions—v_1 is referred to as the degrees of freedom in the numerator and v_2 is the degrees of freedom in the denominator. The mean and variance of the F-distribution are

$$E\{F(v_1, v_2)\} = \frac{v_2}{(v_2 - 2)}$$

when $v_2 > 2$, and

$$\sigma^2\{F(v_1, v_2)\} = \frac{2\, v_2^2\, (v_1 + v_2 - 2)}{v_1\, (v_2 - 2)^2\, (v_2 - 4)}$$

when $v_2 > 4$. The probability density function for an F-distribution with 10 degrees of freedom in the numerator and 8 degrees of freedom in the

denominator is plotted in Figure 7.2. The thick vertical line gives the mean and the two thin vertical lines give the critical values for $\alpha = .90$. The percentiles for this distribution can be found in the F-tables at the back of any textbook in statistics.[2] These tables give only the percentiles above 50 percent. To obtain the percentiles below 50 percent we must utilize the fact that the lower tail for the F-value

$$\frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2} = F(v_1, v_2)$$

is the same as the upper tail for the F-value

$$\frac{\chi^2(v_2)/v_2}{\chi^2(v_1)/v_1} = F(v_2, v_1).$$

This implies that

$$F(\alpha/2; v_1, v_2) = \frac{1}{F(1-\alpha/2; v_2, v_1)}.$$

Equation (7.5) can be written more generally as

$$\frac{v\, s^2}{\sigma^2} \;=\; \chi^2(v) \tag{7.7}$$

which implies that

$$\frac{s^2}{\sigma^2} = \frac{\chi^2(v)}{v}.$$

This expression can be substituted appropriately into the numerator and denominator of equation (7.6) to yield

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \;=\; F(v_1, v_2) \;=\; F(n_1 - 1, n_2 - 1). \tag{7.8}$$

To establish confidence intervals for the technician problem, we can manipulate (7.8) to yield

$$\frac{\sigma_2^2}{\sigma_1^2} \;=\; F(n_1 - 1, n_2 - 1)\,\frac{s_2^2}{s_1^2} \;=\; F(10, 8)\,\frac{s_2^2}{s_1^2}$$

$$= \frac{21.7^2}{38.6^2}\,F(10, 8) \;=\; \frac{470.89}{1489.96}\,F(10, 8) \;=\; .31604\,F(10, 8). \tag{7.9}$$

[2]Or calculated using XlispStat or another statistical computer program.

To calculate a 90 percent confidence interval we find the values of $F(10, 8)$ at $\alpha/2 = .05$ and $1 - \alpha/2 = .95$. These are

$$F(.95; 10, 8) = 3.35$$

and

$$F(.05; 10, 8) = \frac{1}{F(.95; 8, 10)} = \frac{1}{3.07} = .3257$$

and are indicated by the thin vertical lines in Figure 7.2. The confidence intervals are thus

$$L = (.3257)(.31604) = .1029$$

and

$$U - (3.35)(.31604) = 1.057$$

so that

$$.1029 \leq \frac{\sigma_2^2}{\sigma_1^2} \leq 1.057.$$

Note that this confidence interval is based on the assumption that the two populations of measurements from which the sample variances are obtained are normally distributed or approximately so.

Since the above confidence interval straddles 1.0, it is clear that there is no indication that the variance of the measurements made by one technician exceeds the variance of the measurements made by the other. Nevertheless, we can test the hypothesis that the variances of the measurements of the two technicians are the same. The null and alternative hypotheses are

$$H_0 \colon \sigma_1^2 = \sigma_2^2$$

and

$$H_1 \colon \sigma_1^2 \neq \sigma_2^2.$$

We want to control the α-risk at 0.1 when $\sigma_1^2 = \sigma_2^2$. Imposing the equal variance assumption on (7.8) we can extract the relationship

$$\frac{s_1^2}{s_2^2} = F(n_1 - 1, n_2 - 1).$$

The statistic on the left of the equality,

$$\frac{s_1^2}{s_2^2} = \frac{38.6^2}{21.7^2} = \frac{1489.96}{470.29} = 3.168$$

is thus distributed as $F(10, 8)$ and is greater than unity. We therefore need only look at the upper critical value $F(.95; 10, 8) = 3.35$ to see that the statistic falls in the acceptance region. We cannot reject the null hypothesis that the variances of the measurements of the two technicians are the same. When performing this test it is always easiest to manipulate the expression to put the largest variance in the numerator and thereby ensure that the sample statistic is bigger than unity. The decision to accept or reject the null hypothesis can then be made on the basis of the easy-to-calculate rejection region in the upper tail of the F-distribution.

7.3 Chi-Square Tests of Goodness of Fit

Statistical tests frequently require that the underlying populations be distributed in accordance with a particular distribution. Our tests of the equality of variances above required, for example, that both the populations involved be normally distributed. Indeed, any tests involving the chi-square or F-distributions require normally distributed populations. A rough way to determine whether a particular population is normally distributed is to examine the frequency distribution of a large sample from the population to see if it has the characteristic shape of a normal distribution. A more precise determination can be made by using a chi-square test on a sample from the population.

Consider, for example, the reported average daily per patient costs for a random sample of 50 hospitals in a particular jurisdiction. These costs were

257	274	319	282	253
315	313	368	306	230
327	267	318	326	255
392	312	265	249	276
318	272	235	241	309
305	254	271	287	258
342	257	252	282	267
308	245	252	318	331
384	276	341	289	249
309	286	268	335	278

with sample statistics $\bar{X} = 290.46$ and $s = 38.21$. We want to test whether the reported average daily costs are normally distributed, controlling the α-risk at .01. The null hypothesis is that they are normally distributed.

The chi-square test is based on a comparison of the sample data with the expected outcome if H_0 is really true. If the hypothesized distribution of the population was a discrete one, we could calculate the probability that each population value X_i will occur and compare that probability with the relative frequency of the population value in the sample. Since the normal distribution is a continuous one, however, the probability that any particular value X_i will occur is zero. So we must compare the probabilities that the X_i could lie in particular intervals with the frequency with which the sample values fall in those intervals.

The standard procedure is to select the intervals or classes to have equal probabilities so that the expected frequencies in all classes will be equal. Also, it is considered desirable to have as many classes as possible consistent with the expected frequencies in the classes being no less than 5. In the above example we therefore need $50/5 = 10$ classes.

To obtain the class intervals we find the values of z in the table of standardised normal values which divide the unit probability weight into 10 equal portions. These will be the z-values for which the cumulative probability density is respectively .1, .2, .3, .4, .5, .6, .7, .8, and .9. The values of X that fall on these dividing lines are thus obtained from the relationship

$$z = \frac{X - \bar{X}}{s}$$

which can be rearranged as

$$X = s\,z + \bar{X} = 38.21\,z + 290.46.$$

This gives us the intervals of z and X in the second and third columns of the table below.

i	z	X	f_i	F_i	$(f_i - F_i)^2$	$(f_i - F_i)^2/F_i$
1	$-\infty$ to -1.28	< 242	3	5	4	0.80
2	-1.28 to -0.84	242 to 258	11	5	36	7.20
3	-0.84 to -0.52	259 to 270	4	5	1	0.20
4	-0.52 to -0.25	271 to 280	6	5	1	0.20
5	-0.25 to -0.00	281 to 290	5	5	0	0.00
6	-0.00 to 0.25	291 to 300	0	5	25	5.00
7	0.25 to 0.52	301 to 310	5	5	0	0.00
8	0.52 to 0.84	311 to 322	7	5	4	0.80
9	0.84 to 1.28	323 to 339	4	5	1	0.20
10	1.28 to ∞	> 339	5	5	0	0.00
		Total	50	50		14.40

Column four gives the actual frequencies with which the sample observations fall in the ith category and column five gives the theoretically expected frequencies. The remaining two columns give the squared differences between the actual and expected frequencies and those squared differences as proportions of the expected frequencies. It turns out that the sum of the right-most column is distributed as χ^2 with 7 degrees of freedom. In general, when there are k classes with equal expected frequencies F_i in all classes and observed frequencies f_i in the ith class,

$$\sum_{k=1}^{k} \frac{(f_i - F_i)^2}{F_i}$$

is distributed as $\chi^2(k-m-1)$ where m is the number of parameters estimated from the sample data. As noted earlier in the definition of the chi-square distribution, the expression $(k-m-1)$ is the number of degrees of freedom. The 10 squared relative deviations give us potentially 10 degrees of freedom, but we have to subtract $m = 2$ because two parameters, $\bar{X}$ and s were estimated from the data, and a further degree of freedom because once we know the frequencies in nine of the ten classes above we can calculate the tenth frequency so only nine of the classes are independent. This leaves us with 7 degrees of freedom.

If the fit were perfect—i.e., the average daily per patient hospital costs were normally distributed—the total at the bottom of the right-most column in the table above would be zero. All the observed frequencies would equal their expected values—i.e., five of the sample elements would fall in each of the 10 classes. Clearly, the greater the deviations of the actual frequencies from expected, the bigger will be the test statistic. The question is then whether the value of the test statistic, 14.4, is large enough to have probability of less than 1% of occurring on the basis of sampling error if the true relative frequencies in the population equal the expected relative frequencies when the population is normally distributed. The critical value for $\chi^2(.99; 7)$ is 18.48, which is substantially above 14.4, so we cannot reject the null hypothesis that the population from which the sample was chosen is normally distributed.

It is interesting to note that the residuals indicate very substantial deviations from normality in two of the classes, 242–258 and 291–300 with the squared deviations from expected frequencies being 36 in the first of these classes and 25 in the second. We might be wise to examine more detailed data for certain of the hospitals to determine whether any reasons for deviations of these two specific magnitudes can be uncovered before we dismiss

these observations as the result of sampling error. Finally, we should keep in mind that in the above test there is only a 1 percent chance that we would reject normality on the basis of sampling error alone if the population is in fact normal. This means that there is up to a 99 percent probability that we will accept normality if the population deviates from it—the β-risk is very high and the power of test is low for small departures from normality. Since it is usually crucial to our research conclusions that the population be normal, the more serious risk would appear to be the risk of accepting normality when it is not true rather than the risk of rejecting normality when it is true. One would like to make H_0 the hypothesis of non-normality and see if the data will lead us to reject it. Unfortunately, this is not possible because there are infinitely many ways to characterize a situation of non-normality. This suggests the importance of using large samples to make these inferences.

7.4 One-Dimensional Count Data: The Multinomial Distribution

Consider a manufacturer of toothpaste who wants to compare the marketability of its own brand as compared to the two leading competitors, A and B. The firm does a survey of the brand preferences of a random sample of 150 consumers, asking them which of the three brands they prefer. The results are presented in the table below.

Brand A	Brand B	Firm's Own Brand
61	53	36

The firm wants to know whether these data indicate that the population of all consumers have a preference for a particular brand.

Notice that the binomial distribution would provide the proper basis for the statistical analysis required here had the question stated "Do you prefer the firm's own brand to its competitors? Yes or No?" Each person's answer—i.e., each random trial—will yield an outcome $X_i = 1$ if the answer is 'yes' and $X_i = 0$ if it is 'no'. If the answer of each consumer in the survey is independent of the answer of all others, and if the probability that the answer of a random person picked will be 'yes' is the same for any person picked at random, then the total number of 'yes' answers,

$$X = \sum_{i=1}^{n} X_i$$

will be distributed as a binomial distribution with parameters p and n. The parameter p is the unknown probability that a person picked at random from the population will say 'yes'.

In the actual example above, the consumer surveyed has to choose between not two options (which would be a simple yes/no comparison) but three—she can prefer either brand A, brand B, or the firm's brand. Each random trial has 3 outcomes instead of 2. There are now three probabilities, p_1, p_2 and p_3, the probabilities of selecting A, B, or the firm's own brand, which must sum to unity. And the firm is interested in the counts n_1, n_2 and n_3 of consumers preferring the respective brands. This experiment is a *multinomial experiment* with k, the number of possible outcomes of each trial, equal to 3. The probabilities of observing various counts n_1, n_2 and n_3, given p_1, p_2 and p_3, is a *multinomial probability distribution.* In the case at hand, p_1, p_2 and p_3 are unknown and we want to make an inference about them on the basis of a sample n. The observed counts will be

$$n_1 + n_2 + n_3 = n.$$

To decide whether the population of consumers prefers a particular brand, we set up the null hypothesis of no preference and see if the data will prompt us to reject it. The null hypothesis is thus

$$H_0\colon\ p_1 = p_2 = p_3.$$

If the null-hypothesis is true we would expect an equal number of the sampled consumers to choose each brand—that is

$$E\{n_1\} = E\{n_2\} = E\{n_3\} = \frac{n}{3} = 50.$$

Notice the similarity of the problem here to the test of normality above. We have three classes each with an expected frequency of 50 and an actual frequency that differs from 50.

i	f_i	F_i	$(f_i - F_i)^2$	$(f_i - F_i)^2/F_i$
A	61	50	121	2.42
B	53	50	9	.18
Own Brand	36	50	196	3.92
Total	150	150		6.52

As in the normality test would expect

$$\sum_{k=1}^{k} \frac{(f_i - F_i)^2}{F_i}$$

to be distributed as $\chi^2(k-m-1)$. The number of classes here is $k=3$, and no parameters were estimated from the sample data so $m=0$. The statistic is thus distributed as $\chi^2(3-1)=\chi^2(2)$. From the chi-square table at the back of any textbook in statistics the critical value for $\chi^2(2)$ for $(\alpha=.05)$ will be found to be 5.99147. Since the total in the right-most column in the table above is 6.52, we can reject the null hypothesis of no brand preference when the α-risk is controlled at .05. The P-value of the statistic is .038. Does this imply a positive or negative preference for the firm's brand of toothpaste as compared to brands A and B? We want now to test whether consumers' preferences for the firm's own brand are greater or less than their preferences for brands A and B. This problem is a binomial one—consumers either prefer the firm's brand or they don't.

We can now use the techniques presented earlier—using a normal approximation to the binomial distribution—to set up a confidence interval for the proportion of the population of consumers choosing the firm's brand of toothpaste. Our sample estimate of p, now the proportion preferring the firm's brand, is

$$\bar{p}=\frac{36}{150}=.24.$$

Using the results in section 9 of Chapter 4, the standard deviation of $\bar{p}$ is

$$s_{\bar{p}}=\sqrt{\frac{\bar{p}\,(1-\bar{p})}{n-1}}=\sqrt{\frac{(.24)(.76)}{149}}=\sqrt{.00122316}=.03497.$$

The 95 percent confidence interval for p is thus (using the critical value $z=1.96$ from the normal distribution table)

$$.24\pm(1.96)(.03497)=.24\pm.0685412,$$

or

$$.17\leq p\leq .30984.$$

It is clear from this confidence interval that less than 1/3 of consumers prefer the firm's own brand of toothpaste, contrary to what one would expect under the null hypothesis of no differences in consumer preference. Indeed, we can test the hypothesis of no difference in preference between the firm's brand of toothpaste and other brands by setting up the null and alternative hypotheses

$$H_0\colon\ p=\frac{1}{3}$$

and

$$H_1\colon\ p\neq\frac{1}{3}$$

and calculating

$$z^* = \frac{\bar{p} - p}{s_p}$$

where

$$s_p = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{(\frac{1}{3})(\frac{2}{3}}{150}} = \sqrt{.00148148148} = .03849.$$

Notice that we use the value of p under the null hypothesis here instead of $\bar{p}$. Thus we have

$$z^* = \frac{\bar{p} - p}{s_p} = \frac{.24 - .333}{.03849} = \frac{.09333}{.03849} = 2.42.$$

The critical value of z for a two-tailed test with $\alpha = .05$ is 1.96. Clearly, we are led to reject the null hypothesis of no difference in preference. Indeed we could reject the null hypotheses that the population of consumers prefers the firm's brand to other brands with an α-risk of less than .01 because the P-value for a one-tailed test is .00776.

7.5 Contingency Tables: Tests of Independence

In the multinomial distribution above the data were classified according to a single criterion—the preferences of consumers for the three brands of toothpaste. Now we turn to multinomial distributions involving data that are classified according to more than one criterion.

Consider, for example, an economist who wants to determine if there is a relationship between occupations of fathers and the occupations of their sons. She interviewed 500 males selected at random to determine their occupation and the occupation of their fathers. Occupations were divided into four classes: professional/business, skilled, unskilled, and farmer. The data are tabulated as follows:

		Occupation of Son				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
Occupation of Father	Prof/Bus	55	38	7	0	100
	Skilled	79	71	25	0	175
	Unskilled	22	75	38	10	145
	Farmer	15	23	10	32	80
	Total	171	207	80	42	500

The problem the economist faces is to determine if this evidence supports the hypothesis that sons' occupations are related to their fathers'. We can visualize there being a joint probability density function over all father-occupation, son-occupation pairs giving the probability that each combination of father and son occupations will occur. Treated as a table of probabilities, the above table would appear as

		Occupation of Son				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
Occupation of Father	Prof/Bus	p_{11}	p_{12}	p_{13}	p_{14}	p_{r1}
	Skilled	p_{21}	p_{22}	p_{23}	p_{24}	p_{r2}
	Unskilled	p_{31}	p_{32}	p_{33}	p_{34}	p_{r3}
	Farmer	p_{41}	p_{42}	p_{43}	p_{44}	p_{r4}
	Total	p_{c1}	p_{c2}	p_{c3}	p_{c4}	1.00

where the probabilities along the right-most column p_{ri} are the marginal probabilities of fathers' occupations—i.e., the sum of the joint probabilities p_{ij} in the ith row and jth column over the j columns—and the probabilities along the bottom row p_{cj} are the marginal probabilities of sons' occupations—i.e., the sum of the joint probabilities p_{ij} over the i rows.

The count data, since they indicate the frequencies for each cell, can be thought of as providing point estimates of these probabilities. The marginal

probabilities along the bottom row and the right-most column are the cell entries divided by 500 as shown in the table below. We know from the definition of statistical independence that if events A and B are independent,

$$P(A|B) = P(A)$$

which implies that

$$P(A \cap B) = P(A|B)P(B) = P(A)P(B).$$

Hence the joint probabilities in each cell of the table below should equal the product of the marginal probabilities for that particular row and column. The joint probabilities under the null hypothesis that fathers' occupations and sons' occupations are independent are as given below.

		Occupation of Son				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
Occupation of Father	Prof/Bus	.0684	.0828	.0320	.0168	.20
	Skilled	.1197	.1449	.0560	.0294	.35
	Unskilled	.0992	.1201	.0464	.0244	.29
	Farmer	.0547	.0662	.0256	.0134	.16
	Total	.342	.414	.16	.084	1.00

This means that if the occupations of sons were independent of the occupations of their fathers the number or frequency of fathers and sons who were both in the professional or business category would be the joint probability of this outcome (.0684) times the number of sons sampled (500). Accordingly, we can calculate the expected number or expected count in each cell by multiplying the joint probability for that cell by 500. This yields the following table of actual and expected outcomes, with the expected outcomes in brackets below the actual outcomes.

		Occupation of Son				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
Occupation of Father	Prof/Bus	55	38	7	0	100
		(34.2)	(41.4)	(16.0)	(8.4)	
	Skilled	79	71	25	0	175
		(59.85)	(72.45)	(28.0)	(14.7)	
	Unskilled	22	75	38	10	145
		(49.6)	(60.05)	(23.2)	(12.2)	
	Farmer	15	23	10	32	80
		(27.34)	(33.10)	(12.8)	(6.7)	
	Total	171	207	80	42	500

From this point the procedure is the same as in the test of normality. The tabulation, working from left to right, row by row, is as follows:

Father–Son	f_i	F_i	$(f_i - F_i)^2$	$(f_i - F_i)^2/F_i$
Prof/Bus–Prof/Bus	55	34.20	432.64	12.65
Prof/Bus–Skilled	38	41.40	11.56	0.28
Prof/Bus–Unskilled	7	16.00	81.00	5.06
Prof/Bus–Farmer	0	8.40	70.56	8.40
Skilled–Prof/Bus	79	59.85	366.72	6.13
Skilled–Skilled	71	72.45	2.10	0.03
Skilled–Unskilled	25	28.00	9.00	0.32
Skilled–Farmer	0	14.70	216.09	14.70
Unskilled–Prof/Bus	22	49.60	761.76	15.35
Unskilled–Skilled	75	60.05	223.50	3.72
Unskilled–Unskilled	38	23.20	219.04	9.44
Unskilled–Farmer	10	12.20	4.84	0.40
Farmer–Prof/Bus	15	27.34	152.28	5.57
Farmer–Skilled	23	33.10	102.01	3.08
Farmer–Unskilled	10	12.80	7.84	0.61
Farmer–Farmer	32	6.70	640.09	95.54
Total	500	500.00		181.28

It turns out that the total sum of squared relative deviations from expected values, represented by the number 181.28 at the bottom of the right-most column,

$$\sum_{k=1}^{k} \frac{(f_i - F_i)^2}{F_i},$$

Table 7.1: Percentage of Sons' Occupations by Father's Occupation

		Father's Occupation				
		Prof/Bus	Skilled	Unskilled	Farmer	Total
Son's Occupation	Prof/Bus	55	45	15	19	34
	Skilled	38	41	52	29	42
	Unskilled	7	14	26	12	16
	Farmer	0	0	7	40	8
	Total	100	100	100	100	100

is distributed according to a chi-square distribution with degrees of freedom equal to the product of the number of rows minus one and the number of columns minus one—i.e., $\chi^2((nr-1)(nc-1))$, where nr and nc are, respectively, the number of rows and columns in the contingency table. In the case at hand, the total is distributed as $\chi^2(9)$. The critical value for α-risk = .01 from the chi-square table for 9 degrees of freedom is 21.6660. Since the total in the right-most column of the table vastly exceeds that critical value, we must reject the hypothesis of independence and conclude that sons' occupations depend on the occupations of their fathers.

The pattern of dependence can be seen more clearly when we take the percentage of sons in each category of father's occupation and compare them with the overall percentage of sons in each occupation. This is done in the table immediately above. Each column of the table gives the percentage of sons of fathers in the occupation indicated at the top of that column who are in the various occupations listed along the left margin of the table. The right-most column gives the percentage of all sons in the respective occupations.

If sons' occupations were independent of their fathers', 34 percent of the sons in each father's-occupation category would be in professional/business occupations. As can be seen from the table, 55 percent of the sons of professional/business fathers and 45 percent of the sons of fathers in skilled trades are in professional/business occupations. Yet only 15 and 19 percent, respec-

tively, of sons of unskilled and farmer fathers work in the professions and business. If sons' occupations were unrelated to their fathers' occupations, 42 percent would be in skilled occupations, regardless of the occupation of the father. It turns out from the table that 52 percent of sons of unskilled fathers are in skilled trades and less than 42 percent of the sons of fathers in each of the other categories are skilled workers. Judging from this and from the 45 percent of sons of skilled fathers who are in professional/business occupations, it would seem that the sons of skilled fathers tend either to move up into the business/professional category or fall back into the unskilled category, although the percentage of sons of skilled fathers who are also skilled is only slightly below the percentage of all sons who are skilled. If there were no occupational dependence between fathers and sons, 16 percent of sons of unskilled fathers would also be in unskilled work. As we can see from the table, 26 percent of the sons of unskilled workers are themselves unskilled and less than 16 percent of the sons of unskilled fathers are in each of the other three occupational categories. Finally, if the occupations of fathers and their sons were statistically independent we would expect that 8 percent of the sons of farmers would be in each occupational category. In fact, 40 percent of the sons of farmers are farmers, 7 percent of the sons of unskilled fathers are farmers, and none of the sons of fathers in the skilled and professional/business occupations are farmers.

The dependence of son's occupation on father's occupation can also be seen from the table by drawing a wide diagonal band across the table from top left to bottom right. The frequencies tend to be higher in this diagonal band than outside it, although there are exceptions. This indicates that sons' occupations tend to be the same or similar to their fathers'. Sons' occupations and the occupations of their fathers are statistically dependent.

7.6 Exercises

1. Independent random samples were selected from each of two normally distributed populations. The sample results are summarized as follows:

Sample 1	Sample 2
$n_1 = 10$	$n_2 = 23$
$\bar{X}_1 = 31.7$	$\bar{X}_2 = 37.4$
$s_1^2 = 3.06$	$s_2^2 = 7.60$

Setting the α-risk at 0.05, test the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis H_1: $\sigma_1^2 \neq \sigma_2^2$,

2. A financial analyst is exploring the relationship between the return earned by stocks and the return earned by bonds. For a period of $n = 25$ months, the analyst records the return on a particular stock, denoted X, and the return on a particular bond, denoted Y. The relevant sample statistics are recorded below:

Monthly Returns	Stock (X)	Bond (Y)
Mean	1.5	1.2
Standard Deviation	1.0	0.8

Assume that X and Y are uncorrelated and perform hypotheses tests to determine whether the two population variances are equal. Then perform a test to determine whether the two population means are equal. How would your answer change if it turned out that the sample correlation between X and Y was $r_{xy} = -0.20$.

3. A labour economist studied the durations of the most recent strikes in the vehicles and construction industries to see whether strikes in the two industries are equally difficult to settle. To achieve approximate normality and equal variances, the economist worked with the logarithms (to the base 10) of the duration data (expressed in days). In the vehicle industry there were 13 strikes having a mean log-duration of 0.593 and a standard deviation of log-duration of 0.294. In the construction industry there were 15 strikes with a mean log-duration was 0.973 and a standard deviation of log-duration of 0.349. The economist believes that it is reasonable to treat the data as constituting independent random samples.

a) Construct and interpret a 90 percent confidence interval for the difference in the mean log-durations of strikes in the two industries.

b) Test whether the strikes in the two industries have the same log-durations, controlling the α risk at 0.10. State the alternatives, the decision rule, the value of the test statistic and the conclusion.

c) Test the economist's assumption that the log-durations of strikes in the two industries have the same variance controlling the α risk at 0.10. State the alternatives, the decision rule, the value of the test statistic and the conclusion.

4. An industrial machine has a 1.5-meter hydraulic hose that ruptures occasionally. The manufacturer has recorded the location of these ruptures for 25 ruptured hoses. These locations, measured in meters from the pump end of the hose, are as follows:

1.32	1.07	1.37	1.19	0.13
1.14	1.21	1.16	1.43	0.97
0.33	1.36	0.64	1.42	1.12
1.46	1.27	0.27	0.80	0.08
1.46	1.37	0.75	0.38	1.22

Using the chi-square procedure, test whether the probability distribution of the rupture locations is uniform with lowest value $a = 0$ and highest value $b = 1.5$.

5. A city expressway utilizing four lanes in each direction was studied to see whether drivers prefer to drive on the inside lanes. A total of 1000 automobiles was observed during the heavy early-morning traffic and their respective lanes recorded. The results were as follows:

Lane	Observed Count
1	294
2	276
3	238
4	192

Do these data present sufficient evidence to indicate that some lanes are preferred over others? Use $\alpha = .05$ in your test.

6. It has been estimated that employee absenteeism costs North American companies more than $100 billion per year. As a first step in addressing the rising cost of absenteeism, the personnel department of a large corporation recorded the weekdays during which individuals in a sample of 362 absentees were away from work over the past several months:

	Number Absent
Monday	87
Tuesday	62
Wednesday	71
Thursday	68
Friday	74

Do these data suggest that absenteeism is higher on some days of the week than others?

7. The trustee of a company's pension plan has solicited the opinions of a sample of the company's employees about a proposed revision of the plan. A breakdown of the responses is shown in the accompanying table. Is there evidence at the 10% level to infer that the responses differ among the three groups of employees?

Responses	Blue-Collar Workers	White Collar Workers	Managers
For	67	32	11
Against	63	18	9

8. A study of the amount of violence viewed on television as it relates to the age of the viewer showed the accompanying results for 81 people. Each person in the study could be classified according to viewing habits as a low-violence or high-violence viewer.

	16–34 yrs. old	35–54 yrs. old	55 yrs. and over
Low Violence	8	12	21
High Violence	18	15	7

Do the data indicate that viewing of violence is not independent of age of viewer at the 5% significance level?

9. To see if there was any dependency between the type of professional job held and one's religious affiliation, a random sample of 638 individuals belonging to a national organization of doctors, lawyers and engineers were chosen in a 1968 study. The results were as follows:

	Doctors	Lawyers	Engineers
Protestant	64	110	152
Catholic	60	86	78
Jewish	57	21	10

Test at the 5 percent level of significance the hypothesis that the profession of individuals in this organization and their religious affiliation are independent. Repeat at the 1 percent level.

10. To study the effect of fluoridated water supplies on tooth decay, two communities of roughly the same socio-economic status were chosen. One of these communities had fluoridated water while the other did not. Random samples of 200 teenagers from both communities were chosen and the numbers of cavities they had were determined. The results were as follows:

Cavities	Fluoridated Town	Nonfluoridated Town
0	154	133
1	20	18
2	14	21
3 or more	12	28

Do these data establish, at the 5 percent level of significance, that the number of dental cavities a person has is not independent of whether that person's water supply is fluoridated? What about at the 1% level?

Chapter 8

Simple Linear Regression

We now turn to the area of statistics that is most relevant to what economists usually do—the analysis of relationships between variables. Here we will concentrate entirely on linear relationships. For example, we might be interested in the relationship between the quantity of money demanded and the volume of transactions that people make as represented by the level of money income. Or we might be interested in the relationship between family expenditures on food and family income and family size. *Regression analysis* is used to analyse and predict the relationship between the *response* or *dependent* variable (money holdings and family expenditure on food in the above examples) and one or more *independent*, *explanatory*, or *predictor* variables. In the demand for money example the single independent variable was the level of income; in the family food expenditure example, there were two independent variables, family income and family size.

We must distinguish two types of relationships between variables. A *deterministic* relationship exists if the value of Y is uniquely determined when the value of X is specified—the relationship between the two variables is exact. For example, we might have

$$Y = \beta X$$

where β is some constant such as 10. On the other hand, there may be a relationship between two variables that involves some random component or random error. This relationship is called a *probabilistic* or *statistical* relationship. In this case we might have

$$Y = \beta X + \epsilon$$

which can be viewed as a *probabilistic model* containing two components—a *deterministic component* βX plus a *random error* ϵ. Figure 8.1 presents

an example of a deterministic straight-line relationship between X and Y along which all observed combinations of the two variables lie. An example of a probabilistic relationship is given in Figure 8.2. There is a *scatter* of observed combinations of X and Y around a straight-line functional or deterministic relationship indicating errors in the fit that result from the influence on Y of unknown factors in addition to X. For each level of X there is a probability distribution of Y. And the means of these probability distributions of Y vary in a systematic way with the level of X.

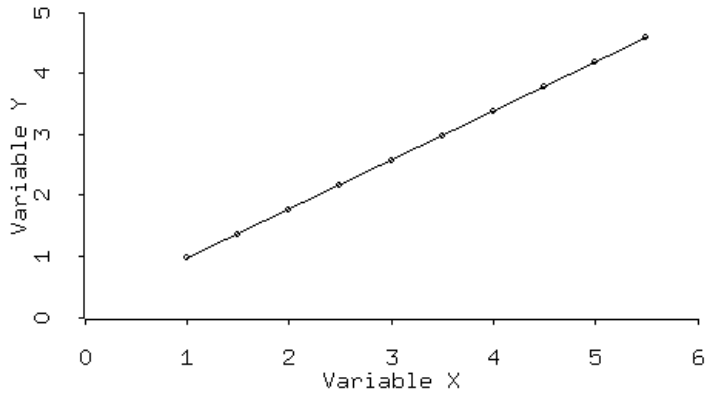


Figure 8.1: A functional or deterministic relationship between two variables X and Y.

8.1 The Simple Linear Regression Model

When the statistical relationship is linear the regression model for the observation Y_i takes the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{8.1}$$

where the functional or deterministic relationship between the variables is given by $\beta_0 + \beta_1 X_i$ and ϵ_i is the random scatter component. Y_i is the dependent variable for the ith observation, X_i is the independent variable for the ith observation, assumed to be non-random, β_0 and β_1 are parameters and the ϵ_i are the deviations of the Y_i from their predicted levels based on X_i, β_0 and β_1.

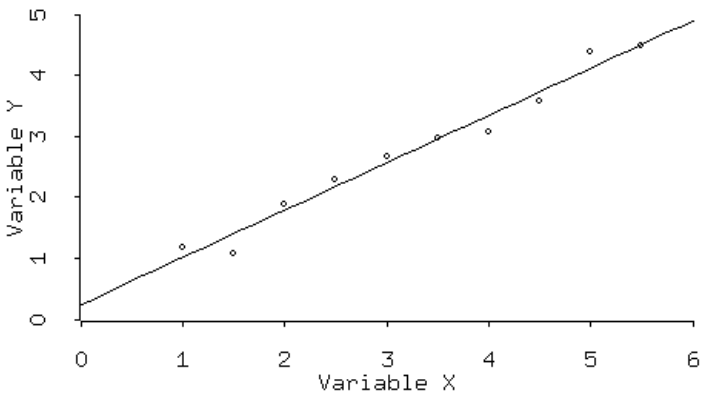


Figure 8.2: An probabilistic or statistical relationship between two variables X and Y.

The error term is assumed to have the following properties:

a) The ϵ_i are normally distributed.

b) The expected value of the error term, denoted by $E\{\epsilon_i\}$, equals zero.

c) The variance of the ϵ_i is a constant, σ^2.

d) The ϵ_i are statistically independent—that is, the covariance between ϵ_i and ϵ_j is zero.

In other words,

$$\epsilon_i = N(0, \sigma^2).$$

This normality assumption for the ϵ_i is quite appropriate in many cases. There are often many factors influencing Y other than the independent variable (or, as we shall see later, variables) in the regression model. Insofar as the effects of these variables are additive and tend to vary with a degree of mutual independence, their mean (and their sum) will tend to normality according to the central limit theorem when the number of these 'missing' factors is large. The distribution of the error term and the resulting levels of Y at various levels of X is given in Figure 8.3.

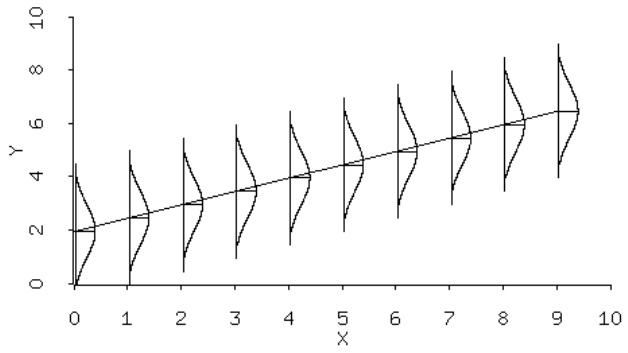


Figure 8.3: Simple regression of Y on X: The probability distribution of Y given X.

Since the error term ϵ_i is a random variable, so is the dependent variable Y_i. The expected value of Y_i equals

$$\begin{aligned} E\{Y_i\} &= E\{\beta_0 + \beta_1 X_i + \epsilon_i\} \\ &= E\{\beta_0\} + E\{\beta_1 X_i\} + E\{\epsilon_i\} \\ &= \beta_0 + \beta_1 E\{X_i\} + 0 \\ &= \beta_0 + \beta_1 X_i \end{aligned} \tag{8.2}$$

where $E\{X_i\} = X_i$ because these X_i are a series of pre-determined non-random numbers. Equation (8.2), the underlying deterministic relationship is called the *regression function*. It the *line of means* that relates the mean of Y to the value of the independent variable X. The parameter β_1 is the slope of this line and β_0 is its intercept.

The variance of Y_i given X_i equals

$$\begin{aligned} Var\{Y_i|X_i\} &= Var\{\beta_0 + \beta_1 X_i + \epsilon_i\} \\ &= Var\{\beta_0 + \beta_1 X_i\} + Var\{\epsilon_i\} \\ &= 0 + Var\{\epsilon_i\} \;=\; \sigma^2 \end{aligned} \tag{8.3}$$

where the regression function $\beta_0 + \beta_1 X_i$ is deterministic and therefore does not vary. Thus the Y_i have the same variability around their means at all X_i.

Finally, since the ϵ_i are assumed to be independent for the various observations, so are the Y_i conditional upon the X_i. Hence it follows that

$$Y_i = N(\beta_0 + \beta_1 X_i, \sigma^2).$$

8.2 Point Estimation of the Regression Parameters

Point estimates of β_0 and β_1 can be obtained using a number of alternative estimators. The most common estimation method is the *method of least squares*. This method involves choosing the estimated regression line so that the sum of the squared deviations of Y_i from the value predicted by the line is minimized. Let us denote the deviations of Y_i from the fitted regression line by e_i and our least-squares estimates of β_0 and β_1 by b_0 and b_1 respectively. Then we have

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 \tag{8.4}$$

where Q is the sum of squared deviations of the Y_i from the values predicted by the line.

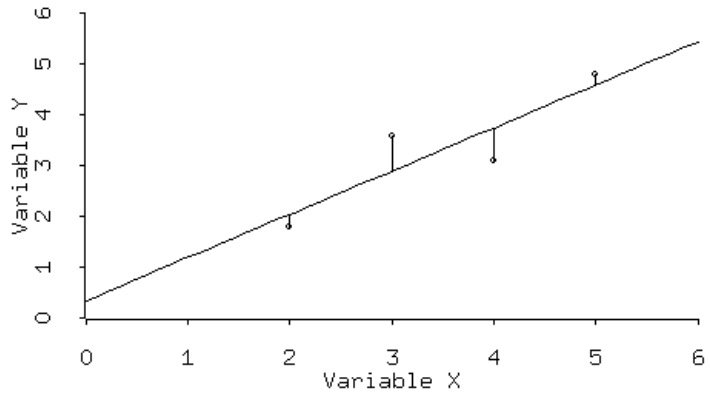


Figure 8.4: A least-squares fit minimizes the sum of the squared vertical distances of the data-points from the least-squares line.

The least-squares estimation procedure involves choosing b_0 and b_1, the intercept and slope of the line, so as to minimize Q. This minimizes the sum

of the squared lengths of the vertical lines in Figure 8.4. Expanding equaton (8.4), we have

$$\begin{aligned} Q &= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2 \\ &= \sum_{i=1}^{n} Y_i^2 + n\, b_0^2 + \sum_{i=1}^{n} b_1^2 X_i^2 - 2\, b_0 \sum_{i=1}^{n} Y_i \\ &\qquad -2\, b_1 \sum_{i=1}^{n} Y_i X_i + 2\, b_0 b_1 \sum_{i=1}^{n} X_i \end{aligned} \tag{8.5}$$

To find the least squares minimizing values of b_0 and b_1 we differentiate Q with respect to each of these parameters and set the resulting derivatives equal to zero. This yields

$$\frac{\partial Q}{\partial b_0} = 2\, n\, b_0 - 2 \sum_{i=1}^{n} Y_i + 2\, b_1 \sum_{i=1}^{n} X_i = 0 \tag{8.6}$$

$$\frac{\partial Q}{\partial b_1} = 2\, b_1 \sum_{i=1}^{n} X_i^2 - 2 \sum_{i=1}^{n} X_i Y_i + 2\, b_0 \sum_{i=1}^{n} X_i = 0 \tag{8.7}$$

which simplify to

$$\sum_{i=1}^{n} Y_i = n\, b_0 + b_1 \sum_{i=1}^{n} X_i \tag{8.8}$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \tag{8.9}$$

These two equations can now be solved simultaneously for b_0 and b_1. Dividing (8.8) by n, rearranging to put b_0 on the left side and noting that $\sum X_i = n\bar{X}$ and $\sum Y_i = n\bar{Y}$ we obtain

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{8.10}$$

Substituting this into (8.9), we obtain

$$\sum_{i=1}^{n} X_i Y_i = \bar{Y} \sum_{i=1}^{n} X_i - b_1 \bar{X} \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2, \tag{8.11}$$

which can be rearranged to yield

$$\begin{aligned} \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i &= b_1 \left[\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i \right] \\ \sum_{i=1}^{n} X_i Y_i - n\, \bar{Y} \bar{X} &= b_1 \left[\sum_{i=1}^{n} X_i^2 - n\, \bar{X}^2 \right] \end{aligned} \tag{8.12}$$

By expansion it can be shown that

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\,\bar{Y}\,\bar{X}$$

and

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2$$

so that by substitution into (8.12) we obtain

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{8.13}$$

This expression can be alternatively written as

$$b_1 = \frac{\sum xy}{x^2} \tag{8.14}$$

where $x = (X_i - \bar{X})$ and $y = (Y_i - \bar{Y})$ are the deviations of the variables from their respective means and the summation is over $i = 1 \ldots n$.

The least-squares estimators b_0 and b_1 are unbiased and, as can be seen from (8.10) and (8.13), linearly dependent on the n sample values Y_i. It can be shown that least-squares estimators are more efficient—that is, have lower variance—than all other possible unbiased estimators of β_0 and β_1 that are linearly dependent on the Y_i. It can also be shown that these desirable properties do not depend upon the assumption that the ϵ_i are normally distributed.

Estimators of β_0 and β_1 can also be developed using the method of maximum likelihood (under the assumption that the ϵ_i are normally distributed). These estimators turn out to be identical with the least-squares estimators.

Calculation of the regression line is straight forward using (8.10) and (8.14). The procedure is to

a) calculate the deviations of X_i and Y_i from their respective means.

b) square the deviations of the X_i and sum them.

c) multiply the X_i deviations with the corresponding Y_i deviations and sum them.

d) plug these sums of squares and cross products into (8.14) to obtain b_1, and

e) plug this value of b_1 into (8.10) along with the means of the X_i and Y_i to obtain b_0.

The regression function $E\{Y\} = \beta_0 + \beta_1 X$ is estimated as

$$\hat{Y} = b_0 + b_1 X \tag{8.15}$$

where $\hat{Y}$ is referred to as the *predicted* value of Y. The mean response or predicted value of Y when X takes some value X_h is

$$\hat{Y}_h = b_0 + b_1 X_h.$$

The point estimate of $E\{Y_h\}$ is thus $\hat{Y}_h$, the value of the estimated regression function when $X = X_h$.

8.3 The Properties of the Residuals

To make inferences (i.e., construct confidence intervals and do statistical tests) in regression analysis we need to estimate the magnitude of the random variation in Y. We measure the scatter of the observations around the regression line by comparing the observed values Y_i with the predicted values associated with the corresponding X_i. The difference between the observed and predicted values for the ith observation is the *residual* for that observation. The residual for the ith observation is thus

$$e_i = Y_i - b_0 - b_1 X_i.$$

Note that e_i is the *estimated residual* while ϵ_i is the *true* residual or *error term* which measures the deviations of Y_i from its true mean $E\{Y\}$.

The least-squares residuals have the following properties.

a) They sum to zero — $\sum e_i = 0$.

b) The sum of the squared residuals $\sum e_i^2$ is a minimum—this follows because the method of least squares minimizes Q.

c) The sum of the weighted residuals is zero when each residual is weighted by the corresponding level of the independent variable — $\sum X_i e_i = 0$.

d) The sum of the weighted residuals is zero when each residual is weighted by the corresponding fitted value — $\sum \hat{Y}_i e_i = 0$.

8.4 The Variance of the Error Term

To conduct statistical inferences about the parameters of the regression we are going to need an estimate of the variance of the error term. An obvious way to proceed is to work with the sum of squared deviations of the observed levels of Y from the predicted levels—i.e.,

$$\sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \sum_{i=1}^{n} e_i^2.$$

It turns out that the mean or average of these squared deviations is the appropriate estimator of σ^2, provided we recognize that all n of of these deviations are not independent. Since we used the sample data to estimate two parameters, b_0 and b_1, we used up two of the n pieces of information contained in the sample. Hence, there are only $n-2$ independent squared deviations—i.e., $n-2$ degrees of freedom. Hence, in taking the average we divide by $n-2$ instead of n. An unbiased estimator of σ^2 is

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{8.16}$$

where *MSE* stands for *mean square error*. In general, a mean square is a sum of squares divided by the degrees of freedom with which it is calculated.

8.5 The Coefficient of Determination

Consider the sum of the squared deviations of the Y_i from their mean $\bar{Y}$, otherwise known as the total sum of squares and denoted by *SSTO*,

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

This total sum of squares can be broken down into components by adding and subtracting $\hat{Y}$ as follows:

$$\begin{aligned} SSTO &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \right]^2 \\
&= \sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \right] \\
&= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}). \quad (8.17)
\end{aligned}
$$

The term

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

equals zero, since

$$
\begin{aligned}
\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i)\hat{Y}_i - (Y_i - \hat{Y}_i)\bar{Y} \right] \\
&= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)\hat{Y}_i - \sum_{i=1}^{n} (Y_i - \hat{Y}_i)\bar{Y} \\
&= \sum_{i=1}^{n} e_i \hat{Y}_i \; - \; \bar{Y} \sum_{i=1}^{n} e_i. \quad (8.18)
\end{aligned}
$$

From the properties a) and d) of the least-squares residuals listed on page 200 above, $\hat{Y} \sum e_i$ and $\sum e_i Y_i$ are both zero. We can thus partition the total sum of squares into the two components,

$$
\begin{aligned}
SSTO &= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
&= \quad SSR \quad + \quad SSE. \quad (8.19)
\end{aligned}
$$

The term

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

is the sum of squares of the deviations of the observed values Y_i from the values predicted by the regression. It is the portion of the total variability of Y that remains as a residual or error after the influence of X is considered, and is referred to as the *sum of squared errors*. The term

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

is the sum of the squared deviations of the predicted values Y_i from the mean of Y. It is the portion of the total variability of Y that is explained by the regression—that is, by variations in the independent variable X. It follows that the sum of squared errors is the total sum of squares minus the portion explained by X—i.e., $SSE = SSTO - SSR$.

The *coefficient of determination*, usually referred to as the R^2, is the fraction of the total variability of the Y_i that is explained by the variability of the X_i. That is,

$$R^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO}. \tag{8.20}$$

8.6 The Correlation Coefficient Between X and Y

The correlation coefficient between two random variables, X and Y has previously been defined as

$$\rho = \frac{Cov\{XY\}}{\sqrt{Var\{X\}Var\{Y\}}}. \tag{8.21}$$

An appropriate estimator of ρ is

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}. \tag{8.22}$$

As in the case of the true correlation coefficient ρ, r can vary between minus unity and plus unity. In the present situation, however, the X_i are assumed fixed—i.e., do not vary from sample to sample—so that X is not a random variable. Nevertheless, r is still a suitable measure of the degree of association between the variable Y and the fixed levels of X. Moreover, when we square r we obtain

$$r^2 = \frac{(\sum(X_i - \bar{X})(Y_i - \bar{Y}))^2}{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2} \tag{8.23}$$

which, it turns out, can be shown to equal R^2 as defined above.

8.7 Confidence Interval for the Predicted Value of Y

Suppose we want to estimate the mean level of Y for a given level of X and establish confidence intervals for that mean level of Y. For example, an admissions officer of a U.S. college might wish to estimate the mean grade point average (GPA) of freshmen students who score 550 on the Scholastic Aptitude Test (SAT).

We have already established that the predicted value Y_h is a good point estimator of $E\{Y_h\}$. In order to obtain confidence intervals for $E\{Y_h\}$, however, we need a measure of the variance of $\hat{Y}_h$. It turns out that

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{8.24}$$

for which an appropriate estimator is

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{8.25}$$

where *MSE* is the mean square error, previously defined as

$$MSE = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}.$$

The magnitude of the estimated variance $s^2\{\hat{Y}_h\}$ is affected by a number of factors:

a) It is larger the greater the variability of the residuals e_i.

b) It is larger the further the specified level of X is from the mean of X in either direction—i.e., the bigger is $(X_h - \bar{X})^2$.

c) It is smaller the greater the variability of the X_i about the mean of X.

d) It is smaller the greater the sample size n. There are two reasons for this. The greater is n, the smaller are both $1/n$ and *MSE* and, in addition, when n is larger the sum of the squared deviations of the X_i from their mean will tend to be larger.

The above points can be seen with reference to Figure 8.5. The true functional relationship is given by the thick solid line and has slope β_1 and intercept β_0. Alternative fitted regression lines are given by the upward

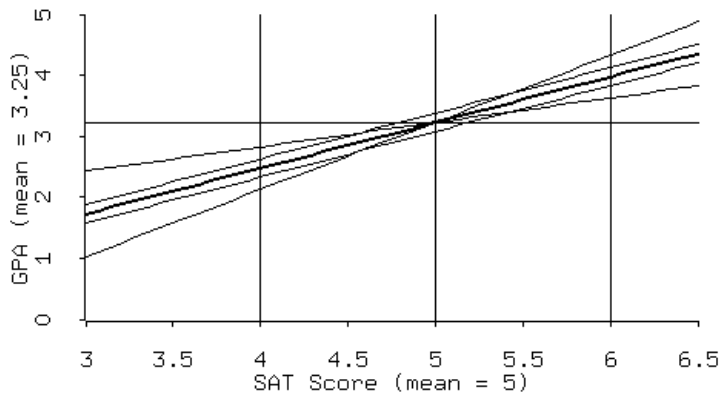


Figure 8.5: The true linear functional relationship (thick line) between Scholastic Aptitude Test (SAT) score and subsequent Grade Point Average (GPA) measured on a 5 point scale in freshman college courses, together with some possible fitted regression lines based on differing samples.

sloping thin lines. Each regression line always passes through the point $(\bar{X}\bar{Y})$ for the sample in question. Different samples of Y_i's drawn for the same set of X_i's yield different regression lines having different slopes since b_1 is a random variable. Also, different samples will yield different mean values of Y, though $\bar{X}$ will be the same because the X_i are fixed from sample to sample. This means that the level of the regression line is also a random variable as shown by the thin lines parallel to the true functional relationship—its variance at $\bar{X}$ is the variance of the error term σ^2 which is estimated by *MSE*.

The estimated variance of the predicted values of Y at $\bar{X}$, associated in the above example with a SAT score of 500, will be equal to *MSE* divided by n and will be determined entirely by the variance of the level of the line. At levels of X above the mean, say for a SAT score of 600, the variance of the predicted value of Y will be larger because there is both variance in the level of the regression line and variance of the slope of the line pivoting on $(\bar{X}\bar{Y})$. The further away one gets from the mean value of X, the bigger is the effect on the variance of the predicted Y of the variation of b_1 from sample to sample. Also, notice that the variance of the predicted Y at a SAT score of 400 will be the same as the variance of the predicted Y at a SAT

score of 600 because the effect of the sampling variation of b_1 is the same at both points (which are equidistant from $\bar{X}$) and the effect of sampling variation on the level of the regression line is the same at all X_i since it depends on $\bar{X}$ which is constant from sample to sample. We can now form the standardised statistic

$$\frac{\hat{Y}_h - E\{\hat{Y}_h\}}{s\{\hat{Y}_h\}}$$

which is distributed according to the t-distribution with $n-2$ degrees of freedom. There are two less degrees of freedom than the number of observations because we used the sample to estimate two parameters, β_0 and β_1. The confidence limits for $E\{Y_h\}$ with confidence coefficient α are thus

$$\hat{Y}_h \;\pm\; t(1-\alpha/2; n-2)\, s\{\hat{Y}_h\}.$$

This confidence interval is interpreted for repeated samples when the X_i are the same from sample to sample. Of many confidence intervals so established based on repeated samples, 100α percent will bracket $E\{Y_h\}$.

8.8 Predictions About the Level of Y

Suppose that we want to predict the grade point average of a student with a SAT score X_h equal to 600. It is important to distinguish this prediction, and the confidence interval associated with it, from predictions about the mean level of Y_h, the point estimator of which was $\hat{Y}_h$. That is, we want to predict the *level* of Y associated with a *new* observation at some X_h, not the *mean value* of Y associated with a whole sample drawn at a value of X equal to X_h. Predicting the grade point average of a randomly selected student who scored 600 on the SAT is very different from predicting what the mean grade point average of students who score 600 on the SAT will be.

If we knew the true values of the regression parameters, β_0, β_1 and σ, the procedure would be quite simple. We could simply calculate

$$E\{Y_h\} = \beta_0 + \beta_1 X_h$$

which might equal, say, 3.7. This would be the point estimate of $Y_{h(new)}$, the newly selected student's grade point average. We could then use the known value of σ to establish a confidence interval for an appropriate value of α.

But we don't know the true regression parameters and so must estimate them. The statistic $\hat{Y}_h$ is an appropriate point estimator of $Y_{h(new)}$. To get a

confidence interval we must estimate the variance of $Y_{h(new)}$. This variance is based on the variance of the difference between Y_h and $\hat{Y}_h$ together with the assumption that the new observation is selected independently of the original sample observation. This yields

$$\begin{aligned} \sigma^2\{\hat{Y}_{h(new)}\} &= \sigma^2\{Y_h - \hat{Y}_h\} \\ &= \sigma^2\{Y_h\} + \sigma^2\{\hat{Y}_h\} \\ &= \sigma^2 + \sigma^2\{\hat{Y}_h\} \end{aligned} \tag{8.26}$$

which is composed of two parts. It is the sum of

a) the variance of the mean predicted level of Y associated with the particular level of X.

b) the variance of the actual level of Y around its predicted mean level, denoted by σ^2.

In the situation above where we knew the true parameters of the regression model we could calculate $\hat{Y}_h$ exactly so that its variance was zero and the grade point average of the new student then varied only because of σ^2.

The variance of $\hat{Y}_{h(new)}$ can be estimated by

$$\begin{aligned} s^2\{\hat{Y}_{h(new)}\} &= MSE + \sigma^2\{\hat{Y}_h\} \\ &= MSE + MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \\ &= MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]. \end{aligned} \tag{8.27}$$

The calculation of the confidence interval is now a routine matter, using the fact that

$$\frac{\hat{Y}_{h(new)} - \hat{Y}_h}{s^2\{\hat{Y}_{h(new)}\}}$$

is distributed according the t-distribution with degrees of freedom equal to $n - 2$. The resulting prediction interval is, of course, much wider than the confidence interval for $E\{\hat{Y}_h\}$ because the variance of $Y_{h(new)}$ contains an additional component consisting of the variance of Y_h around $E\{\hat{Y}_h\}$.

8.9 Inferences Concerning the Slope and Intercept Parameters

In most regression analysis in economics the primary objective is to estimate β_1. The regression slope b_1 is an efficient and unbiased estimate of that parameter. To obtain confidence intervals for β_1, however, and test hypotheses about it, we need the variance of the sampling distribution of b_1. This variance, it turns out, equals

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \tag{8.28}$$

which can be estimated by the statistic

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2}. \tag{8.29}$$

The confidence interval for β_1 can be obtained from the fact that

$$\frac{b_1 - \beta_1}{s\{b_1\}}$$

is distributed according to the t-distribution with $n-2$ degrees of freedom. As explained in Chapter 4, the t-distribution is symmetrical and flatter than the standard-normal distribution, becoming equivalent to that distribution as the degrees of freedom become large. The confidence intervals for β_1 with confidence coefficient α are then

$$b_1 \pm t(1-\alpha/2, n-2)\, s\{b_1\}.$$

where $t(1-\alpha/2, n-2)$ is the t-statistic associated with a cumulative probability of $(1-\alpha)$ when the degrees of freedom are $(n-2)$.

Now suppose that we want to test whether there is any relationship between Y and X. If there is no relationship, β_1 will be zero. Accordingly, we set the null hypothesis as

$$H_0\colon\ \beta_1 = 0$$

and the alternative hypothesis as

$$H_1\colon\ \beta_1 \neq 0.$$

Using our sample data we calculate the standardised test statistic

$$t^* = \frac{b_1}{s\{b_1\}},$$

which is distributed according to the t-distribution with $n-2$ degrees of freedom, and compare it with the critical values of t for the appropriate degree of α-risk from the table of t-values in the back of our statistics textbook. When the standardised test statistic is in the critical range—i.e., in the range for rejecting the null hypothesis—we say that β_1 is *significantly different from zero* at the $100\,\alpha$ percent level. Also we can calculate the P-value of the test statistic t, which equals the probability that a value of b_1 as different from zero as the one observed could have occurred on the basis of pure chance.

Frequently we want to test whether or not β_1 exceeds or falls short of some particular value, say β_1^o. This can be done by setting the null and alternative hypotheses as, for example,

$$H_0 : \beta_1 \leq \beta_1^o$$

and

$$H_1 : \beta_1 > \beta_1^o,$$

expressing the standardised test statistic as

$$t^* = \frac{b_1 - \beta_1^o}{s\{b_1\}},$$

and applying the critical values from the t-table for the appropriate level of α. When the standardised test statistic is in the critical range we can say that β_1 is significantly greater than β_1^o at the $100\,\alpha$ percent level.

Occasionally, inferences concerning the intercept parameter β_0 are also of interest. The regression intercept coefficient b_0 is an unbiased and efficient estimator of β_0. To obtain confidence intervals and conduct hypotheses tests we need an estimator of the sampling variance $\sigma^2\{b_0\}$. It turns out that $b_0 = \hat{Y}_h$ where $X_h = 0$ so we can use the estimator

$$\begin{aligned} s^2\{\hat{Y}_h\} &= MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \\ = \; s^2\{b_0\} &= MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]. \end{aligned} \tag{8.30}$$

Statistical tests can now be undertaken and confidence intervals calculated using the statistic

$$\frac{b_0 - \beta_0}{s\{b_0\}}$$

which is distributed as t(n - 2).

It turns out that these tests are quite robust—that is, the actual α-risk and confidence coefficient remain close to their specified values even when the error terms in the regression model are not exactly normally distributed as long as the departure from normality is not too great.

8.10 Evaluation of the Aptness of the Model

It must now be reemphasized that the application of this regression model to practical problems involves some very critical assumptions—namely, that the true residuals are independently normally distributed with zero mean and constant variance. We can never be sure in advance that in any particular application these assumptions will be close enough to the truth to make our application of the model valid. A basic approach to investigating the aptness or applicability of the model to a particular situation is to *analyse the residuals* from the regression—$e_i = Y_i - \hat{Y}_i$.

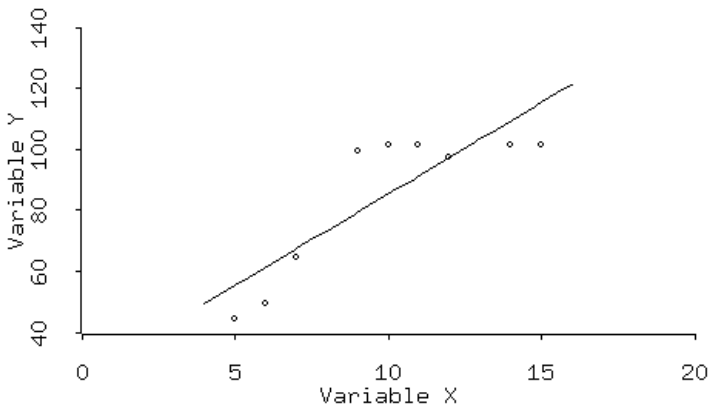


Figure 8.6: The actual and fitted values for a particular regression.

A number of important departures from the regression model may occur. First, the regression function we are trying to estimate may not be linear. We can get a good sense of whether or not this may be a problem by plotting the actual and predicted values of Y against the independent variable X, as is done in Figure 8.6, or plotting the residuals against the predicted values of Y as is done for the same regression in Figure 8.7. When the true relationship between the variables is linear the residuals will scatter

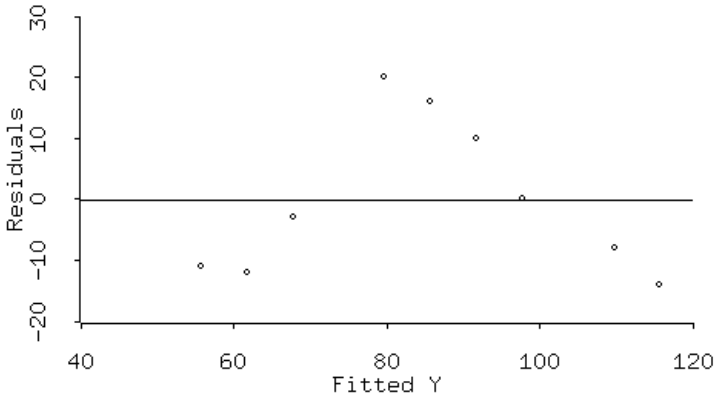


Figure 8.7: The residuals of the regression in Figure 8.6 plotted against the fitted values.

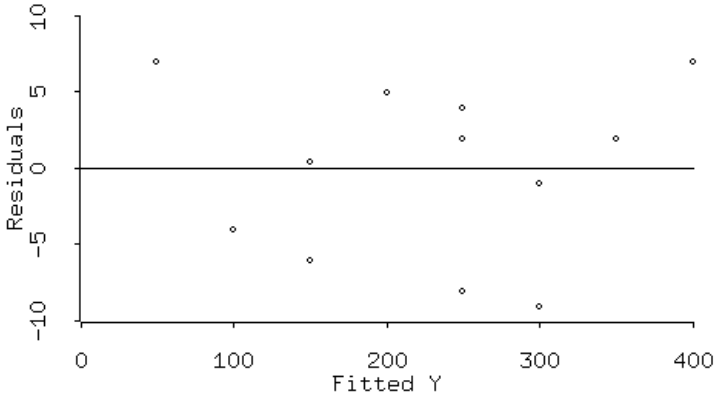


Figure 8.8: Well-behaved regression residuals plotted against the fitted values.

at random around the fitted straight line or around the zero line when plotted against the predicted values of the dependent variable. Obviously, the underlying functional relationship in Figure 8.6 is non-linear. An example of well-behaved residuals is given in Figure 8.8.

A second problem is that the variance of the e_i may not be constant

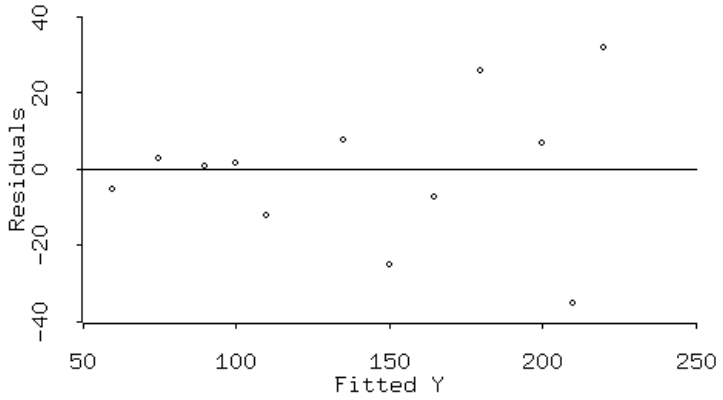


Figure 8.9: An example of heteroscedasticity—regression residuals plotted against the fitted values.

with respect to $\hat{Y}$ but may vary systematically with it. This problem is called *heteroscedasticity*. This is illustrated in Figure 8.9 where the residuals obviously increase as the predicted value of Y becomes larger.

Third, there may be lack of normality in the error terms. One way of checking the error term for normality is to standardise it by dividing it by its standard deviation—the square root of *MSE*—and checking to see whether approximately 2/3 of the errors lie within one standard deviation of zero. Alternatively, we could apply the chi-square test for normality developed in the previous chapter. Less formally, we can compare the observed frequencies of the standardised errors with the theoretically expected frequencies.

Finally, the errors may not be independent of each other. This happens frequently in time-series analysis where there is *autocorrelation* or *serial correlation* in the residuals—when the residual associated with one value of X or its predicted value of Y is high, the residual associated with the adjacent values of X or Y will also be high. This problem is discussed in detail in the next chapter.

To get around these problems it is sometimes useful to transform the variables. The residuals from estimating $Y = \beta_0 + \beta_1 X$ may be heteroscedastic, but those from estimating $log(Y) = \beta_0 + \beta_1 X$ may not be. Similarly, the relationship between $log(X)$ and $log(Y)$, or $1/X$ and $1/Y$, may be linear even though the relationship between X and Y may not be. Sometimes the residuals from the regression may not be well-behaved because, in truth, Y

depends on two variables X and Z instead of just X. By leaving Z out of the model, we are attempting to force the single variable X to explain more than it is capable of, resulting in deviations of the predicted from the actual levels of Y that reflect the influence of the absent variable Z.

8.11 Randomness of the Independent Variable

In some regression analyses it is more reasonable to treat both X and Y as random variables instead of taking the X_i as fixed from sample to sample. When X is random, the distribution of Y at a given level of X is a conditional distribution with a conditional mean and a conditional variance (i.e., conditional upon the level of X). In this case all of the results presented above for the regression model with X fixed continue to apply as long as

a) the conditional distribution of Y is normal with conditional mean $\beta_0 + \beta_1 X$ and conditional variance σ^2, and

b) the X_i are independent random variables whose probability distribution does not depend on the parameters β_0, β_1 and σ^2.

The interpretations of confidence intervals and risks of errors now refer to repeated sampling where *both* the X and Y variables change from one sample to the next. For example the confidence coefficient would now refer to the proportion of times that the interval brackets the true parameter when a large number of repeated samples of n pairs (X_i, Y_i) are taken and the confidence interval is calculated for each sample. Also, when both X and Y are random variables the correlation coefficient r is an estimator of the population correlation coefficient ρ rather than only a descriptive measure of the degree of linear relation between X and Y. And a test for $\beta_1 = 0$ is now equivalent to a test of whether or not X and Y are uncorrelated random variables.

8.12 An Example

During the first part of this century classical economics held that the real quantity of money demanded tends to be a constant fraction of real income—that is

$$\frac{M}{P} = k\, R_Y \tag{8.31}$$

where M is the nominal quantity of money held by the public, P is the general price level, R_Y is real national income and k is a constant, sometimes called the *Cambridge-k*. We want to use some data on nominal money holdings, nominal income and the consumer price index for Canada to test this idea. The data are presented in the worksheet below.

WORKSHEET FOR REGRESSION ANALYSIS OF CANADIAN DEMAND FOR MONEY

DATE	MON	GDP	CPI	RMON	RGDP	D-RMON	D-RGDP	Col. (6) Sq.	Col. (7) Sq.	(6) X (7)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1957	5.07	34.47	88.65	5.72	38.88	-7.90	-57.34	62.45	3288.30	453.17
1958	5.55	35.69	90.88	6.11	39.27	-7.51	-56.95	56.47	3243.81	428.00
1959	5.66	37.88	91.82	6.16	41.25	-7.46	-54.97	55.65	3021.77	410.09
1960	5.75	39.45	92.99	6.18	42.42	-7.44	-53.80	55.33	2894.42	400.18
1961	6.31	40.89	93.93	6.71	43.53	-6.91	-52.70	47.74	2776.85	364.09
1962	6.67	44.41	94.99	7.02	46.75	-6.60	-49.47	43.62	2447.38	326.75
1963	7.17	47.68	96.51	7.42	49.40	-6.20	-46.82	38.41	2192.23	290.18
1964	7.72	52.19	98.27	7.85	53.11	-5.77	-43.11	33.27	1858.56	248.68
1965	8.98	57.53	100.73	8.92	57.11	-4.70	-39.12	22.13	1530.04	184.02
1966	9.71	64.39	104.49	9.29	61.62	-4.33	-34.60	18.78	1197.08	149.93
1967	12.33	69.06	108.24	11.39	63.81	-2.23	-32.42	4.96	1050.80	72.22
1968	15.78	75.42	112.81	13.98	66.85	0.36	-29.37	0.13	862.57	-10.63
1969	15.40	83.03	117.74	13.08	70.52	-0.54	-25.70	0.29	660.62	13.87
1970	14.92	89.12	121.72	12.26	73.21	-1.36	-23.01	1.86	529.53	31.40
1971	16.52	97.29	125.12	13.20	77.75	-0.42	-18.47	0.18	341.06	7.81
1972	18.54	108.63	131.11	14.14	82.86	0.52	-13.37	0.27	178.68	-6.90
1973	20.61	127.37	141.07	14.61	90.29	0.99	-5.94	0.98	35.23	-5.87
1974	21.62	152.11	156.44	13.82	97.24	0.20	1.01	0.04	1.03	0.20
1975	24.06	171.54	173.44	13.87	98.91	0.25	2.68	0.06	7.20	0.67
1976	25.37	197.93	186.34	13.62	106.22	-0.01	10.00	0.00	99.92	-0.07
1977	27.44	217.88	201.35	13.63	108.21	0.00	11.99	0.00	143.71	0.05
1978	29.69	241.61	219.17	13.55	110.23	-0.08	14.01	0.01	196.35	-1.07
1979	30.97	276.10	239.23	12.94	115.41	-0.68	19.19	0.46	368.24	-12.99
1980	32.25	309.89	263.73	12.23	117.50	-1.40	21.28	1.95	452.77	-29.69
1981	33.64	356.00	296.57	11.34	120.04	-2.28	23.82	5.20	567.16	-54.31
1982	36.64	374.44	328.58	11.15	113.96	-2.47	17.73	6.11	314.47	-43.83
1983	42.32	405.72	347.58	12.17	116.73	-1.45	20.50	2.10	420.40	-29.68
1984	47.42	444.74	362.71	13.07	122.62	-0.55	26.39	0.30	696.61	-14.48
1985	62.25	477.99	377.13	16.51	126.74	2.88	30.52	8.32	931.49	88.04
1986	74.38	505.67	392.73	18.94	128.76	5.32	32.53	28.27	1058.53	172.99
1987	83.87	551.60	409.97	20.46	134.55	6.84	38.32	46.72	1468.75	261.97
1988	87.81	605.91	426.39	20.59	142.10	6.97	45.88	48.62	2105.07	319.93
1989	91.45	650.75	447.73	20.42	145.34	6.80	49.12	46.28	2412.99	334.19
1990	92.26	669.51	468.95	19.67	142.77	6.05	46.55	36.61	2166.45	281.64
1991	97.88	676.48	495.46	19.76	136.54	6.13	40.31	37.63	1625.23	247.30
1992	102.79	690.12	502.84	20.44	137.24	6.82	41.02	46.51	1682.78	279.77
1993	108.98	712.86	512.11	21.28	139.20	7.66	42.98	58.66	1847.10	329.17
1994	118.83	747.26	513.05	23.16	145.65	9.54	49.43	90.99	2443.32	471.52
1995	128.83	776.30	524.19	24.58	148.10	10.96	51.87	120.01	2690.85	568.28
SUM	1583.36	11316.84	9656.76	531.24	3752.67	-0.00	-0.00	1027.40	51809.36	6526.58
MEAN	40.60	290.18	247.61	13.62	96.22	-0.00	-0.00			

Columns (1) and (2) of the worksheet give the Canadian nominal money supply and Canadian nominal Gross Domestic Product (GDP) in billions of

current dollars. Gross Domestic Product is a measure of aggregate nominal income produced in the domestic economy. Column (3) gives the Canadian Consumer Price Index (CPI) on a base of 1963-66 = 100. The theory specifies a relationship between real money holdings and real income. Accordingly, real money holdings and real GDP are calculated in columns (4) and (5) by dividing the nominal values of these variables by the CPI and then multiplying by 100. Thus RMON and RGDP measure the Canadian real money supply and Canadian real GDP in constant 1963-66 dollars. So equation (8.31) above specifies that the numbers in column (4) should be a constant fraction of the numbers in column (5) plus a random error. So we want to run the following simple linear regression:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon \tag{8.32}$$

where Y_t is RMON (column (4)) and X_t is RGDP (column (5)). Because the observations occur through time we designate them by subscript t rather than subscript i.

To obtain a fitted line to these data we perform the calculations shown in columns (6) through (10). The columns D-RMON and D-RGDP give the deviations of RMON and RGDP from their respective means, 13.62 and 96.22, calculated at the bottom of columns (4) and (5). Column (8) gives D-RMON squared and column (9) gives D-RGDP squared. The sums at the bottom of these columns thus give

$$\sum_{t=1957}^{1995} (Y_t - \bar{Y})^2 = 1027.40$$

and

$$\sum_{t=1957}^{1995} (X_t - \bar{X})^2 = 51809.36$$

respectively. Column (10) gives the product of D-RMON and D-RGDP and the sum at the bottom gives

$$\sum_{t=1957}^{1995} (Y_t - \bar{Y})(X_t - \bar{X}) = 6526.58.$$

The estimate b_1 of β_1 can thus be calculated as

$$b_1 = \frac{\sum_{t=1957}^{1995}(Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1957}^{1995}(X_t - \bar{X})^2} = \frac{6526.58}{51809.36} = .126$$

and the estimate b_0 of β_0 becomes

$$b_0 = \bar{Y} - b_1\bar{X} = 13.62 - (.126)(96.22) = 1.5.$$

Next we need the R^2. This equals the square of

$$r = \frac{\sum_{t=1957}^{1995}(Y_t - \bar{Y})(X_t - \bar{X})}{\sqrt{\sum_{t=1957}^{1995}(X_t - \bar{X})^2}\sqrt{\sum_{t=1957}^{1995}(Y_t - \bar{Y})^2}} = \frac{6526.58}{\sqrt{51809.36}\sqrt{1027.40}} = .8946,$$

or $R^2 = (.8946)^2 = .8$. This means that the sum of squares explained by the regression is

$$SSR = R^2 \sum_{t=1957}^{1995} (Y_t - \bar{Y})^2 = (.8)(1027.40) = 821.92$$

and the sum of squared errors is

$$SSE = (1 - R^2) \sum_{t=1957}^{1995} (Y_t - \bar{Y})^2 = (.2)(1027.40) = 205.48.$$

The mean square error is then

$$MSE = \frac{SSE}{n-2} = \frac{205.48}{37} = 5.55.$$

To test whether there is a statistically significant relationship between real money holdings and real GNP we form a t statistic by dividing b_1 by its standard deviation. The latter equals

$$s\{b_1\} = \sqrt{\frac{MSE}{\sum_{t=1957}^{1995}(X_t - \bar{X})^2}} = \sqrt{\frac{5.55}{51809.36}} = .01035.$$

The t-statistic for the test of the null hypothesis that $\beta_1 = 0$ thus equals

$$t^* = \frac{b_1 - 0}{s\{b_1\}} = \frac{.126}{.01035} = 12.17.$$

Since this exceeds the critical value of t of 3.325 for $\alpha = .01$, the null-hypothesis of no relation between real money holdings and real income must be rejected.

To test the null-hypothesis that the constant term β_0 equals zero we obtain the standard deviation of b_0 from (8.30),

$$s^2\{b_0\} = MSE \left[\frac{1}{n} + \frac{\bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

which yields

$$s\{b_0\} = \sqrt{5.55\left[\frac{1}{39} + \frac{96.22^2}{51809.36}\right]} = \sqrt{(5.55)\left[.02564 + \frac{9258.29}{51809.36}\right]}$$

$$= \sqrt{(5.55)(.02564 + .1787)} = 1.064935.$$

The t-statistic for the test of the null hypothesis that $\beta_0 = 0$ is thus

$$t^* = \frac{b_0 - 0}{s\{b_0\}} = \frac{1.5}{1.064935} = 1.409$$

for which the P-value for a two-tailed test is .1672. We cannot reject the null hypothesis that β_0 equals zero at a reasonable significance level.

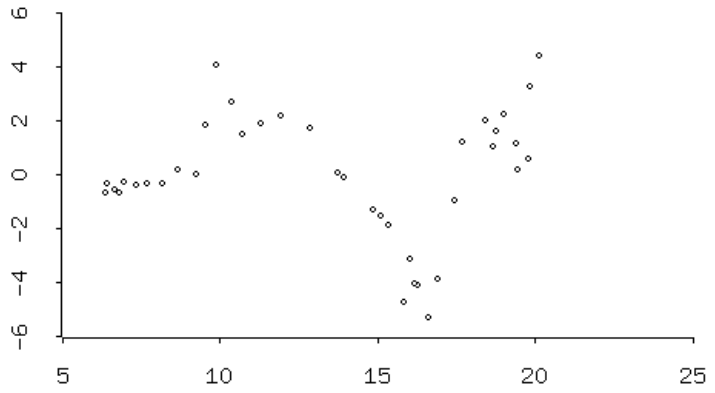


Figure 8.10: The residuals of the regression of Canadian real money holdings on Canadian real GDP, plotted against the fitted values.

The classical hypothesis that the public's real money holdings tend to be a constant fraction of their real income cannot be rejected on the basis of the data used here, because we cannot reject the hypothesis that the true relationship between RMON and RGNP is a straight line passing through the origin. Nevertheless, we must be open to the possibility that the ratio of RMON to RGNP, though perhaps independent of the level of real income, could depend on other variables not in the regression, such as the rate of interest (which equals the opportunity cost of holding money instead of

interest-bearing assets). If this were the case, we might expect the residuals from the regression to be poorly behaved. Figure (8.10) plots the residuals against the fitted values. The residuals are clearly not randomly scattered about zero. It is useful to check for serial correlation in these residuals by plotting them against time. This is done in Figure (8.11). There is obvious serial correlation in the residuals from the regression. We will address this problem again in the next chapter when we investigate the Canadian demand function for money using multiple regression.

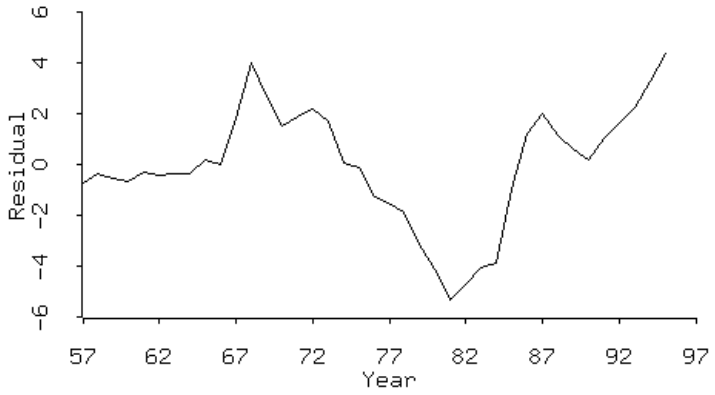


Figure 8.11: The residuals of the regression of Canadian real money holdings on Canadian real GDP, plotted against time.

8.13 Exercises

1. The following data relate to the model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where the X_i are assumed non-stochastic and the ϵ_i are assumed to be independently identically normally distributed with zero mean and constant variance.

i	Y_i	X_i
1	21	10
2	18	9
3	17	8
4	24	11
5	20	11
6	20	10
7	22	12
8	21	11
9	17	9
10	20	9

a) Calculate the regression estimates of α and β. (5.71, 1.43)

b) Calculate a 95% confidence interval for β. (0.56, 2.27)

2. Insect flight ability can be measured in a laboratory by attaching the insect to a nearly frictionless rotating arm by means of a very thin wire. The "tethered" insect then files in circles until exhausted. The non-stop distance flown can easily be calculated from the number of revolutions made by the arm. Shown below are measurements of this sort made on *Culex tarsalis* mosquitos of four different ages. The response variable is the average (tethered) distance flown until exhaustion for 40 females of the species.

Age, X_i (weeks)	Distance Flown, Y_i (thousands of meters)
1	12.6
2	11.6
3	6.8
4	9.2

Estimate α and β and test the hypothesis that distance flown depends upon age. Use a two-sided alternative and the 0.05 level of significance.

3. A random sample of size $n = 5$ is to be used to estimate the values of the unknown parameters of the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the random error term ϵ_i is $N(0, \sigma^2)$. The sample values for (X_i, Y_i) are

X_i	Y_i
-2	-6
-1	-2
0	-2
1	4
2	6

a) Compute the values of the least-squares estimators for β_0 and β_1.

b) Compute the value of the least-squares estimator for σ^2 and the coefficient of determination, R^2.

c) Conduct a test of the null hypothesis H_0: $\beta_1 \leq 2.0$ versus the alternative hypothesis H_1: $\beta_1 > 2.0$ using $\alpha = .05$ and find the approximate P-value for the test.

d) Compute a 95% confidence interval for the expected value of Y when $X = 5$.

4. The District Medical Care Commission wants to find out whether the total expenditures per hospital bed for a particular item tends to vary with the number of beds in the hospital. Accordingly they collected data on number of beds for the 10 hospitals in the district (Y_i) and the total expenditures per hospital bed (X_i). Some simple calculations yielded the following magnitudes:

$$\bar{Y} = 333.0 \quad \bar{X} = 273.4$$

$$\sum_{i=1}^{10}(Y_i - \bar{Y})^2 = 10756.0$$

$$\sum_{i=1}^{10}(X_i - \bar{X})^2 = 301748.4$$

$$\sum_{i=1}^{10}(X_i - \bar{X})(Y_i - \bar{Y}) = -37498$$

Use simple regression analysis to analyse the effect of number of beds on cost of the item per bed. Can you conclude that there is a relationship between these two variables. Is that relationship positive or negative? Calculate the R^2 and the significance of the regression coefficients. Is the overall relationship between the number of hospitals in a district and total expenditures per hospital bed statistically significant at reasonable levels of α-risk?

Chapter 9

Multiple Regression

While simple regression analysis is useful for many purposes, the assumption that the dependent variable Y depends on only one independent variable is very restrictive. For example, if we want to develop a model to estimate the quantity of bread demanded we can expect the latter to depend, at the very minimum, on the price of bread, on the prices of at least some substitutes and on real income.

9.1 The Basic Model

The basic linear multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots\cdots + \beta_K X_{Ki} + \epsilon_i \qquad (9.1)$$

where $i = 1 \ldots n$ and the ϵ_i are independently normally distributed with mean zero and constant variance σ^2. Actually, we can often get away with less restrictive assumptions about the ϵ_i, namely

$$E\{\epsilon_i\} = 0$$

and

$$E\{\epsilon_i \epsilon_j\} = 0, \quad i \neq j$$
$$E\{\epsilon_i \epsilon_j\} = \sigma^2, \quad i = j.$$

This says that the ϵ_i must be independently distributed with constant variance but not necessarily normally distributed. Our problem is to estimate the parameters β_k, $k = 0 \ldots K$, and σ and to establish confidence intervals and conduct appropriate statistical tests with respect to these parameters.

The n-observations on the dependent variable and the K independent variables can be represented as follows:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \cdots\cdots + \beta_K X_{K1} + \epsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \cdots\cdots + \beta_K X_{K2} + \epsilon_2$$

$$Y_3 = \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \beta_3 X_{33} + \cdots\cdots + \beta_K X_{K3} + \epsilon_3$$

$$\cdots\cdots\cdots$$

$$\cdots\cdots\cdots$$

$$\cdots\cdots\cdots$$

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots\cdots + \beta_K X_{Kn} + \epsilon_n$$

This appears in matrix form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{31} & \cdots & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & X_{32} & \cdots & \cdots & X_{K2} \\ 1 & X_{13} & X_{23} & X_{33} & \cdots & \cdots & X_{K3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & X_{3n} & \cdots & \cdots & X_{Kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

and can be written

$$\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E} \tag{9.2}$$

where $\mathbf{Y}$ is an n by 1 column vector, $\mathbf{X}$ is an n by $K+1$ matrix (i.e., a matrix with n rows and $K+1$ columns), $\mathcal{B}$ is a $K+1$ by 1 column vector and $\mathcal{E}$ is an n by 1 column vector. The first column of the matrix $\mathbf{X}$ is a column of 1's.

9.2 Estimation of the Model

Our problem is now to choose an estimate of (9.2) of the form

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \tag{9.3}$$

where $\mathbf{b}$ is a $K+1$ by 1 column vector of point estimates of the vector $\mathcal{B}$ and $\mathbf{e}$ is an n by 1 column vector of residuals. According to the method of least squares we choose the vector $\mathbf{b}$ so as to minimize the sum of squared residuals which appears in matrix form as

$$\begin{bmatrix} \epsilon_1 & \epsilon_2 & \epsilon_3 & \cdots & \cdots & \cdots & \cdots & \epsilon_n \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or

$$\mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} e_i^2,$$

where $\mathbf{e}'$ is the transpose of $\mathbf{e}$ and thereby consists of a row vector containing the n errors e_i. This sum of squares can be further represented as

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) \\ &= (\mathbf{Y}' - \mathbf{b}'\mathbf{X}')(\mathbf{Y} - \mathbf{Xb}) \\ &= (\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{Xb} - \mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}) \\ &= (\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}) \end{aligned} \tag{9.4}$$

where the second line uses the facts that the transpose of the sum of two matrices (vectors) is the sum of the transposes and the transpose of the product of two matrices (vectors) is the product of the transposes in reverse order, and the fourth line uses the fact that $\mathbf{Y}'\mathbf{Xb}$ and $\mathbf{b}'\mathbf{X}'\mathbf{Y}$ are identical

scalars—this can be seen by noting that $\mathbf{Y}'\mathbf{X}\mathbf{b}$ is

$$\begin{bmatrix} Y_1 & Y_2 & Y_3 & \cdots & Y_n \end{bmatrix} \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_K X_{K1} \\ \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_K X_{K2} \\ \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \beta_3 X_{33} + \cdots + \beta_K X_{K3} \\ \vdots\vdots\vdots \\ \vdots\vdots\vdots \\ \vdots\vdots\vdots \\ \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_K X_{Kn} \end{bmatrix}$$

and $\mathbf{b}'\mathbf{X}'\mathbf{Y}$ is

$$\begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \cdots & \beta_K \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \cdots & \cdots & 1 \\ X_{11} & X_{12} & X_{13} & \cdots & \cdots & X_{1n} \\ X_{21} & X_{22} & X_{23} & \cdots & \cdots & X_{2n} \\ X_{31} & X_{32} & X_{33} & \cdots & \cdots & X_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{K1} & X_{K2} & X_{K3} & \cdots & \cdots & X_{Kn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}$$

We now differentiate this system with respect to the vector $\mathbf{b}$ and choose that value of the vector $\hat{\mathbf{b}}$ for which $\partial \mathbf{e}'\mathbf{e}/\partial \mathbf{b} = 0$. We thus obtain

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{Y} \tag{9.5}$$

which yields

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{9.6}$$

where $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of the matrix $\mathbf{X}'\mathbf{X}$.

The system of equations (9.5) is called the *least-squares normal equations*. In the case where there are only two independent variables plus a constant term (i.e., $K = 2$), these equations are

$$\begin{array}{ccccccc} n\,\hat{b}_0 & + & \hat{b}_1 \sum X_{1i} & + & \hat{b}_2 \sum X_{2i} & = & \sum Y_i \\ \hat{b}_0 \sum X_{1i} & + & \hat{b}_1 \sum X_{1i}^2 & + & \hat{b}_2 \sum X_{1i}X_{2i} & = & \sum X_{1i}Y_i \\ \hat{b}_0 \sum X_{2i} & + & \hat{b}_1 \sum X_{1i}X_{2i} & + & \hat{b}_2 \sum X_{2i}^2 & = & \sum X_{2i}Y_i \end{array}$$

The coefficients $\hat{b}_k$ can be obtained by actually calculating all of these sums of squares and cross products, substituting the resulting numbers into

the above system of equations, and solving that system simultaneously for the $\hat{b}_k$'s. Alternatively, the data can be expressed in matrix form (i.e., as a vector $\mathbf{Y}$ and matrix $\mathbf{X}$) and the vector $\hat{\mathbf{b}}$ obtained by applying equation (9.6) to $\mathbf{Y}$ and $\mathbf{X}$ using a standard computer linear algebra program.[1] The easiest way to obtain the $\hat{b}_k$, however, is to read the variables X_k and Y into one of the many standard statistical software packages and apply the linear-regression procedure contained in that package. This has the computer do everything—except determine what regression to run and interpret the results! Remember that a computer performs fast calculations but cannot do our thinking for us. It does exactly what it is told—whence the fundamental *gigo* principle, "garbage in → garbage out".[2]

Along with the vector of estimated regression coefficients, the standard statistical packages give the *standard deviations* (or *standard errors*) of these coefficients, the appropriate t-statistics and sometimes the P-values, the minimized sum of squared deviations of the dependent variable from the regression line, and the coefficient of determination or R^2.[3]

9.3 Confidence Intervals and Statistical Tests

To construct confidence intervals and perform statistical tests regarding the regression coefficients we need estimates of the standard deviations or standard errors of these coefficients. The matrix of variances and covariances of the regression coefficients (from which the standard statistical packages present their standard errors) is

$$\begin{bmatrix} Var\{b_0\} & Cov\{b_0b_1\} & Cov\{b_0b_2\} & \cdots\cdots & \cdots\cdots & Cov\{b_0b_K\} \\ Cov\{b_0b_1\} & Var\{b_1\} & Cov\{b_1b_2\} & \cdots\cdots & \cdots\cdots & Cov\{b_1b_K\} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ Cov\{b_0b_K\} & Cov\{b_1b_K\} & Cov\{b_2b_K\} & \cdots\cdots & \cdots\cdots & Var\{b_K\} \end{bmatrix}$$

$$= E\{(\hat{\mathbf{b}} - \mathcal{B})(\hat{\mathbf{b}} - \mathcal{B})'\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

[1] Such as, for example, MATLAB, MAPLE or OCTAVE. The first two of these are commercial programs while the latter one is freely available over the internet.

[2] Attention should also be paid to a second important principle of computing, *rtfm*. The first letter of this acronym stands for the word "read" and the last letter stands for the word "manual"!

[3] XlispStat has been used for most of the regression calculations, as well as the graphics, in this book

As in the case of simple regression the appropriate estimator for σ^2 is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{df} = MSE$$

where $df = n - K - 1$ is the degrees of freedom and

$$\mathbf{e}'\mathbf{e} = \sum e_i^2 = SSE$$

is the minimized sum of squared deviations of Y_i from the regression line. The degrees of freedom is $n-K-1$ because we are using the data to estimate $K + 1$ parameters (for K dependent variables plus a constant term). The sum of squares 'explained' by the independent variables is

$$SSR = (\mathbf{Y} - \bar{Y})'(\mathbf{Y} - \bar{Y}) - \mathbf{e}'\mathbf{e} = \sum (Y_i - \bar{Y})^2 - \sum e_i^2 = SSTO - SSE.$$

where $\bar{Y}$ is the mean value of the dependent variable–i.e., the mean of the elements of $\mathbf{Y}$. As in the case of simple linear regression, the fraction of the variation in the dependent variable explained by the independent variables—the R^2—is equal to

$$R^2 = 1 - \frac{SSE}{SSTO}.$$

Notice that the addition of new independent variables to the regression will always increase the R^2. To see this, think of an experiment whereby we keep adding independent variables until the total number of these variables plus the constant equals the total number of observations—this would yield an R^2 equal to unity. We can thus 'explain' more and more of the variation in the dependent variable by adding additional independent variables, paying little attention to whether the variables added are relevant determinants of the dependent variable. To obtain a more meaningful measure of how much of the variation in the dependent variable is being explained, the R^2 is frequently adjusted to compensate for the loss in the degrees of freedom associated with the inclusion of additional independent variables. This adjusted R^2, called the $\bar{R}^2$ is calculated according to the formula

$$\bar{R}^2 = 1 - \frac{n-1}{n-K-1}\frac{SSE}{SSTO}. \tag{9.7}$$

For $\bar{R}^2$ to rise as the result of the addition of another independent variable, the sum of squares of the residuals must fall sufficiently to compensate for the effect of the addition of that variable on the number of degrees of freedom.

The ratio of $(\hat{b}_k - \beta_k)$ to its standard deviation

$$t^* = \frac{\hat{b}_k - \beta_k}{\sqrt{Var\{\hat{b}_k\}}} \tag{9.8}$$

is distributed according to the t-distribution with degrees of freedom

$$df = n - K - 1.$$

The t-table at the back of any textbook in statistics can be used to establish critical values and confidence intervals. The t-values associated with the null hypothesis H_0: $\beta_k = 0$ are also given in most standard statistics computer packages. To test the null hypothesis that β_k takes a particular hypothesized value, or exceeds or falls short of a particular hypothesized value, we divide the difference between the estimated value $\hat{b}_k$ and the hypothesized value β_k by the standard error of $\hat{b}_k$, also given in most statistics computer packages. The P-values given by standard computer packages are the probabilities of observing values of the coefficients as different from zero (in either direction since the test is two-tailed) as are the respective estimated coefficients when the true value of the coefficient in question is zero. It should be noted here that, when conducting tests and setting up confidence intervals, the constant term is treated simply as another coefficient.

9.4 Testing for Significance of the Regression

Frequently we want to test whether the regression itself is significant—that is, whether the independent variables taken as a group explain any of the variation in the dependent variable. The R^2 measures this, but it is a point estimate which could be as high as it is simply because of sampling error. What we want to test is the null hypothesis

$$H_0\colon\ \beta_1 = \beta_2 = \beta_3 = \ldots\ldots = \beta_K = 0$$

against the alternative hypothesis that at least one of these coefficients is different from zero. Notice that this null-hypothesis does not require that β_0, the constant term, be zero—indeed, when there is no relationship between all K independent variables and Y_i, the constant term will be $\beta_0 = \bar{Y}$.

When we run the regression we choose the coefficients $\hat{b}_k$ that minimize the sum of squared residuals $\sum e_i^2$. If the independent variables do not contribute at all to explaining the variations in Y_i we would expect the

minimized sum of squared residuals to be the same as the sum of squared deviations of the Y_i about their mean, $\sum (Y_i - \bar{Y})^2$. That is, we would expect *SSE* to equal *SSTO*. To the extent that

$$\sum e_i^2 \leq \sum (Y_i - \bar{Y})^2$$

there is evidence that the independent variables included in the regression have some explanatory power. The trouble is, however, that *SSE* could be less than *SSTO* strictly as a result of sampling error. We must therefore test whether the observed excess of *SSTO* over *SSE* is bigger than could reasonably be expected to occur on the basis of sampling error alone.

We have already seen that a sum of squares of independently and identically distributed normal random variables divided by their variance is distributed as χ^2 with degrees of freedom equal to the number of independent squared normal deviations being summed. This means that

$$\frac{\sum e_i^2}{\sigma^2} = \chi^2(n - K - 1) \tag{9.9}$$

and

$$\frac{\sum (Y_i - \bar{Y})^2}{\sigma_y^2} = \chi^2(n - 1). \tag{9.10}$$

It can be shown (though we will not do it here) that the difference between two χ^2 variables is also distributed according to the χ^2 distribution, but with degrees of freedom equal to the difference between the degrees of freedom of the two χ^2 variables. This implies that

$$\frac{\sum e_i^2}{\sigma^2} - \frac{\sum (Y_i - \bar{Y})^2}{\sigma_y^2} = \frac{\sum e_i^2 - \sum (Y_i - \bar{Y})^2}{\sigma^2} = \chi^2(K). \tag{9.11}$$

Here $\sigma_y^2 = \sigma^2$ under the null hypothesis that adding the independent variables to the regression has no effect on the residual variance.

We have also learned earlier that the ratio of two independent χ^2 distributions divided by their respective degrees of freedom is distributed according to the F-distribution with two parameters equal to the number of degrees of freedom in the numerator and denominator respectively. Thus, using (9.9) and (9.11) we obtain

$$\frac{\sum (Y_i - \bar{Y})^2 - \sum e_i^2}{K} \div \frac{\sum e_i^2}{n - K - 1} = F(K, n - K - 1) \tag{9.12}$$

where the σ^2 variables in the denominators of (9.9) and (9.11) cancel out. If the independent variables contribute nothing to the explanation of the dependent variable we would expect $\sum (Y_i - \bar{Y})^2$ to approximately equal $\sum e_i^2$ and the calculated F-statistic to be close to zero. On the other hand, if the independent variables do explain some of the variation in the dependent variable the F-statistic will be substantially positive. The question then is whether the probability of observing a value of F as high as the one observed for this particular sample, given that the independent variables truly explain none of the variation in the dependent variable, is small enough that we can reject the null hypothesis of no effect. We choose a critical value of F based on the desired α-risk and reject the null hypothesis if the value of the F-statistic obtained from the sample exceeds this critical value.

Notice now that we can substitute

$$SSR = \sum (Y_i - \bar{Y})^2 - \sum e_i^2$$

and

$$SSE = \sum e_i^2$$

into (9.12) to obtain

$$\frac{n-K-1}{K}\frac{SSR}{SSE} = F(K, n-K-1) \tag{9.13}$$

which can be further simplified using the facts that

$$SSR = R^2 SSTO$$

and

$$SSE = (1-R^2)\, SSTO$$

to produce

$$\left[\frac{n-K-1}{K}\right]\left[\frac{R^2}{1-R^2}\right] = F(K, n-K-1). \tag{9.14}$$

We can thus calculate the F-statistic using the values for R^2, n and K without calculating the total sum of squares and the sum of squared errors.

The basic principle behind (9.12) can be generalized to test the significance of subsets of the β_k and of relationships between various β_k. The test of the significance of a regression involves a comparison of the residuals obtained from the regression and the residuals obtained from the same

regression with everything but the constant term omitted (i.e., with all coefficients but the constant term set equal to zero). We could test the joint significance of, say, two of the K independent variables, X_2 and X_3, by running the regression with these two variables omitted and comparing the residuals so obtained with the residuals from the regression with the two variables included. This is called a *test of restrictions*. The two restrictions in this example are $\beta_2 = 0$ and $\beta_3 = 0$. The null hypothesis is

$$H_0 \colon \beta_2 = \beta_3 = 0$$

against the alternative hypothesis that either β_2 or β_3 is non-zero. We call the sum of squared residuals from the regression that excludes X_2 and X_3 the *restricted residual sum of squares*, $\sum e_{iR}^2$, and the sum of squares of the residuals from the full regression the *unrestricted residual sum of squares*, $\sum e_i^2$. The question is then whether imposing the restrictions raises the residual sum of squares by a 'significant' amount—that is, by an amount which would have a probability less than α of occurring if the restrictions truly have no effect on the explanatory power of the regression. The relevant F-statistic is

$$\frac{\sum e_{iR}^2 - \sum e_i^2}{v} \div \frac{\sum e_i^2}{n-K-1} = F(v, n-K-1) \tag{9.15}$$

where v (= 2 in this example) is the number of restrictions imposed on the regression. If the resulting F-statistic is above the critical value we can reject the null hypothesis that two coefficients β_2 and β_3 are both equal to zero and accept the alternative hypothesis that at least one of them is non-zero.

The same approach can be used to test particular hypotheses about the relationship between two coefficients. Suppose we have reason to believe that β_3 should be the negative of β_2. We can test this single restriction by formulating the null hypothesis

$$H_0 \colon \beta_3 = -\beta_2$$

and testing it against the alternative hypothesis

$$H_1 \colon \beta_3 \neq -\beta_2.$$

The null hypothesis implies the regression model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots\cdots + \beta_K X_{Ki} + \epsilon_i \\ &= \beta_0 + \beta_1 X_{1i} - \beta_3 X_{2i} + \beta_3 X_{3i} + \cdots\cdots + \beta_K X_{Ki} + \epsilon_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_3 (X_{3i} - X_{2i}) + \cdots\cdots + \beta_K X_{Ki} + \epsilon_i. \end{aligned} \tag{9.16}$$

We therefore construct the new variable $(X_3 - X_2)$ and replace the two variables (X_2 and X_3) in the regression with it. The residuals from this new regression can be designated $\sum e_{iR}^2$ and inserted into (9.15) together with a value of v equal to 1, representing the single restriction, and a sample F-statistic so obtained. If this statistic exceeds the critical value of F for the appropriate degree of α-risk we reject the null hypothesis and conclude that β_3 is not equal to the negative of β_2.

9.5 Dummy Variables

The independent variables in a multiple regression need not be quantitative. For example, suppose we have some data on the salaries of managers in industry and their years of education and want to investigate whether individuals' years of education affect their salaries. We run a simple regression of salary on years of education for the data in question and obtain the following results (the standard errors of the coefficients are given in brackets and $\hat{\sigma}$ is a point estimate of σ):

Dependent Variable: Salary in \$000's

Constant	38.91	(12.88)
Years of Education	.064	(0.898)
R-Squared	.00036	
Standard Error ($\hat{\sigma}$)	8.97	
Number of Observations	8	
Degrees of Freedom	6	

The null hypothesis that years of education has a zero or negative effect on salary cannot be rejected at any reasonable level of significance given the test statistic

$$t^* = \frac{.064}{.898} = .071269.$$

When we plot the data and impose the fitted regression line on it we get the data points and the virtually horizontal regression line in Figure 9.1.

Upon examining the data, it turns out that all the data points above the nearly horizontal fitted line are for individuals who are sales managers and all the data points below the line are managers who are not in sales. Our regression should obviously contain a variable specifying whether or not the individual in the sample is a sales manager. This variable is a qualitative

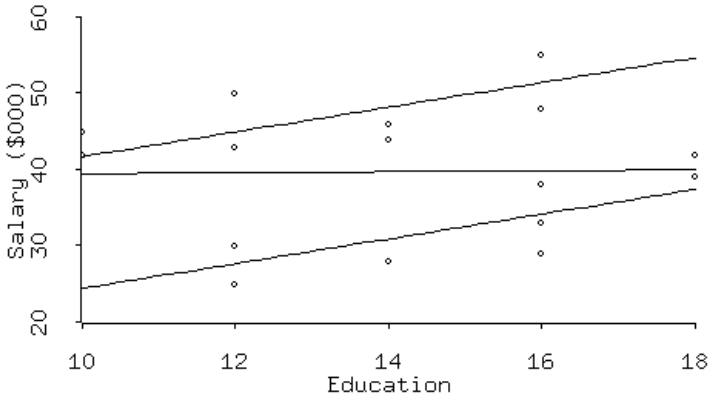


Figure 9.1: Plot and fitted lines of regression of salaries of sales managers on years of education (top line), other managers on years of education (bottom line) and all managers on years of education (middle line).

variable, usually referred to as a *dummy variable*. It consists entirely of zeros or ones—with the variable taking the value of 1 if the individual is a sales manager and 0 if the individual is not a sales manager.

Our regression model now takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon \tag{9.17}$$

where the variable X_1 is salary and X_2 is the dummy variable.

Consider the individual sample elements that do not represent sales managers. For these elements $X_{2i} = 0$ so the equation being fitted yields the predicted values

$$\hat{Y}_i \quad = \quad b_0 + b_1 X_{1i}. \tag{9.18}$$

For the individual sample elements that do represent sales managers, $X_{2i} = 1$ so the fitted equation becomes

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} = b_0 + b_1 X_{1i} + b_2$$

or

$$\hat{Y}_i \quad = \quad \tilde{b}_0 + b_1 X_{1i} \tag{9.19}$$

where $\tilde{b}_0 = b_0 + b_2$. Adding the dummy variable essentially allows the regression to have different constant terms for those managers who are sales managers and for those who are not sales managers. When we run this regression we get the following results:

Dependent Variable: Salary in $000's

Constant	8.254	(6.40)
Years of Education	1.62	(0.41)
Sales Manager Dummy	17.28	(2.05)
R-Squared	.845	
Standard Error ($\hat{\sigma}$)	3.66	
Number of Observations	16	
Degrees of Freedom	13	

Notice how the R^2 increases and the standard error of the regression falls when we add the dummy variable. Notice also that the test statistic for the null hypothesis that the true coefficient of the years-of-education variable is zero or less is now

$$t^* = \frac{1.62}{0.41} = 3.95$$

which has a P-value equal to .00083, so we can easily reject the null hypothesis at an α-risk of .001.

The predicted salary levels for each level of education for sales managers is given by the top upward-sloping line in Figure 9.1 and the predicted salary levels for each education level for non-sales managers is given by the lower upward-sloping line. These lines are very close to the fitted lines that would be obtained by running separate regressions for sales managers and for other managers.

We could include a second dummy variable to account for differences in the slope of the relationship between education and salary for the two groups of managers. This variable would be the product of the sales-manager dummy and the years of education—when the data element is a manager not in sales this variable would take a zero value and when the data element is a sales manager the variable would take a value equal to years of education. This dummy variable can be referred to as an *interaction* between years of education and whether the manager was sales vs. non-sales. The regression model would then be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon. \tag{9.20}$$

For data elements representing non-sales managers the predicted values will be

$$\hat{Y}_i = b_0 + b_1 X_{1i} \tag{9.21}$$

since both X_{2i} and X_{3i} will be zero for these elements. For data elements representing sales managers the predicted values will be

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 + b_3 X_{1i}$$

since for these elements $X_{3i} = X_{1i}$ and $X_{2i} = 1$, so we have

$$\hat{Y}_i = \tilde{b}_0 + \tilde{b}_1 X_{1i} \tag{9.22}$$

where $\tilde{b}_0 = b_0 + b_2$ and $\tilde{b}_1 = b_1 + b_3$.

The inclusion of dummies for both the constant term and the slope coefficient turns out to be equivalent to running two separate regressions—one for sales managers and one for other managers—except that by pooling the data and running a single regression with dummy variables included for the constant term and slope parameters we are imposing the assumption that the variance of the error term is the same in the separate regression models. Unless we have prior information about the variance of the errors there is no gain to pooling the data for the two types of managers in this case. When we include only a single dummy variable to allow, say, for differences in the constant term there is a gain from pooling the data and running a single regression provided we are prepared to force upon the model the assumption that the response of salary to years of education is the same for sales managers as for other managers. If we are not prepared to assume that the response of salary to education is the same for both groups we should run two regressions. It would still be appropriate to add two dummy variables, one for the constant term and one for the slope of salary with respect to education of sales vs. other managers, if we also have additional variables in the regression such as, for example, education of the individual manager's parents and race or religion. In this case, of course, the pooled regression will be appropriate only if we are willing to impose on the estimation the assumption that the effects of parents' education, race and religion are the same for sales managers and other managers.

9.6 Left-Out Variables

Frequently we do not have the data to include in a regression a variable that should be there. When this is the case we can often form an opinion, based on casual knowledge, about the effects of the coefficients of the included variables of leaving out a variable that should be in the regression. Suppose that the correct specification of the regression equation is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon \tag{9.23}$$

but we estimate

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon^* \tag{9.24}$$

instead.

Since in the case we are examining the regression actually estimated is a simple regression, our least-squares estimate of β_1 is

$$\hat{b}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{9.25}$$

From the true relationship we know that

$$\begin{aligned} Y_i - \bar{Y} &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon^* - \beta_0 - \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 - \bar{\epsilon}^* \\ &= \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + \epsilon^* \end{aligned} \tag{9.26}$$

Upon substitution of this equation into (9.25), the expected value for $\hat{b}_1$ becomes

$$\begin{aligned} E\{\hat{b}_1\} = \hat{\beta}_1 &= \frac{\beta_1 \sum_{i=1}^n (X_i - \bar{X})^2 + \beta_2 \sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_1 + \beta_2 \left[\frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]. \end{aligned} \tag{9.27}$$

The term in the big square brackets will be recognized as the slope coefficient of a regression of the variable X_2 on the variable X_1. Let us denote this coefficient by d_{21}. Then (9.27) becomes

$$\hat{\beta}_1 = \beta_1 + \beta_2 \, d_{21} \tag{9.28}$$

Suppose that the left-out variable is positively correlated with the included variable X_1 and positively related to the dependent variable. Then β_2 and d_{21} will both be positive and our least-squares estimate of β_1 will

be biased upward. If the left-out variable is negatively correlated with the included variable and positively related to the dependent variable, β_2 will be negative and d_{21} positive so our least-squares estimate of β_1 will be biased downward. If the left-out variable is negatively related to the dependent variable the bias will be upward when the left-out and included variables are negatively related and downward when the left-out and included variables are positively related.

9.7 Multicollinearity

Suppose a young researcher wants to estimate the demand function for money for Canada. She has learned in her intermediate macroeconomics class that the demand for real money holdings can be expressed

$$\frac{M}{P} = L(r_N, Y_R) \tag{9.29}$$

where M is the nominal money stock, P is the price level (so that M/P is the real money stock), r_N is the nominal interest rate and Y_R is the level of real income. This suggest a regression equation of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon \tag{9.30}$$

where Y is Canadian real money holdings, X_1 is the nominal interest rate and X_2 is Canadian real income. In the process of collecting her data, our researcher discovered two different measures of real income, GNP and GDP.[4] Not knowing which to use as her measure of real income, she did the easy thing and simply included both in the regression. Her regression model now becomes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon \tag{9.31}$$

where X_2 is real Canadian GDP and X_3 is real Canadian GNP. She used the Canadian 90-day commercial paper rate as a measure of the Canadian nominal interest rate.[5] All the data series were annual (as opposed to quarterly or monthly) for the years 1957 to 1996 inclusive.

[4]GDP or gross domestic product measures the level of aggregate real output produced by resources employed in the country while GNP or gross national product measures the level of aggregate real output produced by resources owned by domestic residents. To calculate GNP from GDP we have to subtract out that part of aggregate domestic output (GDP) produced by resources that are owned by foreigners and then add in the part of aggregate output abroad that is produced by resources owned by domestic residents.

[5]The 90-day commercial paper rate is the rate of interest charged on commercial paper—that is, on securities issued by major corporations for short-term borrowing—that becomes due 90 days after issue.

The researcher obtained the following regression results:

Dependent Variable: Canadian Real Money Holdings

Constant	8.50	(4.47)
90-Day Paper Rate	-2.65	(0.39)
Real GDP	-0.32	(0.50)
Real GNP	0.51	(0.53)
R-Squared	.91	
Standard Error ($\hat{\sigma}$)	6.75	
Number of Observations	40	
Degrees of Freedom	36	

Surprised that both real income coefficients were insignificant, the researcher decided to perform an F-test of the null hypothesis that both are simultaneously zero (H_0: $\beta_2 = \beta_3 = 0$). So she ran the same regression with both variables omitted, obtaining the following results:

Dependent Variable: Canadian Real Money Holdings

Constant	43.35	(8.60)
90-Day Paper Rate	1.46	(1.01)
R-Squared	.05	
Standard Error ($\hat{\sigma}$)	21.44	
Number of Observations	40	
Degrees of Freedom	38	

The mean squared errors for the respective regressions are equal to their sums of squared residuals divided by their respective degrees of freedom. Thus, the sum of squared residuals for the unrestricted regression (i.e., the one that included the two real income variables) is

$$\sum e_i^2 = df\hat{\sigma}^2 = (36)(6.75)^2 = 1640$$

and the sum of squared residuals for the restricted regression (the one that excluded the two real income variables) is

$$\sum e_{Ri}^2 = df\hat{\sigma}^2 = (38)(21.44)^2 = 17467$$

The appropriate test statistic is therefore

$$\frac{\sum e_{Ri}^2 - \sum e_i^2}{v} \div \frac{\sum e_i^2}{n-K-1} = \frac{17467-1640}{2} \div \frac{1640}{36}$$

$$= \frac{7913.5}{45.55} = 173.73 = F(v, n-K-1) = F(2,36)$$

where v is the number of restrictions, equal to 2 in this case. The critical value for $F(2,36)$ setting the α-risk at .01 is 5.18 so the researcher rejected the null hypothesis that both of the coefficients are zero.

What is happening here? Neither of the income variables is statistically significant in the regression but the two together are significant at far below the 1% level!

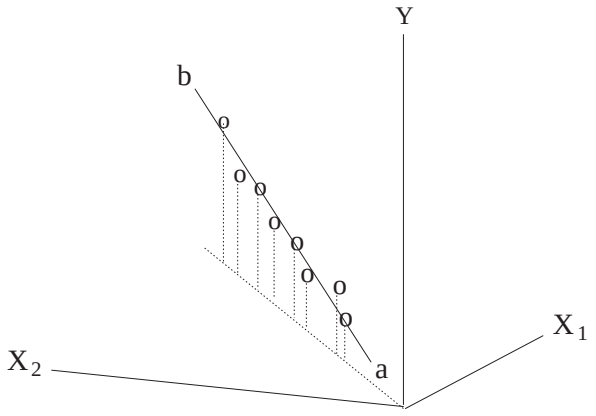


Figure 9.2: An illustration of multicolinearity of X_1 and X_2 in predicting Y.

This is an example of *multicollinearity*. The problem is that GDP and GNP are so highly correlated with each other that they are virtually the same variable. Had they been perfectly correlated, of course, the computer would not have been able to run the regression. Including two perfectly correlated variables in the regression is equivalent to including the same variable twice. This would mean that the $\mathbf{X}$ matrix would have two identical columns so that it would be non-singular and the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix would not exist. The problem here is that the two variables are not identical but

nevertheless highly correlated. This makes it impossible to determine their separate influences in the regression. The situation can be seen from Figure 9.2 for a multiple regression containing a constant term and two highly collinear independent variables X_1 and X_2. The purpose of the regression is to identify a plane in X_1, X_2, Y space that indicates how the dependent variable Y responds to changes in X_1 and X_2. When X_1 and X_2 are highly correlated, however, all the points lie very close to a ray projecting outward into X_1, X_2, Y space. It is possible to identify a relationship between X_1 and Y and between X_2 and Y but not between both X_1 and X_2 together and Y. Any estimated plane resting on the line ab in Figure 9.2 will be very unstable in the dimensions X_1, Y and X_2, Y—slightly different placements of the points in different samples will lead to planes with very different slopes in the X_1, Y and X_2, Y dimensions.

The researcher's solution to the problem in this case is easy—simply drop one of the income variables from the regression, since both are measuring the same thing, real income. Dropping real GDP, she obtains the following results:

Dependent Variable: Canadian Real Money Holdings

Constant	10.47	(3.21)
90-Day Paper Rate	-2.62	(0.38)
Real GNP	0.17	(0.01)
R-Squared	.91	
Standard Error ($\hat{\sigma}$)	6.70	
Number of Observations	40	
Degrees of Freedom	37	

Situations arise, however, in which two collinear variables really measure different things and we therefore want to identify their separate effects on the dependent variable. Suppose, for example, that we want to measure the effects of domestic and foreign real incomes and domestic relative to foreign prices on a country's balance of trade. The theoretical equation takes the form

$$B_T = B(Y_R^D, Y_R^F, P_R) \tag{9.32}$$

where Y_R^D is domestic real income, Y_R^F is foreign real income and P_R is the relative price of domestically produced goods in terms of foreign produced

goods with all prices measured in a common currency.[6] The appropriate regression model would be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon \tag{9.33}$$

where the dependent variable Y is the real balance of trade, X_1 is domestic real income Y_R^D, X_2 is foreign real income Y_R^F is foreign real income, and X_3 is the relative price of domestic goods, P_R. Since a rise in the relative price of domestic in terms of foreign goods will cause both domestic and foreign residents to switch their purchases away from domestic goods, increasing imports and reducing exports, we would expect the real balance of trade to be negatively affected, so the expected sign of β_3 is negative. An increase in domestic income might be expected to cause domestic residents to buy more foreign goods, increasing imports and reducing the real balance of trade. We would therefore expect β_1 to also be negative. An increase in foreign income, on the other hand, might be expected to cause foreigners to import more, resulting in an expansion of domestic exports and an increase in the balance of trade. The coefficient β_2 would thus expected to take on a positive sign.

When we estimate equation (9.33) for some country pairs we might find that the domestic and foreign real income variables are so highly collinear that our estimates of β_1 and β_2 will be statistically insignificant. If we drop one of the variables, the remaining real income variable acts as a measure of world real income and the response of the real balance of trade to that variable will measure the effect of a proportional rise in both domestic and foreign income on net domestic exports. Our purpose, however, is to measure the separate effects of the two income variables on the domestic real trade balance. There is no way that we can do this on the basis of the information provided by the data we are using. The only way to solve our problem is to obtain more information.

[6]The variable P_R is called the *real exchange rate*. The nominal exchange rate is the price of one country's money in terms of another country's money while the real exchange rate is the price of one country's output in terms of another country's output.

9.8 Serially Correlated Residuals

Perhaps the most important basic assumption of the linear regression model is that the errors ϵ_i are *independently* distributed. This means that the error associated with the i-th observation does not in any way depend on the error associated with the j-th observation. This assumption is frequently violated in regressions involving time series because the errors are correlated through time. As noted earlier, this situation is called *serial correlation* or *autocorrelation*. High (low) values at any point in time are associated with high (low) values in neighbouring points in time when there is positive autocorrelation.

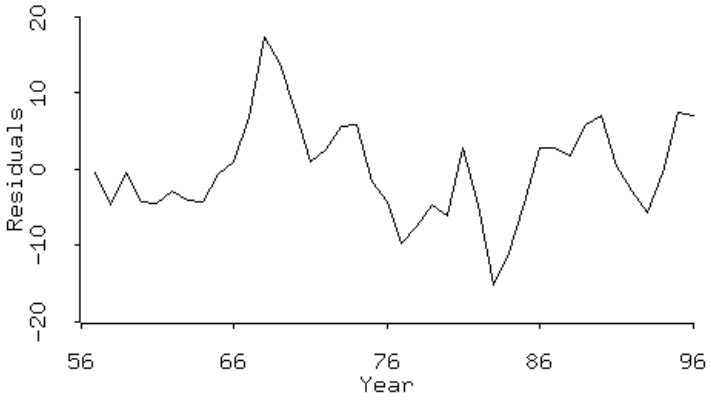


Figure 9.3: Residuals from the regression of Canadian real money holdings on the country's 90-day commercial paper rate and real GNP plotted against time.

Consider the regression of Canadian real money holdings on the 90-day commercial paper rate and real GNP reported above. The residuals from that regression are reported in Figure 9.3. It is clear from looking at the figure that these residuals are serially correlated—high values in one period are clearly associated with high values in immediately adjacent periods. To demonstrate formally that serial correlation is present, we can regress each year's residual on the residuals for several previous years. Using three lags, we obtain

Dependent Variable: Residual

Constant	.0544	(0.749122)
Residual-lag-1	1.0279	(0.171394)
Residual-lag-2	-0.4834	(0.233737)
Residual-lag-3	.0959	(0.175700)
R-Squared	.5849	
Standard Error ($\hat{\sigma}$)	4.5437	
Number of Observations	37	
Degrees of Freedom	33	

Statistically significant coefficients were obtained for one and two lags of the residuals—based on t-ratios of 6.0 and -2.06, respectively. The third lag is clearly insignificant. When the residual is correlated with the immediately previous residual, the serial correlation is called *first-order* serial correlation, when it is correlated with the residual two periods previous it is called *second-order* serial correlation, and so forth. In the above case, there is first- and second-order serial correlation in the residuals but not third-order. We do not know whether fourth-order serial correlation is present because we did not test for it—it is possible to have fourth- (or any other) order serial correlation in the residuals without having serial correlation of lower orders.

The standard procedure for detecting first-order (and only first-order) serial correlation in the residuals is to calculate the *Durbin-Watson Statistic*. This equals

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}. \tag{9.34}$$

The sampling distribution of d is a complex one. It turns out that d can take values between 0 and 4, and will differ from 2 when first-order serial correlation is present. When the first-order serial correlation is positive, d will be less than 2 and when it is negative d will be greater than 2. There is, however, a wide range of indeterminacy. In the case of positive serial correlation, one cannot clearly reject the null hypothesis of zero autocorrelation unless d is below the lower bound for the chosen level of α-risk in the table of critical values for the Durbin-Watson d statistic in the back of one's statistics textbook. And one can only accept the hypothesis of zero autocorrelation if d is above the upper bound in the table. For values of d between the lower and upper bounds we cannot draw any conclusion. For

negative serial correlation (which is present when $d > 2$) the same limits are used except we compare the numbers in the table with $4 - d$. In the regression above, the Durbin-Watson statistic is .58 which is well below the lower bound for $\alpha = .01$ and indicates positive first-order serial correlation.

What do we do when first-order serial correlation is present in the residuals? (Dealing with higher order serial correlation is beyond the technical level of the analysis here.) The answer to this question depends on why the autocorrelation is present. One possibility is that the true errors are serially correlated. This implies that the standard linear model is the incorrect one to apply to the data. An appropriate error term might be

$$\epsilon_t = \rho\,\epsilon_{t-1} + u_t$$

which implies that

$$\epsilon_t - \rho\,\epsilon_{t-1} = u_t,$$

where u_t is independently normally distributed with zero mean and variance σ^2. Assuming that the residuals actually behave in this way, we can lag the original regression equation

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t$$

once to yield

$$Y_{t-1} = \beta_0 + \beta_1 X_{1(t-1)} + \beta_2 X_{2(t-1)} + \epsilon_{t-1}$$

and then subtract ρ times the second equation from the first to obtain

$$Y_t - \rho\,Y_{t-1} = \beta_0 + \beta_1(X_{1t} - \rho\,X_{1(t-1)}) + \beta_2(X_{2t} - \rho\,X_{2(t-1)}) + u_t. \quad (9.35)$$

In this equation $(Y_t - \rho\,Y_{t-1})$, $(X_{1t} - \rho\,X_{1(t-1)})$ and $(X_{2t} - \rho\,X_{2(t-1)})$ are related according to the standard linear model with the independently and normally distributed error term u_t.

To estimate equation (9.35), we need an estimator of ρ. A natural way to proceed is to regress the residuals from the original regression on themselves lagged one period and use the slope coefficient as that estimator. Our regression model would be

$$e_t = \gamma + \rho\,e_{t-1} + \upsilon_t$$

where υ_t is an independent draw from the true constant-variance error term and we would expect our estimate of γ to be zero. The results from this regression are as follows:

Dependent Variable: Residual

Constant	0.140	(0.761913)
Residual-lagged	0.716	(0.118507)
R Squared:	0.497	
Standard Error ($\hat{\sigma}$)	4.7562	
Number of Observations	39	
Degrees of Freedom	37	

We can apply the resulting estimate of ρ (= 0.716) to obtain the new variables

$$\tilde{Y}_t = (Y_t - .716\, Y_{t-1})$$

$$\tilde{X}_{1t} = (X_{1t} - .716\, X_{1(t-1)})$$

and

$$\tilde{X}_{2t} = (X_{2t} - .716\, X_{2(t-1)}).$$

A new regression of the form

$$\tilde{Y}_t = \beta_0 + \beta_1 \tilde{X}_{1t} + \beta_2 \tilde{X}_{2t} + u_t$$

can then be run, yielding the following results:

Dependent Variable: Real Money Variable

Constant	1.03	(2.03)
Interest rate variable	-1.38	(0.33)
Real GNP variable	0.17	(0.02)
R-Squared	.73	
Standard Error ($\hat{\sigma}$)	4.05	
Number of Observations	39	
Degrees of Freedom	36	

It turns out that the effects of this 'correction' for serial correlation in the residuals, comparing the before and after regressions, reduces the absolute value of the slope coefficient of the interest rate variable from -2.62 to -1.38 and also reduces its standard error slightly. A sophisticated extension of this procedure is to regress the residuals of this new equation on themselves lagged and modify the estimate of ρ accordingly, doing this repeatedly until the estimates of ρ change by less than some minimal amount. When this is done, we obtain the following results:

Dependent Variable: Real Money Variable

Constant	-4.24	(24.62)
Interest rate variable	-1.09	(0.31)
Real GNP variable	0.18	(0.05)
ρ	0.928	(0.07)
R-Squared	.97	
Standard Error ($\hat{\sigma}$)	3.75	
Number of Observations	39	
Degrees of Freedom	35	

These refinements reduce further the absolute value of the slope coefficient of the interest rate variable and its standard error and raise slightly the coefficient of the real income variable and more substantially its standard error.

The 'optimal' value of ρ obtained by the above iterative method is very close to unity. In fact, a long-standing traditional approach to dealing with serial correlation in the residuals has been to take the first differences of the variables and run the regression in the form

$$Y_t - Y_{t-1} = \beta_0 + \beta_1(X_{1t} - X_{1(t-1)}) + \beta_2(X_{2t} - X_{2(t-1)}) + \vartheta_t. \qquad (9.36)$$

This assumes that

$$\vartheta_t = \epsilon_t - \epsilon_{t-1}$$

is independently and normally distributed and therefore that $\rho = 1$. When we impose this assumption on the residuals we obtain the following results:

Dependent Variable: Real Money Variable

Constant	0.496	(0.89)
Interest rate variable	-0.96	(0.32)
Real GNP variable	0.15	(0.06)
R-Squared	.22	
Standard Error ($\hat{\sigma}$)	3.73	
Number of Observations	39	
Degrees of Freedom	36	

The results differ little from those obtained when ρ was estimated iteratively.

Which coefficients are we to believe, those with no 'correction' of the residuals for serial correlation or those with a 'correction' imposed? To answer this question we must know the reason for the residuals being serially correlated. One possibility, of course, is that the residuals of the 'true' model are serially correlated. The problem with this explanation is that there is no reason in economic theory for the residuals to be serially correlated if we have correctly modeled the economic process we seek to explain. The reason why we have serial correlation in the residuals is that we have left variables that are correlated with time out of the model because we either could not measure them or could not correctly specify the underlying theory given the current state of knowledge. Obviously, the best approach is to try to better specify the model and to be sure that all variables that should be in it are included in the estimating equation. If we cannot do so our coefficients are likely to be biased for reasons outlined in section 9.6 above on left-out variables. Whether we improve things by correcting the residuals for first-order serial correlation is a question that econometricians will debate on a case-by-case basis. Clearly, however, it is inappropriate to routinely and unthinkingly impose a 'correction' on the residuals every time serial correlation is present.

9.9 Non-Linear and Interaction Models

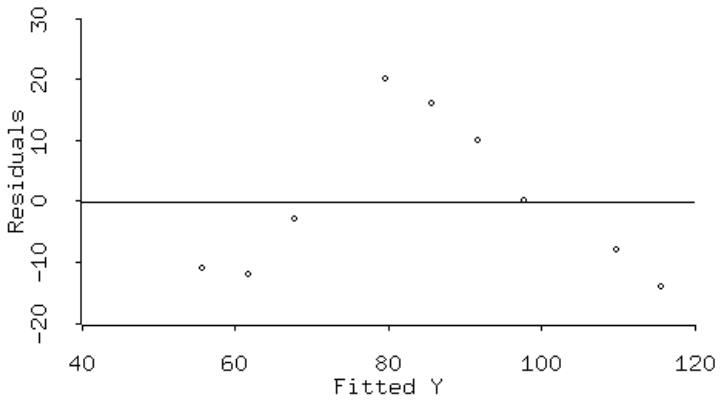


Figure 9.4: Residuals from a linear regression that suggest the underlying relationship is nonlinear.

It frequently arises that the residuals show a non-linear pattern as is illustrated in Figure 9.4. There are a number of simple ways of fitting non-linear relationships—either the dependent or independent variables or both can be transformed by inverting them or taking logarithms and using these non-linear transformations of the variables in a linear regression model. Another way is to include in the regression model squares of the independent variables along with their levels. For example, we might have

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \epsilon_t. \tag{9.37}$$

Interaction models arise when the relationship between the dependent variable and one of the independent variables depends on the level of a second independent variable. In this case, the appropriate regression model would be of the form

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{1t} X_{2t} + \epsilon_t. \tag{9.38}$$

Let us work through an example that illustrates both of these modifications to the standard linear model. It is quite common for colleges and universities to develop regression models for predicting the grade point averages (GPA's) of incoming freshmen. This evidence is subsequently used to decide which students to admit in future years. Two obvious variables that should be predictors of subsequent student performance are the verbal and mathematics scores on college entrance examinations. Data for a randomly-selected group of 40 freshmen were used to obtain the following regression results:

Dependent Variable: Freshman Grade Point Average

Constant	-1.570	(0.4937
Verbal Score (percentile)	0.026	(0.0040)
Math Score (percentile)	0.034	(0.0049)
R-Squared	.68	
Standard Error ($\hat{\sigma}$)	.402	
Number of Observations	40	
Degrees of Freedom	37	

These results indicate that students' scores on both the verbal and mathematical college entrance tests are significant positive predictors of freshman success (with t-ratios 6.3 and 6.8, respectively). An increase in a student's

verbal score by 10 percentiles will lead on average to a .26 increase in his/her GPA. For example, a student in the 70th percentile on both the verbal and mathematics sections of the entrance exam will have an expected freshman GPA of

$$-1.57 + (70)(.026) + (70)(.034) = 2.58.$$

An increase in her verbal score on the entrance exam from the 70th to the 80th percentile will increase her expected GPA by 0.26 to 2.84. And an increase in her math score from the 70th to the 80th percentile will increase her expected GPA by .34 to 2.92. An increase in both of her scores from the 70th to the 80th percentile will increase her expected GPA by .6 (= .26 + .34) to 3.18. The increase in expected GPA predicted by an increase in the percentiles achieved on the mathematics and verbal college entrance exams will be independent of the initial levels of the student's scores.

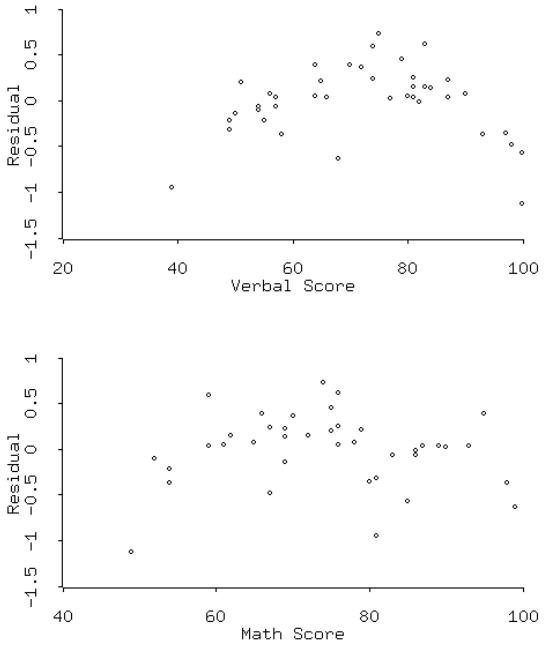


Figure 9.5: Residuals from first order regression model of grade point average on test scores.

The residuals from this regression are plotted against the two independent variables in Figure 9.5. Plotted against verbal score, they have an

inverse parabolic pattern, suggestive of non-linearity.[7] To check this out we run a second regression that includes the squared verbal and mathematics scores as additional variables together with an interactive variable consisting of the product of the verbal and math scores. The results are as follows:

Dependent Variable: Freshman Grade Point Average

Constant	-9.9167	(1.35441)
verbal score	0.1668	(0.02124)
math score	0.1376	(0.02673)
verbal score squared	-0.0011	(0.00011)
math score squared	-0.0008	(0.00016)
verb score x math score	0.0002	(0.00014)
R-Squared	.94	
Standard Error ($\hat{\sigma}$)	.187	
Number of Observations	40	
Degrees of Freedom	34	

As expected, the verbal score squared has a significant negative sign indicative of an inverse parabolic relationship (the t-statistic equals -10). The squared mathematical score is also statistically significant with a negative sign (the t-ratio equals -5). The interactive term, verbal score times math score, is not statistically significant, with a t statistic of only 1.43. The residuals from this extended regression, plotted in Figure 9.6 are very well behaved. An F-test of the null hypothesis of no effect of the squared and interactive terms yields the statistic

$$\frac{\sum e_{iR}^2 - \sum e_i^2}{3} \div \frac{\sum e_i^2}{34} = \frac{(37)(.402)^2 - (34)(.187)^2}{3} \div \frac{(34)(.187)^2}{34}$$

$$= \frac{5.98 - 1.19}{3} \div \frac{1.19}{34} = \frac{1.60}{.035} = 45.71 = F(3, 34).$$

We can reject the null hypothesis at any reasonable level of α-risk.

Notice how the addition of these second order terms (squares and cross-products) affects the response of GPA to verbal and mathematical test

[7] A parabola takes the mathematical form

$$y = ax^2 - bx - c.$$

When $a < 0$ the parabola will be inverted with the arms extending downward.

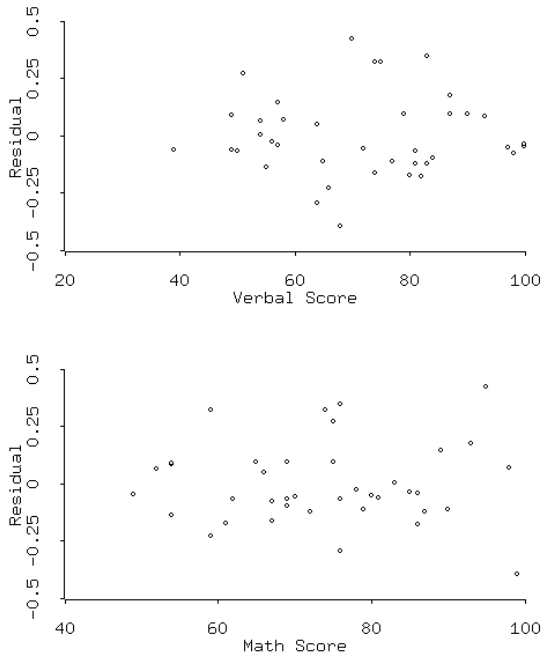


Figure 9.6: Residuals from second order regression model of grade point average on test scores.

scores. A student with scores in the 70th percentile on both the verbal and mathematical tests will have a predicted GPA of

$$-9.9167 + (.1668)(70) + (.1376)(70) - (.0011)(70)^2$$

$$-(.0008)(70)^2 + (.0002)(70)(70) = 3.04$$

which is higher than the predicted value from the regression that did not include the second order terms. Now suppose that the student's verbal test score increases to the 80th percentile. This will increase his expected GPA by

$$(.1668)(80-70) - (.0011)(80^2 - 70^2) + (.0002)(80-70)(70) = .158$$

to 3.198. An increase in the mathematical score from the 70th to the 80th percentile with his verbal score unchanged would increase his expected GPA by

$$(.1376)(80-70) - (.0008)(80^2 - 70^2) + (.0002)(70)(80-70) = .316$$

to 3.356. Given the interaction term, an increase in both the verbal and mathematical scores of the student from the 70th to the 80th percentile would increase his expected GPA by more than the sum of the two separate effects above (= .158 + .316 = .474). The increase would be

$$(.1668)(80-70) - (.0011)(80^2 - 70^2) + (.1376)(80-70) - (.0008)(80^2 - 70^2)$$

$$+(.0002)[(80-70)(70) + (70)(80-70) + (80-70)(80-70)$$

$$= .158 + .316 + (.0002)(100) = .158 + .316 + .02 = .494$$

to 3.534. Notice the difference in the levels and predicted changes in the GPA's under the second order as opposed to the first order model. Given that the interaction term is statistically insignificant, however, we might decide to make our predictions on the basis of a regression model that includes the squared terms but excludes the interaction term.

9.10 Prediction Outside the Experimental Region: Forecasting

A major purpose of regression analysis is to make predictions. Problems arise, however, when the fitted models are used to make predictions outside the range of the sample from which the regression model was estimated—i.e., outside the experimental region. The fit within sample is based on the surrounding sample points. Outside the range of the sample there is no opportunity for the fitted regression parameters to be influenced by sample observations—we simply do not know what values of the dependent variable would be associated with levels of the independent variables in this range were they to occur. As a result, the farther outside the sample range we extrapolate using the estimated model the more inaccurate we can expect those predictions to be.

Predicting outside the sample range in time series regressions is called *forecasting*. We have data on, say, the consumer price index, up to and including the current year and want to predict the level of the consumer price index next year. We develop a regression model 'explaining' past movements in the consumer price index through time and then use that model to forecast the level of the consumer price index in future periods beyond the sample used to estimate the model. To the extent that we use independent variables other than time we have to forecast the levels of those variables because their realization has not yet occurred. Errors in those forecasts will produce errors in predicting the future values of the dependent variable. These will additional to the errors that will result because we are using the regression parameters to predict values of the dependent variable outside the sample range in which those parameters were estimated.

Alternatively, we could forecast the consumer price index based on a simple regression of a range of its previous realized values against time using a model such as

$$Y_T = \beta_0 + \beta_1 T + \epsilon_t$$

where Y_T is the consumer price index at time T. This is the simplest *time-series* model we could fit to the data—time-series econometricians typically use much more sophisticated ones. The regression model is estimated for the period $T = 1, 2, \ldots N$ and then a prediction of Y for period $N + 1$ is obtained as

$$Y_{N+1} = b_0 + b_1 (N + 1).$$

Obviously, if the time-period $N + 1$ could have been used in the estimation of the model, the estimates b_0 and b_1 would be different. The further we forecast beyond period N the less the expected accuracy of our forecasts.

9.11 Exercises

1. A random sample of size $n = 20$ families is used to conduct a multiple regression analysis of how family i's annual savings S_i depends on its annual income I_i and its home-ownership status H_i. Both S_i and I_i are measured in thousands of dollars. Variable H_i is equal to 1 if family i owns its home and equal to 0 if family i rents. The regression results

Coefficient	Estimate	Standard Error
Constant — β_0	-0.320	0.620
Annual Income — β_1	0.0675	0.004
Home Ownership — β_2	0.827	0.075
Sum of Squared Errors	0.230	
Total Sum of Squares	15.725	

yield the fitted equation

$$\hat{S}_i = -0.320 + 0.0675\, I_i + 0.827\, H_i.$$

a) The value of the coefficient associated with the variable I is estimated to be 0.0675. Provide a one-sentence explanation of what this number implies about the relationship between family income and saving. Also, provide a one-sentence explanation of what the coefficient estimate 0.827 implies about the relationship between home ownership and saving.

b) Using $\alpha = .05$, conduct a test of the null hypothesis H_0: $\beta_1 = \beta_2 = 0$ versus the alternative hypothesis that at least one of β_1, β_2 is not equal to zero.

2. A shoe store owner estimated the following regression equation to explain sales as a function of the size of investment in inventories (X_1) and advertising expenditures (X_2). The sample consisted of 10 stores. All variables are measured in thousands of dollars.

$$\hat{Y} = 29.1270 + .5906\, X_1 + .4980\, X_2$$

The estimated R^2 was .92448, $\Sigma(Y_i - \bar{Y})^2 = 6,724.125$, and the standard deviations of the coefficients of X_1 and X_2 obtained from the regression were .0813 and .0567 respectively.

a) Find the sum of squared residuals and present a point estimate of the variance of the error term. (507.81, 72.54)

b) Can we conclude that sales are dependent to a significant degree on the size of stores' inventory investments?

c) Can we conclude that advertising expenditures have a significant effect on sales?

d) Can we conclude that the regression has uncovered a significant overall relationship between the two independent variables and sales?

e) What do we mean by the term 'significant' in b), c) and d) above?

3. Quality control officers at the Goodyear Tire and Rubber Company are interested in the factors that influence the performance of their Goodyear TA All Season Radial Tires. To this end, they performed a multiple regression analysis based on a random sample of 64 automobiles. Each vehicle was equipped with new tires and driven for one year. Following the test period, Goodyear experts evaluated tire wear by estimating the number of additional months for which the tire could be used. For the regression study, the dependent variable *TIRE* measures this estimated remaining lifetime in months. A totally worn out tire will report $TIRE = 0$. Independent variables selected for the study include *WEIGHT* which measures the test vehicle's weight in pounds, *CITY* which measures the number of miles driven in city traffic in thousands and *MILES* which measures the total number of miles driven (city and highway), also in thousands. The statistical software package Xlispstat reports multiple regression results and a simple regression of *TIRE* on *WEIGHT*. The standard errors of the coefficients are given in brackets.

Dependent Variable: *TIRE*

Constant	60.000	(15.000)
WEIGHT	-0.003	(0.001)
CITY	0.020	(0.008)
MILES	-0.400	(0.100)
R-Squared	.86	
Standard Error ($\hat{\sigma}$)	1.542	
Number of Observations	64	
Degrees of Freedom	60	

Dependent Variable: *TIRE*

Constant	72.000	(36.000)
WEIGHT	-0.005	(0.001)
R-Squared	.79	
Standard Error ($\hat{\sigma}$)	1.732	
Number of Observations	64	
Degrees of Freedom	62	

a) Interpret each of the estimated parameters in the multiple regression model (i.e., what does $\beta_2 = 0.020$ tell you about the relationship between city miles and tire wear?)

b) Briefly discuss why the estimated coefficient on *WEIGHT* differs between the simple and multiple regression models.

c) Perform an hypothesis test to evaluate whether the coefficient on *CITY* is significantly greater than zero. Manage the α-risk at 5%. Interpret the results of this test.

d) Test whether the estimated coefficients on *CITY* and *MILES* are jointly equal to zero. Manage the α-risk at 5%. Interpret the results of this test.

4. J. M. Keynes postulated that aggregate real consumption (RCONS) is positively related to aggregate real GNP (RGNP) in such a way that the marginal propensity to consume—the change in consumption resulting from a one-unit change in income—is less than the average propensity to consume—the ratio of consumption to income. There remains the question of whether consumption is negatively related to the rate of interest (or, which is the same thing, savings is positively related to the interest rate). The table on the next page presents some data on consumption, real GNP and interest rates in Canada, along with the LOTUS-123 regression output using these data. A dummy variable is included to test whether consumption depends on whether the exchange rate is fixed or flexible. The column PRED gives the level of consumption predicted by the regression that includes the dummy variable and the column ERROR gives the difference between the actual value of consumption and the value predicted by that regression. SQERR is the error squared and the right-most column gives the error times itself lagged.

WORKSHEET FOR REGRESSION ANALYSIS OF CANADIAN CONSUMPTION

	RCONS	RGNP	INTRATE	DUMMY	PRED	ERROR	SQERR	ERROR TIMES ERROR LAGGED
1961	105.4	161.4	3.37	0	99.7	5.7	32.3	
1962	111.1	173.4	4.38	0	105.7	5.4	29.0	30.629
1963	116.4	182.8	4.01	1	114.1	2.3	5.4	12.530
1964	122.8	196.6	4.20	1	121.9	0.9	0.8	2.023
1965	129.7	211.5	5.01	1	129.8	-0.1	0.0	-0.079
1966	136.8	228.2	6.27	1	138.4	-1.6	2.5	0.143
1967	142.9	236.3	5.84	1	143.5	-0.5	0.3	0.831
1968	150.0	248.4	6.82	1	149.6	0.4	0.2	-0.207
1969	157.1	262.0	7.84	1	156.5	0.5	0.3	0.212
1970	160.6	271.9	7.34	0	160.2	0.4	0.1	0.196
1971	169.3	288.5	4.51	0	172.4	-3.1	9.5	-1.122
1972	181.0	308.0	5.10	0	183.2	-2.2	4.8	6.763
1973	192.4	335.8	7.45	0	197.2	-4.8	22.6	10.447
1974	202.7	361.1	10.50	0	209.1	-6.4	40.6	30.310
1975	211.9	367.5	7.93	0	215.1	-3.3	10.7	20.837
1976	225.3	393.2	9.17	0	228.9	-3.7	13.6	12.036
1977	231.3	399.7	7.47	0	234.2	-2.9	8.4	10.691
1978	236.2	405.4	8.83	0	236.3	-0.1	0.0	0.294
1979	241.5	423.8	12.07	0	244.0	-2.6	6.5	0.259
1980	246.3	432.0	13.15	0	247.9	-1.6	2.5	4.012
1981	249.3	438.3	18.33	0	246.8	2.4	5.9	-3.810
1982	241.4	415.2	14.15	0	237.2	4.2	17.6	10.189
1983	250.8	427.5	9.45	0	248.6	2.3	5.1	9.495
1984	261.8	449.2	11.18	0	259.6	2.3	5.1	5.106
1985	276.1	465.9	9.56	0	270.7	5.3	28.4	12.037
1986	287.1	474.2	9.16	0	275.9	11.2	126.3	59.940
SUM	5037.1	8557.8	213.1	7.0		10.6	378.6	233.8
MEAN	193.7	329.1	8.2	0.3		0.0		
VAR	3164.17	10421.87	12.6228	0.204				

****

Regression Output:

Constant	9.12
R Squared	0.99527
No. of Observations	26

	RGNP	INTRATE	DUMMY
X Coefficient(s)	0.58	-0.90	2.53
Std Err of Coef.	0.02	0.38	2.40

Regression Output: Dummy Variable Excluded:

Constant	12.21
R Squared	0.99504
No. of Observations	26

	RGNP	INTRATE
X Coefficient(s)	0.57	-0.84
Std Err of Coef.	0.01	0.38

a) Can we conclude that consumption is positively related to income?

b) How would you test the proposition that the marginal propensity to consume equals the average propensity to consume?

c) Can we conclude that the interest rate has a negative effect on consumption?

d) Is aggregate consumption affected by whether the country was on fixed as opposed to flexible exchange rates?

e) Test whether the regression that includes all three independent variables is statistically significant.

f) Do an F-test of the proposition that consumption depends on whether the country was on a fixed or flexible exchange rate. Show that the F-statistic so obtained is equal to the square of the relevant t-statistic in the regression that includes the dummy variable.

g) Perform a crude test of whether residuals of the regression are serially correlated.

Chapter 10

Analysis of Variance

Analysis of variance (ANOVA) models study the relationship between a dependent variable and one or more independent variables within the same framework as do linear regression models but from a different perspective. We begin by viewing from an ANOVA perspective the results of a regression explaining the response of Canadian real money holdings to Canadian real GNP and the interest rate on Canadian 90-day commercial paper.

10.1 Regression Results in an ANOVA Framework

The regression results were as follows:

Dependent Variable: Canadian Real Money Holdings

Constant	10.47	(3.21)
90-Day Paper Rate	-2.62	(0.38)
Real GNP	0.17	(0.01)
R-Squared	.91	
Standard Error ($\hat{\sigma}$)	6.70	
Number of Observations	40	
Degrees of Freedom	37	

The regression model can be seen as attempting to explain the total sum of squares of the dependent variable, real money holdings, using two independent variables, real GNP and the nominal interest rate. The residual sum of squares *SSE* represents the portion of the total sum of squares *SSTO* that cannot be explained by the independent variables. And the sum of

squares due to the regression *SSR* represented the portion of the total sum of squares explained by the regressors. It will be recalled that the R^2 is the ratio of *SSR* to *SSTO*. The regression results above give the standard error of the regression $\hat{\sigma}$ which is a point estimate of σ—it is the square root of the mean square error *MSE*. The mean square error in the regression above is thus

$$MSE = \hat{\sigma}^2 = 6.70^2 = 44.89$$

so the sum of squared errors is

$$SSE = (n - K - 1)\, MSE = (37)(44.89) = 1660.93.$$

Since the coefficient of determination, R^2, equals

$$R^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$$

it follows that

$$R^2 SSTO = SSTO - SSE \quad \Longrightarrow \quad (1 - R^2)\, SSTO = SSE,$$

so that, given R^2 and *SSE*, we can calculate *SSTO* from the relationship

$$SSTO = \frac{SSE}{1 - R^2} = \frac{1660.93}{1 - .91} = \frac{1660.93}{.09} = 18454.78.$$

The sum of squares due to regression then becomes

$$SSR = SSTO - SSE = 18454.78 - 1660.93 = 16793.85.$$

Now the variance of the dependent variable, real money holdings, is the total sum of squares divided by $(n - 1)$, the degrees of freedom relevant for calculating it—one observation out of the n available is used up calculating the mean of the dependent variable. And we have seen that the error variance is estimated by dividing the sum of squared errors by $(n - K - 1)$, the number of degrees of freedom relevant for its calculation—here we have used up K pieces of information calculating the regression coefficients of the independent variables and one piece of information to calculate the constant term, leaving only $(n - K - 1)$ independent squared residuals.

Finally, we can identify the degrees of freedom used in calculating the sum of squares due to regression (*SSR*). *SSR* is the sum of squared deviations of the fitted values of the dependent variable from the mean of the dependent variable—in terms of our regression notation,

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

There are n fitted values $\hat{Y}$ that by the nature of the calculations are constrained to lie along the fitted line. The potential degrees of freedom in calculating this line are its K + 1 parameters—the slopes with respect to the K independent variables, and the intercept. One of these degrees of freedom is lost because only $n - 1$ of the $(\hat{Y}_i - \bar{Y})$ are independent—the deviations must satisfy the constraint

$$\sum_{1=1}^{n}(\hat{Y}_i - \bar{Y}) = 0$$

so if we know any $n - 1$ of these deviations we also know the remaining deviation. The sum of squares due to regression is thus calculated with K degrees of freedom (two in the above example). So we can calculate the variance due to the regression (i.e., the regression mean square) as

$$MSR = \frac{SSR}{K} = \frac{16793.85}{2} = 8396.925.$$

These analysis of variance results can be set out in the following ANOVA table:

Analysis of Variance: Canadian Real Money Holdings

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Regression	16793.85	2	8396.925
Error	1660.93	37	44.89
Total	18454.78	39	

Notice how the total degrees of freedom is the sum of the degrees of freedom for calculating the regression sum of squares and the degrees of freedom for calculating the sum of squared errors. And, as shown in the previous two chapters as well as above, the total sum of squares is equal to the sum of squares due to regression plus the error sum of squares. It is especially important to notice, however, that the mean square due to regression and the mean square error do not add up to equal the variance of the dependent variable, which in the case above is 18454.78/39 = 473.2. The F-Statistic

for testing the null hypothesis of no relationship between the regressors and the dependent variable is

$$\begin{aligned} F &= \frac{\sum(Y_i - \hat{Y})^2 - \sum e_i^2}{K} \div \frac{\sum e_i^2}{n-K-1} \\ &= \frac{SSI0 - SSE}{K} \div \frac{SSE}{n-K-1} \\ &= \frac{MSR}{MSE} = \frac{8396.925}{44.89} = 185.72 \end{aligned}$$

which far exceeds the value of $F(2, 37)$ in the statistical tables for at any reasonable level of α.

10.2 Single-Factor Analysis of Variance

Let us now take a fresh problem and approach it strictly from an ANOVA perspective. Suppose we randomly select 5 male students and 5 female students from a large class and give each student an achievement test. Our objective is to investigate whether male students do better than their female counterparts on such a test. The resulting data are

	Student i				
Gender j	1	2	3	4	5
Male	86	82	94	77	86
Female	89	75	97	80	82

This is a *designed sampling experiment* because we control (and randomize) the selection of male and female participants. It would be an *observational sampling experiment* if we were to simply take a class of 10 students, half of whom turn out to be female, and give them an achievement test.

Analysis of variance has its own terminology. The achievement test score is the response or dependent variable as it would be in a linear regression. The independent variables, whose effects on the response variable we are interested in determining, are called *factors*. In the case at hand, there is a single factor, gender, and it is qualitative—i.e., not measured naturally on a numerical scale. We could add additional factors such as, say, the race of the student. The values of the factors utilized in the experiment are called *factor levels*. In this single factor experiment, we have two factor levels, male and female. In the single factor case the factor levels are also called *treatments*. In an experiment with more than one factor, the treatments are the factor-level combinations utilized. For example, if we take the race

of the students as a second factor, the treatments might be male-white, female-white, male-non-white and female-non-white. The objects on which the response variables are observed—i.e., the individual students in the case considered here—are referred to as *experimental units*. These are called elements in regression analysis.

The objective of a completely randomized design is usually to compare the treatment means—these are the mean achievement scores of male and female students respectively. The means of the two treatments (male and female) are, respectively,

$$\frac{86+82+94+77+86}{5} = 85$$

and

$$\frac{89+75+97+80+82)}{5} = 84.6$$

and the overall mean is 84.8. Some thought suggests that if the response variable (achievement test score) is not much affected by treatment (i.e., by whether the student is male or female) the means for the two treatments will not differ very much as compared to the variability of the achievement test scores around their treatment means. On the other hand, if test score responds to gender, there should be a large degree of variability of the treatment means around their common mean as compared to the variability of the within-group test scores around their treatment means.

We thus calculate the *Sum of Squares for Treatments* by squaring the distance between each treatment mean and the overall mean of all sample measurements, multiplying each squared difference by the number of sample measurements for the treatment, and adding the results over all treatments. This yields

$$SST = \sum_{j=1}^{p}(n_j)(\bar{x}_j - \bar{x})^2 = (5)(85-84.8)^2 + (5)(84.6-84.8)^2$$

$$= (5)(.04) + (5)(.04) = .2 + .2 = .4.$$

In the above expression $p = 2$ is the number of treatments, n_j is the number of sample elements receiving the j-th treatment, $\bar{x}_j$ is the mean response for the jth treatment and $\bar{x}$ is the mean response for the entire sample.

Next we calculate the *Sum of Squares for Error*, which measures the sampling variability within the treatments—that is, the variability around the treatment means, which is attributed to sampling error. This is computed by summing the squared distance between each response measurement and

the corresponding treatment mean and then adding these sums of squared differences for all (both) treatments. This yields

$$SSE = \sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2$$

$$= [(86-85)^2 + (82-85)^2 + (94-85)^2 + (77-85)^2 + (86-85)^2]$$
$$+[(89-84.6)^2 + (75-84.6)^2 + (97-84.6)^2 + (80-84.6)^2 + (82-84.6)^2]$$
$$= [1+9+81+64+1] + [19.36 + 92.16 + 153.76 + 21.16 + 6.76]$$
$$= 156 + 293.2 = 449.2.$$

Again, n_j is the number of sample measurements for the jth treatment and x_{ij} is the ith measurement for the jth treatment.

Finally, the *Total Sum of Squares* is the sum of squares for treatments plus the sum of squares for error. That is

$$SSTO = SST + SSE = .4 + 449.2 = 449.6.$$

Now we calculate the *Mean Square for Treatments* which equals the sum of squares for treatments divided by the appropriate degrees of freedom. We are summing p squared deviations (of each of the p treatment means from the overall mean) but only $p-1$ of these squared deviations are independent because we lose one piece of information in calculating the overall mean. So for the above example we have

$$MST = \frac{SST}{p-1} = \frac{0.4}{1} = 0.4.$$

Next we calculate the *Mean Square Error* which equals the sum of the squared deviations of the sample measurements from their respective treatment means for all measurements, again divided by the appropriate degrees of freedom. Here we have n cases (or sample measurements), where

$$n = n_1 + n_2 + n_3 + \ldots + n_p$$

but we had to calculate the p treatment means from the data, so the degrees of freedom will be $n - p$. We thus obtain

$$MSE = \frac{SSE}{n-p} = \frac{SSTO - SST}{n-p} = \frac{449.2}{10-2} = 56.15.$$

The above numbers can be used to construct the following ANOVA table:

Analysis of Variance: Achievement Test Scores

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Treatments	0.4	1	0.4
Error	449.2	8	56.15
Total	449.6	9	

The purpose of this whole exercise, of course, is to determine whether gender (given by treatments) has any effect on achievement test scores (the response variable). If there is no effect we would expect the error sum of squares to be nearly as big as the total sum of squares and the treatment sum of squares to be very small. This appears to be the case in the ANOVA table above. The sum of squares for treatments (which measures the variability of the treatment means around the overall mean) is extremely low relative to the error sum of squares. But is it low enough for us to conclude that there is no significant relationship of achievement scores to gender? Is the observed treatment sum of squares as high as it is purely because of sampling error?

The statistical test for significance is straight forward. From the discussions in the previous chapter it is evident that under the null hypothesis of no relationship

$$\frac{SST}{\sigma^2} = \frac{SSTO - SSE}{\sigma^2} = \chi^2(p-1)$$

where $(p-1)[= (n-1) - (n-p)]$ is the degrees of freedom for treatment and σ^2 is the common variance of the individual achievement scores around the overall mean achievement score and of the individual scores around their treatment means. The two types of variation have a common variance under the null hypotheses that the achievement test scores are independent of treatment. Also,

$$\frac{SSE}{\sigma^2} = \chi^2(n-p).$$

We can now apply the principle that the ratio of two independent χ^2 variables, each divided by its degrees of freedom, will be distributed according to the F-distribution with parameters equal to the number of degrees of freedom in the numerator and number of degrees of freedom in the denominator.

Thus we have

$$\frac{SSTO - SSE}{(n-1)-(n-p)} \div \frac{SSE}{n-p} = \frac{SST}{p-1} \div \frac{SSE}{n-p} = \frac{MST}{MSE} = F(p-1, n-p)$$

where the σ^2 terms cancel out. In the example under consideration, this yields

$$\frac{MST}{MSE} = \frac{.4}{56.15} = .007123778 = F(1,8).$$

The critical value of F with one degree of freedom in the numerator and 8 degrees of freedom in the denominator for $\alpha = .1$ is 3.46. So we cannot reject the null hypothesis of no effect of gender on achievement test scores.

You might recognize the similarity of this analysis of variance test to the tests we did in Chapter 6 for differences in the means of two populations. Indeed, the tests are identical. In Chapter 6 we expressed the difference between the two population means as

$$E\{\bar{Y} - \bar{X}\} = E\{\bar{Y}\} - E\{\bar{X}\} = \mu_2 - \mu_1$$

and the variance of the difference between the two means as

$$\sigma^2\{\bar{Y} - \bar{X}\} = \sigma^2\{\bar{Y}\} + \sigma^2\{\bar{X}\},$$

using

$$s^2\{\bar{Y} - \bar{X}\} = s^2\{\bar{Y}\} + s^2\{\bar{X}\}$$

as an unbiased point estimator of $\sigma^2\{\bar{Y} - \bar{X}\}$. We then used in this formula the expressions for the variances of the means,

$$s^2\{\bar{Y}\} = s^2\{Y/n\}$$

and

$$s^2\{\bar{X}\} = s^2\{X/n\}.$$

The difference in means in the case above is $85 - 84.6 = 0.4$. The sample population variances can be obtained by noting that the sums of the squared deviations of the achievement scores of the male and female students around their respective means are, respectively, 156 and 293.2. Dividing each of these by the degrees of freedom relevant for their calculation ($n_i - 1 = 5 - 1 = 4$), we obtain sample population variances for male and female students of 39 and 73.3 respectively. Imposing the condition that the true variances of the two groups are the same, we then obtain a pooled estimator of this common variance by calculating a weighted average of the two estimated variances

with the weights being the ratios of their respective degrees of freedom to the total. That is

$$s_P^2 = \frac{(4)(39) + (4)(73.3)}{8} = \frac{156 + 293.2}{8} = 56.15$$

which, you will note, equals *MSE*. The variance of the difference between the two means (which we denote using the subscripts m for male and f for female) equals

$$\sigma_{m-f}^2 = \frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f} = s_P^2 \left[\frac{1}{n_m} + \frac{1}{n_f} \right] = (56.15)(.2 + .2) = 22.46.$$

The standard deviation of the difference between the two means then equals $\sqrt{22.46} = 4.739198$. Given the point estimate of the difference in the means of 0.4, the t-statistic for testing the null-hypothesis of no difference between the means is

$$t^* = \frac{.4}{47392} = .08440246.$$

This statistic will be within the acceptance region for any reasonable level of significance. The result is the same as we obtained from the analysis of variance.

As a matter of fact, this test and the analysis of variance test are identical. Squaring t^*, we obtain .007123778 which equals the F-statistic obtained in the analysis of variance procedure. This is consistent with the principle, already noted, that when there is one degree of freedom in the numerator, $F = t^2$.

A third way of approaching this same problem is from the point of view of regression analysis. We have $n = 10$ observations on gender and want to determine the response of achievement test score to gender. Gender is a qualitative variable which we can introduce as a dummy variable taking a value of 0 for elements that are male and 1 for elements that are female. Our regression model becomes

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where Y_i , $i = 1 \ldots 10$, is the test score for the i-th student, and X_i is the dummy variable taking a value of zero for male students and unity for female students. The regression results obtained are:

Dependent Variable: Achievement Test Score

Constant	85	(3.35112)
Female Dummy	-.40	(4.73920)
R-Squared	.000889	
Standard Error ($\hat{\sigma}$)	7.4933	
Number of Observations	10	
Degrees of Freedom	8	

The dummy variable for female indicates that the 'constant term for females' is $85 - 0.4 = 84.6$, which is the treatment mean for females obtained by the analysis of variance procedure. The $t-$ratio for the hypothesis that the female dummy is zero (i.e., the female treatment mean equals the male treatment mean) is -.4/473920, which is the same as was obtained for the above test for difference between the means. And the square of $\hat{\sigma}$ is 56.15, the mean squared error obtained in the analysis of variance procedure.

Now let us take a more complicated problem. Suppose we randomly divide fifteen male students enrolled in a mathematics course into three groups of five students each. We then randomly assign each group to one of three instructional modes: (1) programmed text, (2) video-taped lecture-style presentation, and (3) interactive computer programs. These modes are all designed to augment a standard textbook which is the same for all three groups. At the end of the course, we give all students the same achievement test, with the following results:

	Student i				
Mode j	1	2	3	4	5
1	86	82	94	77	86
2	90	79	88	87	96
3	78	70	65	74	63

Again we have a *designed sampling experiment* because we were able to control the details of the instructional modes for the three groups and make sure that students were randomly assigned to groups and groups were randomly assigned to instructional modes. The experiment is completely *randomized* because the allocation of the students to the three groups is random and the allocation of the groups to the instructional modes is random. In contrast, an *observational sampling experiment* would be one where we, for example, observe the test scores of three groups of students, perhaps of different sizes,

who for reasons beyond our control happen to have been instructed in accordance with three alternative instructional modes of the above types. In this single factor study there are three factor levels or treatments representing the three modes of instruction.

Our objective in this completely randomized design is to compare the treatment means—the mean achievement scores of the students in the three groups taught using the different instructional modes. The means of the three modes are

$$\frac{86+82+94+77+86}{5} = 85$$

$$\frac{90+79+88+87+96}{5} = 88$$

$$\frac{78+70+65+74+63}{5} = 70$$

And the overall mean is 81. Again we note that if the response variable (achievement test score) is not much affected by treatment (instructional mode) the means for the three treatments will not differ very much as compared to the variability of the achievement test scores around their treatment means. On the other hand, if test score responds to instructional mode, there should be a large degree of variability of the treatment means around their common mean as compared to the variability of the within-group test scores around their treatment means.

We again calculate the *Sum of Squares for Treatments* by squaring the distance between each treatment mean and the overall mean of all sample measurements, multiplying each squared distance by the number of sample measurements for the treatment, and adding the results over all treatments.

$$SST = \sum_{j=1}^{p} n_j \left(\bar{x}_j - \bar{x}\right)^2 = (5)(85-81)^2 + (5)(88-81)^2 + (5)(70-81)^2$$

$$= (5)(16) + (5)(49) + (5)(121) = 80 + 245 + 605 = 930.$$

In the above expression $p = 3$ is the number of treatments, $\bar{x}_j$ is the mean response for the jth treatment and $\bar{x}$ is the mean response for the entire sample.

Next we calculate the *Sum of Squares for Error*, which measures the sampling variability within the treatments—the variability around the treatment means that we attribute to sampling error. This is computed by summing the squared distance between each response measurement and the corresponding treatment mean and then adding the squared differences over all

measurements in the entire sample.

$$SSE = \sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2 + \sum_{i=1}^{n_3}(x_{i3} - \bar{x}_3)^2$$

$$= (86-85)^2 + (82-85)^2 + (94-85)^2 + (77-85)^2 + (86-85)^2$$
$$+(90-88)^2 + (79-88)^2 + (88-88)^2 + (87-88)^2 + (96-88)^2$$
$$+(78-70)^2 + (70-70)^2 + (65-70)^2 + (74-70)^2 + (63-70)^2$$
$$= [1+9+81+64+1] + [4+81+0+1+64] + [64+0+25+16+49]$$
$$= 156 + 150 + 154 = 460.$$

Again, n_j is the number of sample measurements for the jth treatment, which turns out to be 5 for all treatments, and x_{ij} is the ith measurement for the jth treatment.

Finally, the *Total Sum of Squares*, which equals the sum of squares for treatments plus the sum of squares for error, is

$$SSTO = SST + SSE = 930 + 460 = 1390.$$

Now we calculate the *Mean Square for Treatments* which equals the sum of squares for treatments divided by the appropriate degrees of freedom. We are summing 3 squared deviations from the overall mean but only 2 of these squared deviations are independent because we lose one piece of information in calculating the overall mean. So we have

$$MST = \frac{SST}{p-1} = \frac{930}{2} = 465.$$

Finally, we calculate the *Mean Square Error* which equals the sum of the squared deviations of the sample measurements from their respective treatment means for all measurements, again divided by the appropriate degrees of freedom. Here we have 15 cases (or sample measurements), but we had to calculate the 3 treatment means from the data, so the degrees of freedom will be 12. We thus obtain

$$MSE = \frac{SSE}{n-p} - \frac{460}{12} = 38.333.$$

The above numbers can be used to construct the following ANOVA table:

Analysis of Variance: Achievement Test Scores

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Treatments	930	2	465
Error	430	12	38.33
Total	1390	14	

Our goal is to determine whether mode of instruction (given by treatments) has any effect on achievement test score (the response variable). If there is no effect we would expect the error sum of squares to be nearly as big as the total sum of squares and the treatment sum of squares to be very small. It turns out that the sum of squares for treatments (which measures the variability of the treatment means around the overall mean) is quite high relative to the error sum of squares. But is it high enough for us to conclude that there is a significant response of achievement scores to instructional mode? We answer this question by doing an F-test. The F-statistic obtained is

$$\frac{MST}{MSE} = \frac{465}{38.33} = 12.13 = F(2, 12),$$

which is well above the critical value of 6.93 for $\alpha = .01$. We reject the null hypothesis of no effect of instruction mode on achievement test score.

The natural question to ask at this point is: Which of the instructional modes are responsible for the significant overall relationship? All our analysis of variance results tell us is that there is a significant effect of at least one of the three modes of instruction, compared to the other two, on achievement test score. We have not established the relative importance of these modes in determining students' achievement test scores. To investigate this, we can approach the problem from the point of view of regression analysis.

The dependent variable for our regression is achievement test score in a sample of 15 students. Taking the programmed text instructional mode as a reference, we create two dummy variables—one that takes a value of 1 when the instructional mode is video-taped lecture and zero otherwise, and a second that takes a value of 1 when the mode is interactive computer programs and zero otherwise. The effect of programmed text, the reference treatment, is thus measured by the constant terms and the differences in the effects of

the other two treatments from the reference treatment are measured by the coefficients of their respective dummy variables. Our regression model is therefore

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

where Y_i , $i = 1 \ldots 15$, is the test score for the i-th student, and X_{1i} is the dummy variable for video-taped lecture and X_{2i} is the dummy variable for computer programs. The regression results obtained are:

Dependent Variable: Achievement Test Score

Constant	85	(2.77)
Dummy-video	3	(3.92)
Dummy-computer	-15	(3.92)
R-Squared	.67	
Standard Error ($\hat{\sigma}$)	6.19139	
Number of Observations	15	
Degrees of Freedom	12	

The mean score for students using programmed text is equal to the constant term, 85. And the mean score for students receiving video-taped lectures is 3 points higher than that for students using programmed text—i.e., 85 + 3 = 88. Finally, the mean score for students using computer programs is 15 points less than those using programmed text—i.e., 85 - 15 = 70. These correspond to the means calculated earlier. The t-statistic for testing the null hypothesis of no difference between the means for programmed text and video-taped lectures—that is $\beta_1 = 0$—is

$$t^* = \frac{3}{3.92} = .765,$$

which is well with any reasonable acceptance region. So we cannot reject the null hypothesis of no difference between the means for programmed text and video-taped lecture. The t-statistic for testing the null hypothesis of no difference between computer program and programmed text is

$$t^* = \frac{-15}{3.92} = -3.83,$$

leading us to conclude that mean test score under computer programmed learning is significantly below that of programmed text—the critical value of $t(12)$ for $\alpha = .005$ is 3.055.

The question arises as to whether there is a significant difference between the test scores under video-taped lecture vs. computer programmed learning. This would seem to be the case. To check this out we rerun the regression letting video-taped lecture be the reference—that is, including dummies for programmed text and computer program but no dummy variable for video-taped lecture. This yields

Dependent Variable: Achievement Test Score

Constant	88	(2.77)
Dummy-text	-3	(3.92)
Dummy-computer	-18	(3.92)
R-Squared	.67	
Standard Error ($\hat{\sigma}$)	6.19139	
Number of Observations	15	
Degrees of Freedom	12	

The computer program dummy is clearly statistically significant, having a $t-$statistic of -4.59. We have to reject the null hypothesis of no difference between the mean test scores under video-taped lecture and computer programmed learning.

Notice how the difference between the coefficient of Dummy-video and Dummy-computer in the regression that uses programmed text as the reference treatment is exactly the same as the coefficient of Dummy-computer in the regression that uses video-taped lectures as the reference treatment, and that the standard errors of the dummy coefficients are the same in both regressions. It would appear that instead of running the second regression we could have simply subtracted the coefficient of Dummy-computer from the coefficient of Dummy-video (to obtain the number 18) and then simply divided that difference by the variance of all dummy coefficients to obtain the correct t-statistic for testing the null hypothesis of no difference between the coefficients of the two dummy variables.

This suggests that we might have approached the problem of testing for a significant difference between the two coefficients in the same way as we approached the problem of comparing two population means in Chapter 6. In the problem at hand, however, the required computations are different than we used in Chapter 6 for two reasons. First, the regression coefficients we are comparing represent the mean responses of the dependent variable to the respective independent variables, so their variances are the variances of

means rather than population variances. We therefore do not need to divide these variances by n. Second, the coefficients of the independent variables in linear regressions are not necessarily statistically independent, so we cannot obtain the variance of the difference between two coefficients simply by adding their variances—we must subtract from this sum an amount equal to twice their covariance. The variance-covariance matrix of the coefficients in the regression that used programmed text as the reference treatment is[1]

	b_0	b_1	b_2
b_0	7.6666	-7.6666	-7.6666
b_1	-7.6666	15.3333	7.6666
b_2	-7.6666	7.6666	15.3333

The variance of the difference between the coefficient estimates b_1 and b_2 is

$$Var\{b_1 - b_2\} = Var\{b_1\} + Var\{b_2\} - 2\,Cov\{b_1, b_2\}$$

$$= 15.3333 + 15.3333 - (2)(7.6666) = 15.3333 + 15.3333 - 15.3333 = 15.3333.$$

The standard deviation of the difference between the two coefficients is therefore equal to the square root of 15.3333, which equals 3.91578, the standard error of the coefficients of both dummy variables. So we can legitimately test whether the coefficients of Dummy-video and Dummy computer differ significantly by taking the difference between the coefficients and dividing it by their common standard error to form an appropriate t-statistic.

It should be noted, however, that although we could have obtained an appropriate test of the difference between the coefficients of the two dummy variables in this case by simply dividing the difference between the coefficients by their common standard error and comparing the resulting t-statistic with the critical values in the table at the back of our textbook, this will not necessarily work under all circumstances. We have not investigated what would be the best procedure to follow when, for example, the numbers of sample elements receiving each of the three treatments differ. We always have to take account of the fact that the covariance between estimated regression coefficients will not in general be zero.

[1]This was obtained from XlispStat, the computer program used to calculate the regression. Using the matrix notation we very briefly developed in Chapter 9, the variance covariance matrix can be written (see page 227) as $s^2(X'X)^{-1}$.

10.3 Two-factor Analysis of Variance

In ending this chapter we examine briefly a *two-factor designed experiment*. We add fifteen randomly selected female students to the fifteen male students in the above single factor experiment. These fifteen female students are also randomly divided into three groups of 5 students each. One group is instructed by programmed text, one by video-taped lecture and one by computer programs. In this two factor experiment the number of treatments expands from three to six according to the six factor combinations—male-text, male-video, male-computer, female-text, female-video and female-computer. The best way to approach this problem for our purposes is to use a regression analysis of the sort immediately above. In setting up the regression, we obviously need a dummy variable to separate the genders—we let it take a value of 0 if the student is male and 1 if the student is female. Letting programmed text be the reference, we also need dummy variables for video-taped lecture (taking the value of 1 if the instructional mode is video-taped lecture and zero otherwise) and for computer programmed learning (taking a value of 1 if the instructional mode is computer programs and zero otherwise). This would give us the following regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

where Y_i is test score, X_{1i} is the female dummy, X_{2i} the video dummy and X_{3i} the computer dummy. The mean test scores identified in the model are as follows:

Males-text	β_0
Females-text	$\beta_0 + \beta_1$
Males-video	$\beta_0 + \beta_2$
Males-computer	$\beta_0 + \beta_3$
Females-video	$\beta_0 + \beta_1 + \beta_2$
Females-computer	$\beta_0 + \beta_1 + \beta_3$

But this imposes the condition that the effects of the different modes of instruction on achievement test scores be the same for males as for females—using video-taped-lectures instead of programmed text will increase the test scores by an amount equal to β_2 for both males and females, and using computer programs instead of programmed text will increase their test scores uniformly by β_3.

This formulation is inadequate because we should be taking account of whether mode of instruction has a differential effect on the achievement

test scores of females and males. We do this by adding *interaction dummy variables* constructed by multiplying the female dummy by the mode-of-instruction dummy. Our regression model then becomes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{2i} + \beta_5 X_{1i} X_{3i} + \epsilon_i.$$

The first three independent variables are the same as before—female dummy, the video dummy, and the computer dummy. The fourth independent variable is the product of the female dummy and the video dummy—it will take the value 1 if the student is both female and using video-taped lecture instruction and 0 otherwise. And the fifth independent variable is the product of the female dummy and computer dummy, which will be equal to 1 if the student is both female and using computer programmed instruction and 0 otherwise. Notice that the five dummy variables together with the constant term represent the six treatments in the two-factor experiment. The mean test scores identified in the model for the six treatments are now

Males-text	β_0
Females-text	$\beta_0 + \beta_1$
Males-video	$\beta_0 + \beta_2$
Males-computer	$\beta_0 + \beta_3$
Females-video	$\beta_0 + \beta_1 + \beta_2 + \beta_4$
Females-computer	$\beta_0 + \beta_1 + \beta_3 + \beta_5$

The regression results obtained are as follows:

Dependent Variable: Achievement Test Score

Constant	85	(2.86)
Dummy-female	-0.4	(4.05)
Dummy-video	3	(4.05)
Dummy-computer	-15	(4.05)
Dummy-video-female	1	(5.73)
Dummy-computer-female	14	(5.73)
R-Squared	.54	
Standard Error ($\hat{\sigma}$)	6.40182	
Number of Observations	30	
Degrees of Freedom	24	

The coefficient of Dummy-video, which is a point estimate of β_2, measures the difference in male scores under video-taped lecture instruction as

compared to programmed text. Its t-ratio of

$$t^* = \frac{3}{4.05} = .74071$$

indicates that β_2 is not significantly different from zero. The coefficient of Dummy-computer is a point estimate of β_3 and measures the difference in male scores under computer-programmed instruction as compared to programmed text. Its t-ratio is

$$t^* = \frac{-15}{4.05} = -3.7037,$$

indicating a significant negative effect of computer programs over programmed text as a method of instruction. The coefficient for Dummy-female is a point estimate of β_1, measuring the effect of being female rather than male on achievement test scores when programmed text is the method of instruction. It should be obvious that β_1 is not significantly different from zero. The point estimate of β_4, the coefficient of Dummy-video-female, measures the estimated effect of being female rather than male when taking video-lecture instruction. The relevant t-ratio, .17452, indicates no significant effect. Finally, the coefficient of Dummy-computer-female, which measures the effect of being female rather than male when taking computer-programmed instruction, is plus 14 with a t-statistic of

$$t^* = \frac{14}{5.73} = 2.4433.$$

This indicates a significantly positive effect of being female rather than male when taking computer programmed instruction. It is clear that females do significantly better than males when the instruction mode is computer programs. In fact, it can be seen from a comparison of the coefficient of Dummy-computer with that of Dummy-computer-female that the negative effect of computer programmed instruction on learning, which is statistically significant for male students, almost vanishes when the student is female.

10.4 Exercises

1. In a completely randomized design experiment with one factor the following data were obtained for two samples:

Sample 1: 5 5 7 11 13 13
Sample 2: 10 10 12 16 18 18

Test the null hypothesis that the two samples were drawn from populations with equal means and draw up the appropriate ANOVA table.

2. A clinical psychologist wished to compare three methods for reducing hostility levels in university students. A certain psychological test (HLT) was used to measure the degree of hostility. High scores on this test indicate great hostility. Eleven students obtaining high and nearly equal scores were used in the experiment. Five were selected at random from among the eleven problem cases and treated by method A. Three were taken at random from the remaining six students and treated by method B. The other three students were treated by method C. All treatments continued throughout a semester. Each student was given the HLT test again at the end of the semester, with the results shown below:

Method A	Method B	Method C
73	54	79
83	74	95
76	71	87
68		
80		

Do the data provide sufficient evidence to indicate that at least one of the methods of treatment produces a mean student response different from the other methods? What would you conclude at the $\alpha = .05$ level of significance?

3. Is eating oat bran an effective way to reduce cholesterol? Early studies indicated that eating oat bran daily reduces cholesterol levels by 5 to 10%. Reports of these studies resulted in the introduction of many new breakfast cereals with various percentages of oat bran as an ingredient. However, a January 1990 experiment performed by medical researchers in Boston, Massachusetts cast doubt on the effectiveness of oat bran. In that study,

20 volunteers ate oat bran for breakfast and another 20 volunteers ate another grain cereal for breakfast. At the end of six weeks the percentage of cholesterol reduction was computed for both groups:

Oat Bran	Other Cereal
14	3
18	3
4	8
9	11
4	9
0	7
12	12
2	13
8	18
12	2
10	7
11	5
12	1
6	5
15	3
17	13
12	11
4	2
14	19
7	9

What can we conclude at the 5% significance level?

4. Prior to general distribution of a successful hardcover novel in paperback form, an experiment was conducted in nine test markets with approximately equal sales potential. The experiment sought to assess the effects of three different price discount levels for the paperback (50, 75, 95 cents off the printed cover price) and the effects of three different cover designs (abstract, photograph, drawing) on sales of the paperback. Each of the nine combinations of price discount and cover design was assigned at random to one of the test markets. The dependent variable was sales, and the independent variables were the discount off cover price, a dummy variable taking a value of 1 if the design was photograph and 0 otherwise, and a dummy variable taking a value of 1 if the design was drawing and 0 otherwise.

The regression results were as follows:

Dependent Variable: Sales

Constant	6.03685	(0.753114)
Discount	0.18363	(0.009418)
Photo-dummy	-0.68333	(0.424682)
Drawing-Dummy	1.60000	(0.424682)
R-Squared	.98970	
Standard Error ($\hat{\sigma}$)	0.520126	
Number of Observations	9	
Degrees of Freedom	5	

The numbers in brackets are the standard errors of the respective coefficients and $\hat{\sigma}$ is the standard error of the regression, a point estimate of the standard deviation of the error term.

a) Is there good evidence that discounting the price increases sales?

b) Is there good evidence that using an abstract cover rather than putting on a photograph or drawing results in less sales?

c) Is the overall regression relationship statistically significant?

d) What would be the expected level of sales if the discount is 75 cents off the printed cover price and a drawing is put on the cover?

Index