

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

©J. E. Floyd, University of Toronto

Contents

1	Introduction to Statistics, Data and Statistical Thinking	1
1.1	What is Statistics?	1
1.2	The Use of Statistics in Economics and Other Social Sciences	1
1.3	Descriptive and Inferential Statistics	4
1.4	A Quick Glimpse at Statistical Inference	5
1.5	Data Sets	7
1.6	Numerical Measures of Position	18
1.7	Numerical Measures of Variability	22
1.8	Numerical Measures of Skewness	24
1.9	Numerical Measures of Relative Position: Standardised Values	25
1.10	Bivariate Data: Covariance and Correlation	27
1.11	Exercises	31
2	Probability	35
2.1	Why Probability?	35
2.2	Sample Spaces and Events	36
2.3	Univariate, Bivariate and Multivariate Sample Spaces	38
2.4	The Meaning of Probability	40
2.5	Probability Assignment	41
2.6	Probability Assignment in Bivariate Sample Spaces	44
2.7	Conditional Probability	45
2.8	Statistical Independence	46
2.9	Bayes Theorem	49
2.10	The AIDS Test	52
2.11	Basic Probability Theorems	54
2.12	Exercises	55

3	Some Common Probability Distributions	63
3.1	Random Variables	63
3.2	Probability Distributions of Random Variables	64
3.3	Expected Value and Variance	67
3.4	Covariance and Correlation	70
3.5	Linear Functions of Random Variables	73
3.6	Sums and Differences of Random Variables	74
3.7	Binomial Probability Distributions	76
3.8	Poisson Probability Distributions	83
3.9	Uniform Probability Distributions	86
3.10	Normal Probability Distributions	89
3.11	Exponential Probability Distributions	94
3.12	Exercises	96
4	Statistical Sampling: Point and Interval Estimation	103
4.1	Populations and Samples	103
4.2	The Sampling Distribution of the Sample Mean	106
4.3	The Central Limit Theorem	110
4.4	Point Estimation	114
4.5	Properties of Good Point Estimators	115
	4.5.1 Unbiasedness	115
	4.5.2 Consistency	116
	4.5.3 Efficiency	116
4.6	Confidence Intervals	117
4.7	Confidence Intervals With Small Samples	119
4.8	One-Sided Confidence Intervals	122
4.9	Estimates of a Population Proportion	122
4.10	The Planning of Sample Size	124
4.11	Prediction Intervals	125
4.12	Exercises	127
4.13	Appendix: Maximum Likelihood Estimators	130
5	Tests of Hypotheses	133
5.1	The Null and Alternative Hypotheses	133
5.2	Statistical Decision Rules	136
5.3	Application of Statistical Decision Rules	138
5.4	P -Values	140

5.5	Tests of Hypotheses about Population Proportions	142
5.6	Power of Test	143
5.7	Planning the Sample Size to Control Both the α and β Risks	148
5.8	Exercises	151
6	Inferences Based on Two Samples	155
6.1	Comparison of Two Population Means	155
6.2	Small Samples: Normal Populations With the Same Variance	157
6.3	Paired Difference Experiments	159
6.4	Comparison of Two Population Proportions	162
6.5	Exercises	164
7	Inferences About Population Variances and Tests of Goodness of Fit and Independence	169
7.1	Inferences About a Population Variance	169
7.2	Comparisons of Two Population Variances	173
7.3	Chi-Square Tests of Goodness of Fit	177
7.4	One-Dimensional Count Data: The Multinomial Distribution	180
7.5	Contingency Tables: Tests of Independence	183
7.6	Exercises	188
8	Simple Linear Regression	193
8.1	The Simple Linear Regression Model	194
8.2	Point Estimation of the Regression Parameters	197
8.3	The Properties of the Residuals	200
8.4	The Variance of the Error Term	201
8.5	The Coefficient of Determination	201
8.6	The Correlation Coefficient Between X and Y	203
8.7	Confidence Interval for the Predicted Value of Y	204
8.8	Predictions About the Level of Y	206
8.9	Inferences Concerning the Slope and Intercept Parameters	208
8.10	Evaluation of the Aptness of the Model	210
8.11	Randomness of the Independent Variable	213
8.12	An Example	213
8.13	Exercises	218

9	Multiple Regression	223
9.1	The Basic Model	223
9.2	Estimation of the Model	225
9.3	Confidence Intervals and Statistical Tests	227
9.4	Testing for Significance of the Regression	229
9.5	Dummy Variables	233
9.6	Left-Out Variables	237
9.7	Multicollinearity	238
9.8	Serially Correlated Residuals	243
9.9	Non-Linear and Interaction Models	248
9.10	Prediction Outside the Experimental Region: Forecasting . .	254
9.11	Exercises	255
10	Analysis of Variance	261
10.1	Regression Results in an ANOVA Framework	261
10.2	Single-Factor Analysis of Variance	264
10.3	Two-factor Analysis of Variance	277
10.4	Exercises	280

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 (25th Percentile)
14	344	F		
15	346	F		
16		
...				
23	469	F		
24	490	M		
25	513	M		
			521.50	Median (50th Percentile)
26	530	M		
27	548	F		
28	552	F		
29		
...				
35	675	M		
36	728	M		
37	745	F		
			748	3rd (Upper) Quartile (75th Percentile)
38	751	F		
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

¹The 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.

²Wrong! 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!

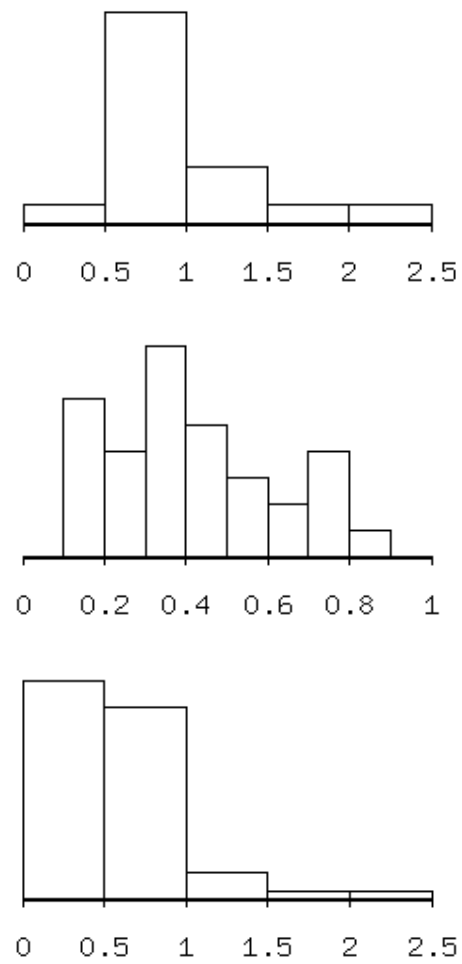


Figure 1.1: Histogram of weekly wages for male (top), female (middle) and all (bottom) employees. The horizontal scale is thousands of U.S. dollars.

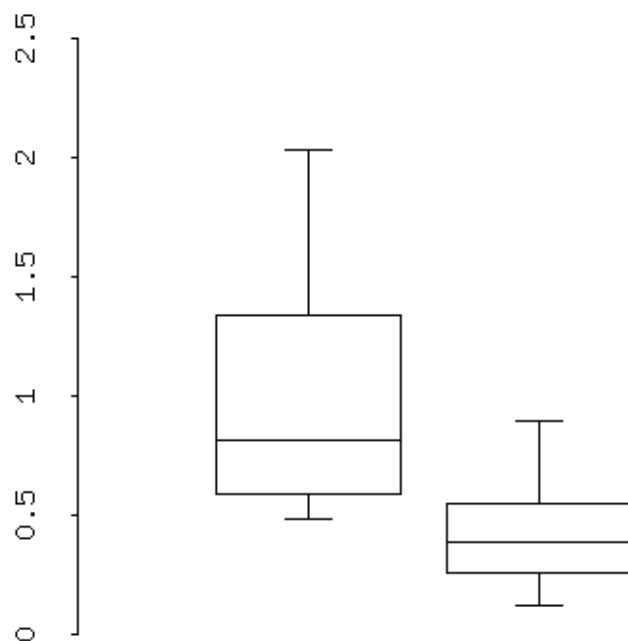


Figure 1.2: Box plot of weekly wages for males (left) and females (right). The vertical scale is thousands of U.S. dollars.

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.

³Consumer 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*.

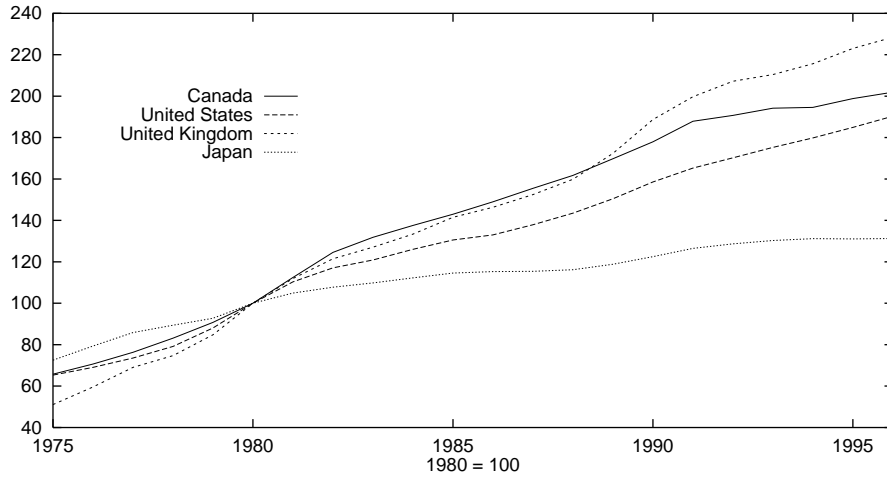


Figure 1.3: Consumer price indexes of selected countries

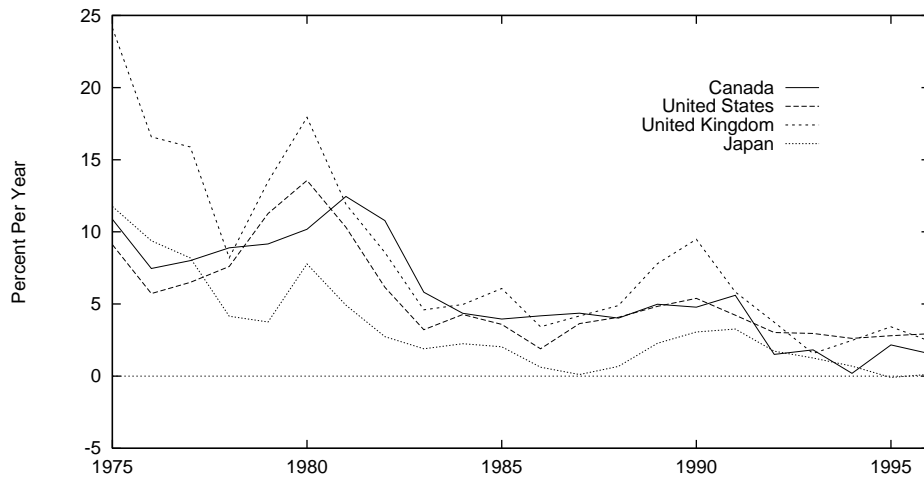


Figure 1.4: Year-over year inflation rates of selected countries

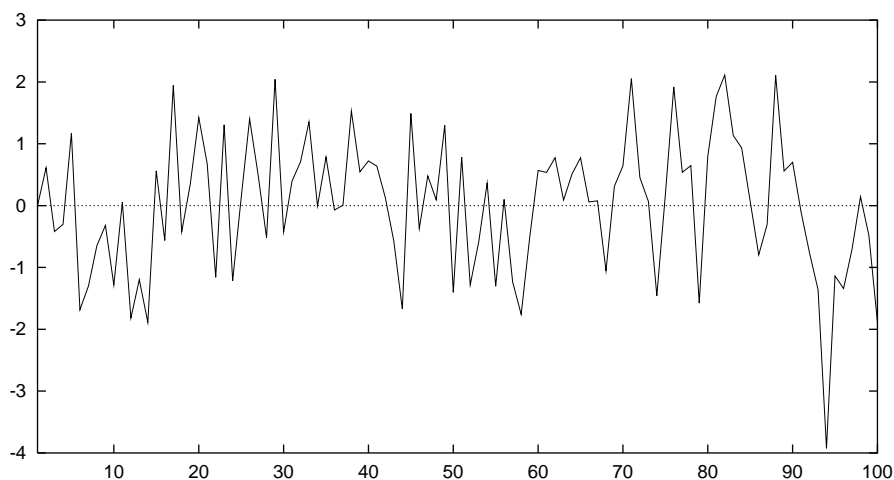


Figure 1.5: A time-series devoid of autocorrelation

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 $X_1, X_2, X_3, \dots, X_N$ is defined as

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N} \quad (1.1)$$

where \bar{X} is the arithmetic mean and

$$\sum_{i=1}^N X_i = X_1 + X_2 + X_3 + \dots + X_N. \quad (1.2)$$

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,

$$\sum_{i=1}^N X_i = N\bar{X}, \quad (1.3)$$

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

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

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,

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

where \bar{X}_i is the mean of the observations in the i th class and f_i is the number (frequency) of observations in the i th 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

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

or the n th root of $X_1X_2X_3\dots X_N$.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

⁴Note from the definition of logarithms that taking the logarithm of the n th root of $(X_1X_2X_3\dots X_N)$, which equals

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

yields

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

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.

⁵The 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

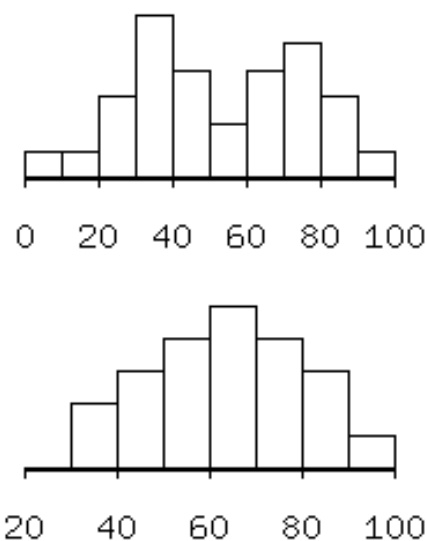


Figure 1.6: Bimodal distribution (top) and unimodal or humped-shaped distribution (bottom).

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

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

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

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

where S is the number of classes, f_i is the frequency of the i th class, \bar{X}_i is the mean of the i th 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,

$$C = \frac{100s}{\bar{X}}, \quad (1.8)$$

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.

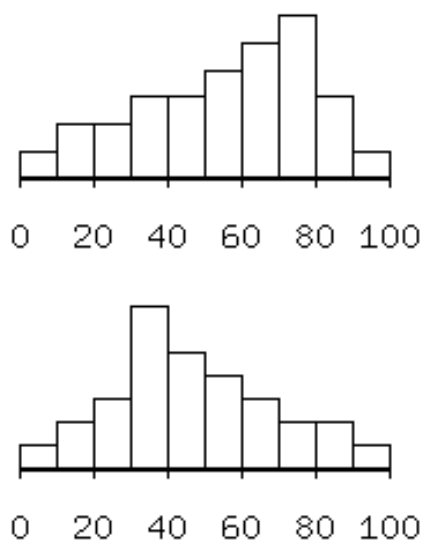


Figure 1.7: Left-skewed distribution (top—mean = 55.1 , median = 58, mode = 75) and right-skewed distribution (bottom —mean = 46.4, median = 43.5, mode = 35).

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,

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

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 $X_1, X_2, X_3 \dots X_N$ are given by

$$Z_i = \frac{X_i - \mu}{\sigma} \quad (1.10)$$

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

$$Z_i = \frac{X_i - \bar{X}}{s} \quad (1.11)$$

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 i th household. Now consider a random sample of N households which yields the paired observations (X_i, Y_i) for $i = 1, 2, 3, \dots, N$.

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

⁶This 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.

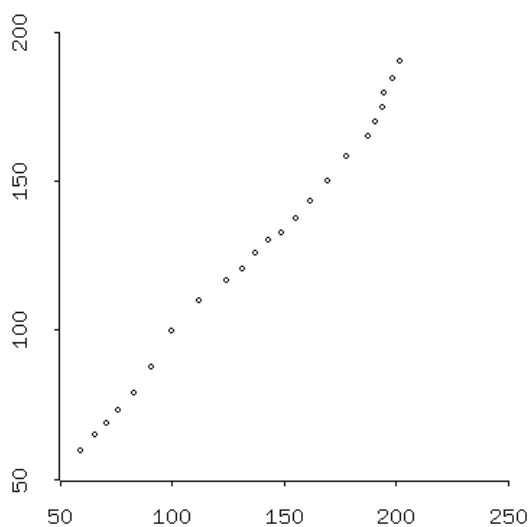


Figure 1.8: Scatterplot of the Canadian consumer price index (horizontal axis) vs. the U.S. consumer price index (vertical axis).

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

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

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

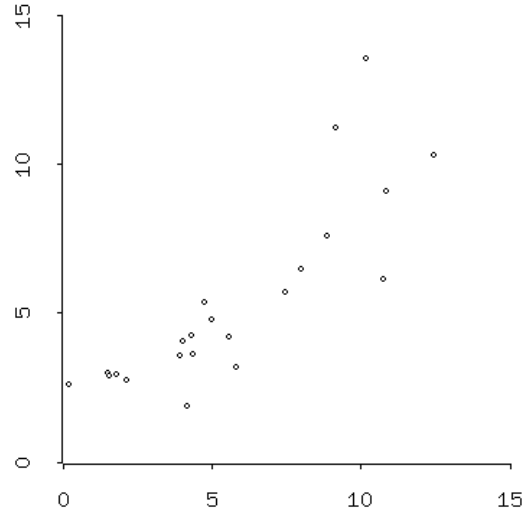


Figure 1.9: Scatterplot of the Canadian year-over-year inflation rate (horizontal axis) against the U.S. year-over-year inflation rate (vertical axis).

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

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} = r_{y,x} \quad (1.13)$$

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)

Index

- P*-value
 - diagrammatic illustration, 142
 - nature of, 142, 229
 - two sided test, 142
- α -risk
 - and power curve, 146
 - choice of, 138
 - level of μ at which
 - controlled, 138
 - nature of, 134
 - varies inversely with β -risk, 146
- β -risk
 - and power curve, 146
 - level of μ at which
 - controlled, 144
 - nature of, 134
 - varies inversely with α -risk, 146
- action limit or critical
 - value, 136
- actual vs. expected outcomes, 185
- AIDS test example, 52
- alternative hypothesis, 134
- analysis of variance (ANOVA)
 - chi-square distribution, 267
 - comparison of treatment
 - means, 265, 271
 - degrees of freedom, 266, 268
 - designed sampling
 - experiment, 264, 270
 - dummy variables in, 275
 - experimental units, 265
- factor
 - factor levels, 264, 271
 - meaning of, 264
 - in regression models, 261–263
 - mean square error, 266, 272
 - mean square for
 - treatment, 266, 272
 - nature of models, 261
 - observational sampling
 - experiment, 264, 271
 - randomized experiment, 270
 - response or dependent
 - variable, 264
 - similarity to tests of differences
 - between population
 - means, 268
 - single factor, 264
 - sum of squares
 - for error, 265, 271
 - sum of squares for
 - treatments, 265, 271
 - table, 266, 272
 - total sum of squares, 266, 272
 - treatments, 264
 - two-factor designed
 - experiment, 277
 - using F-distribution, 268, 273
 - using regression
 - analysis, 269, 274
- arithmetic mean, 19
- autocorrelation, 14

- basic events, 36
- basic outcome, 36, 39
- basic probability theorems, 54
- Bayes theorem, 49, 50
- Bernoulli process, 77
- bimodal distributions, 21
- binomial expansion, 82
- binomial probability distribution
 - binomial coefficient, 77, 82
 - definition, 76, 77
 - deriving, 80
 - mean of, 79
 - normal approximation to, 182
 - variance of, 79
- binomial probability function, 77
- binomial random variables
 - definition, 77
 - sum of, 79
- box plot, 11, 18

- census, 104
- central limit theorem
 - definition of, 113
 - implication of, 158
- central tendency
 - measures of, 18
- Chebyshev's inequality, 26
- chi-square distribution
 - assumptions underlying, 170
 - degrees of freedom, 170, 171
 - difference between two chi-square variables, 230
 - goodness of fit tests, 178, 179
 - in analysis of variance, 267
 - multinomial data, 182
 - plot of, 172
 - shape of, 172
 - source of, 170, 230
 - test of independence using, 187
- coefficient of
 - determination, 203, 228
- coefficient of correlation,
 - definition, 72
- coefficient of variation, 23
- comparison of two
 - population means
 - large sample, 155
 - small sample, 158
- comparison of two population variances, 173
- complementary event, 36, 39
- conditional probability, 45
- confidence coefficient, 118
- confidence interval
 - and sample size, 119
 - calculating, 118
 - correct, 118
 - for difference between two population means, 156
 - for difference between two population proportions, 162
 - for fitted (mean) value in regression analysis, 204
 - for intercept in simple regression, 209
 - for population proportion, 123
 - for population variance, 172
 - for predicted level in regression analysis, 207
 - for ratio of two variances, 175, 176
 - for regression
 - parameter, 208, 227
 - interpreting, 118
 - one-sided vs. two-sided, 122
 - precision of estimators, 117
 - using the t-distribution, 120
 - when sample size small, 119
- confidence limits, 118

- consistency, 116
- consumer price index
 - calculating, 14
 - data for four countries, 27
 - plots, 31
- contingency tables, 183
- continuous uniform probability distribution
 - mean and variance of, 88
 - density function, 87
 - nature of, 87
- correction for continuity, 94
- correlation
 - coefficient between standardised variables, 76
 - coefficient of, 30, 31
 - concept of, 30
 - of random variables, 70
 - statistical independence, 72
- count data
 - multi-dimensional, 184
 - one-dimensional, 180
- covariance
 - and statistical independence, 72
 - calculation of, 71
 - nature of, 28
 - of continuous random variables, 72
 - of discrete random variables, 70
 - of random variables, 70
- covariation, 70
- critical value or action
 - limit, 136
- cross-sectional data, 14
- cumulative probabilities, calculating, 84
- cumulative probability distribution, 64
- function, 64, 66
- data
 - cross-sectional, 14
 - multi-dimensional count, 184
 - one-dimensional count, 180
 - panel, 18
 - quantitative vs. qualitative, 7
 - skewed, 24
 - sorting, 8
 - time-series, 14
 - time-series vs. cross-sectional, 14
 - univariate vs. multivariate, 27
- data generating process, 42
- data point, 8
- degrees of freedom
 - chi-square distribution, 170
 - concept and meaning of, 170
 - F-distribution, 174
 - goodness of fit tests, 179
 - regression analysis, 201
 - t-distribution, 120
- dependence, statistical, 47
- descriptive vs. inferential statistics, 4
- deterministic relationship, 193
- discrete uniform distribution
 - mean of, 87
 - plot of, 87
 - variance of, 87
- discrete uniform random variable, 86
- distributions
 - bimodal, 21
 - hump-shaped, 21, 26
 - normal, 26
 - of sample mean, 106
- dummy variables
 - and constant term, 235

- as interaction terms, 235
 - for slope parameters, 235
 - nature of, 234
 - vs. separate regressions, 236
- economic theories, nature of, 2
- efficiency, 116
- element, of data set, 8
- estimating regression parameters
 - simple linear
 - regression, 197, 199
- estimators
 - alternative, 115
 - consistent, 116
 - efficient, 116
 - least squares, 199
 - properties of, 115
 - unbiased, 116
 - vs. estimates, 115
- event space, 37, 40
- events
 - basic, 36
 - complementary, 36, 39
 - intersection, 37, 39
 - nature of, 36, 40
 - null, 37, 40
 - simple, 36
- expectation operator, 67
- expected value
 - of continuous random variable, 69
 - of discrete random variable, 67
- exponential probability distribution
 - density function, 94
 - mean and variance of, 94
 - plots of, 94
 - relationship to poisson, 96
- F-distribution
 - assumptions underlying, 176
 - confidence intervals
 - using, 175, 176
 - degrees of freedom, 174, 230
 - hypothesis tests using, 176, 177
 - in analysis of variance, 268, 273
 - mean and variance of, 174
 - obtaining percentiles of, 175
 - plot of, 175
 - probability density function, 175
 - shape of, 175
 - source of, 174, 230
 - test of restrictions on
 - regression, 232, 233, 240
 - test of significance
 - of regression, 231
- forecasting, 254
- frequency distribution, 14, 20, 80
- game-show example, 60
- geometric mean, 20
- goodness of fit tests
 - actual vs. expected
 - frequencies, 178
 - degrees of freedom, 179
 - nature of, 177
 - using chi-square
 - distribution, 178
- histogram, 11
- hump-shaped distributions, 21, 26
- hypotheses
 - null vs. alternative, 134
 - one-sided vs. two-sided, 135
- hypothesis test
 - P -value, 142
 - diagrammatic illustration, 140
 - matched samples, 161
 - multinomial distribution, 181
 - of difference between
 - population means, 156

- of population variance, 173
- one-sided lower tail, 139
- one-sided upper tail, 139
- two-sided, 139
- hypothesis tests
 - goodness of fit, 177
 - using F-distribution, 176, 177
- independence
 - condition for statistical, 48, 185
 - of sample items, 110
 - statistical, 47, 48
 - tabular portrayal of, 188
 - test of, 184
- independently and identically distributed variables, 77
- inference
 - about population variance, 169
 - measuring reliability of, 6
 - nature of, 5, 35
- inflation rates, calculating, 14
- interquartile range, 11, 18, 19, 22
- intersection of events, 37, 39
- joint probability, 44, 45
- joint probability distribution, 50
- judgment sample, 104
- law of large numbers, 42
- least-squares estimation, 198, 199, 226, 227
- linear regression
 - nature of, 193
- low-power tests, 143
- marginal probability, 44
- matched samples, 160
- maximum, 18
- maximum likelihood
 - estimators, 130
 - likelihood function, 130
 - linear regression estimator, 199
 - method, 130
- mean
 - arithmetic, 19, 21
 - comparison of two population means, 155, 268
 - exact sampling distribution of, 108, 114
 - expected value, 68
 - geometric, 20, 21
 - more efficient estimator than median, 117
 - nature of, 19
 - sample vs. population, 20
 - trimmed, 21
- mean square error, 201
- median
 - less efficient estimator than mean, 117
 - measure of central tendency, 18
 - measure of position, 18, 19
 - middle observation, 8
- minimum, 18
- Minitab, 11
- modal class, 21
- mode, 21
- multicollinearity
 - dealing with, 241, 242
 - nature of, 240
- mutually exclusive, 36, 37
- mutually exhaustive, 36
- normal approximation to binomial distribution, 91, 93
- normal probability distribution
 - density function, 89
 - family of, 89
 - mean and variance of, 89
 - plots of, 91

- vs. hump-shaped, 26
- normal random variables,
 - sum of, 91
- null event, 37, 40
- null hypothesis, 134
- null set, 37

- observation, 8, 106
- odds ratio, 41, 42
- one-sided test
 - lower tail, 139
 - upper tail, 139
- outcomes
 - actual vs. expected, 185

- paired difference experiments, 159
- panel data, 18
- parameter vs. statistic, 104
- Pascal's triangle, 82
- percentiles, 8, 25
- point estimate, 114
- point estimator, 115
- poisson probability distribution
 - mean, 84
 - calculation of, 84
 - nature of, 83
 - plots of, 86
 - relationship to
 - exponential, 94, 96
 - variance, 84
- poisson probability function, 83
- poisson process, 86
- poisson random variable, 83
- poisson random variables,
 - sum of, 86
- population
 - concept of, 5, 35, 103
 - parameters, 104
- population proportion
 - estimates of, 123
 - pooled estimator, 163
 - tests of hypotheses about, 142
- population proportions, tests of
 - difference between, 162
- posterior probability, 51
- power curve, 146
- power of test
 - concept of, 143, 144
 - for hypothesis about
 - population proportion, 147
 - goodness of fit tests, 180
 - two-sided, 147
- prediction interval
 - calculating, 125
 - compared to confidence
 - interval, 126
- prior probability, 49, 51
- probabilistic relationship, 193
- probability
 - addition, 54
 - basic theorems, 54
 - complementation, 55
 - conditional, 45
 - joint, 44, 45
 - joint density function, 72
 - marginal, 44
 - multiplication, 55
 - nature of, 40
 - prior, 49
 - reliability of subjective
 - assignment, 44
- probability assignment
 - bivariate, 44
 - nature of, 41
 - objective, 42, 43
 - rules for, 40
 - subjective, 42, 43
- probability density function
 - F-distribution, 175
 - continuous uniform, 87

- exponential, 94
 - nature of, 64
- probability distribution
 - binomial, 76, 77, 180
 - chi-square, 170
 - conditional, 50
 - continuous uniform, 87
 - cumulative, 64
 - exponential, 94
 - F-distribution, 174
 - joint, 50, 184, 185
 - marginal, 50, 184
 - meaning of, 64
 - multinomial, 180, 181
 - normal, 89
 - of random variable, 68
 - of sample mean, 106
 - poisson, 94
 - posterior, 50, 51
 - prior, 49–51
 - standardised normal, 89
 - t-distribution, 120
 - uniform, 86
- probability function
 - binomial, 77
 - cumulative, 64, 66
 - poisson, 83
- probability mass function, 64
- probability sample, 104
- processes vs.
 - populations, 5, 103
- qualitative data, 7
- quantitative data, 7
- quartiles, 11
- random numbers
 - table of, 105, 106
- random sample, 104–106
- random trial, 36, 40, 63
- random trials, sequences
 - of, 77
- random variable
 - binomial, 77
 - definition of, 63
 - discrete uniform, 86
 - discrete vs. continuous, 63
 - normally distributed, 91
 - poisson, 83
- random variables
 - linear functions of, 73
 - sums and differences
 - of, 74
- range, 8, 18, 19, 22
- range, interquartile, 22
- regression analysis
 - R^2 , 203, 228
 - aptness of model, 210
 - autocorrelated
 - residuals, 212, 243
 - coefficient of
 - determination, 203, 228
 - confidence interval for \hat{Y} , 204
 - confidence interval for
 - predicted level, 207
 - confidence intervals for
 - parameters, 209
 - correcting residuals for
 - serial correlation, 245–248
 - degrees of freedom, 201, 262, 264
 - Durbin-Watson statistic, 244
 - error or residual sum
 - of squares, 202
 - fitted line, 197
 - forecasting, 254
 - heteroscedasticity, 211
 - left-out variables, 237
 - maximum likelihood
 - estimators, 199
 - mean square error, 201, 228, 262

- nature of, 193
- non-linear models, 249–251
- non-linearity, 210
- non-normality of error
 - term, 212
- normality of error term, 195
- prediction outside
 - experimental region, 254
- prediction outside sample
 - range, 254
- properties of error
 - term, 195, 197, 223
- properties of residuals, 200
- randomness of independent
 - variables, 213
- regression function, 196
- regression mean square, 263
- serially correlated
 - residuals, 212, 243
- statistical significance, 209
- sum of squares due to
 - regression, 202, 203, 262
- t-statistic, 229
- tests of hypotheses about
 - parameters, 209
- time-series models, 254
- total sum of squares, 202, 262
- unbiased and efficient
 - estimators, 199
- variance of error term, 201, 205
- variance of fitted (mean)
 - value, 204–206
- variance of predicted
 - level, 206, 207
- regression analysis (multiple)
 - dummy variables, 234
 - \bar{R}^2 , 228
 - basic model, 223
 - confidence intervals for
 - parameters, 227
 - constant term in, 229
 - dealing with
 - multicollinearity, 241, 242
 - degrees of freedom, 228
 - dummy variable for slope, 235
 - dummy variables, 270, 278
 - estimated coefficients not
 - statistically independent, 276
 - estimation of model, 225–227
 - F-test of
 - restrictions, 232, 233, 240
 - F-test of significance
 - of regression, 231
 - in matrix form, 224
 - in two-factor analysis of
 - variance, 277
 - interaction dummy
 - variables, 278
 - left-out variables, 237
 - multicollinearity, 240
 - non-linear interaction terms, 251
 - non-linear models, 251
 - second-order terms, 251
 - statistical tests, 227
 - sum of squares due to
 - regression, 228
 - testing for significance
 - of regression, 229
 - variance-covariance matrix
 - of coefficients, 276
 - variance-covariance matrix of
 - coefficients, 227
- regression analysis (simple)
 - R^2 , 203
 - calculating parameter
 - estimates, 200
 - coefficient of determination, 203
 - confidence interval
 - for intercept, 209
 - for slope coefficient, 208

- estimating parameters, 197, 199
 - linear model, 194
 - significance of slope parameter, 208, 209
 - variance of slope coefficient, 208
 - worked-out example, 213
- rejection probability, 144
- relationship between variables
 - deterministic, 193
 - linear, 194
 - probabilistic, 193
 - statistical, 193
- sample, 6, 35, 104
- sample mean
 - expectation of, 108
 - variance of, 110
- sample point, 36, 39
- sample points, enumerating, 64
- sample size
 - planning of, 124, 125
 - planning of to control α and β risks, 148–150
- sample space
 - and basic outcomes, 36
 - and event space, 40
 - union of all events, 37
 - univariate vs.
 - multivariate, 38
- sample statistics, 104
- sample statistics vs. population parameters, 106
- sample, matched, 160
- sample, representative, 6
- sampling a process, 106
- sampling error
 - interpretation of, 180
- sampling methods, 105, 106
- SAS, 11
- scatter plots, 28
- serial correlation, 14
- simple events, 36
- simple random sample, 104–106
- skewness, 19, 24
- skewness, measuring, 25
- sorting data, 8
- SPSS, 11
- standard deviation
 - calculation of, 23
 - definition of, 69
 - matched samples, 160
 - measure of variability, 22
 - of difference between sample means, 156
 - of estimated population proportion, 123
 - of sample mean, 114
 - of standardised random variables, 72
 - pooled or combined estimator, 157
- standard error of difference between sample means, 156
- standardised form
 - of continuous random variable, 70
 - of discrete random variable, 69
- standardised normal probability distribution, 89
- standardised values, 25, 26
- statistic vs. parameter, 104
- statistical decision rule
 - acceptance region, 136
 - critical values, 136
 - diagrammatic illustration, 140
 - nature of, 136
 - rejection region, 136
- statistical dependence, 47
- statistical independence
 - and conditional probability, 48

- checking for, 48
 - matched samples, 160
 - nature of, 47
- statistical test, 133
- sum of squares
 - restricted vs. unrestricted, 232
- t-distribution
 - compared to normal
 - distribution, 120, 121
 - degrees of freedom, 120
 - nature of, 120
 - when population non-normal, 122
- testable propositions, 3
- theories, truth of, 3
- theory of demand, 2
- time-series data, 14
- two-sided test, 139
- two-tailed hypothesis test, 140
- Type I error, 134
- Type II error, 134
- unbiasedness, 116
- uniform probability
 - distributions, 86
- union of events, 37
- universal event, 37
- variable
 - concept of, 8
 - dependent or response, 193
 - independent, explanatory
 - or predictor, 193
 - quantitative vs. qualitative, 8
- variance
 - calculation of, 23
 - matched samples, 160
 - measure of variability, 22
 - of continuous random
 - variable, 70
 - of difference between sample
 - means, 156
 - of discrete random
 - variable, 68, 69
 - of sample mean, 108
 - of sample proportion, 142
 - of sums and differences of
 - variables, 75
 - pooled or combined
 - estimator, 157
 - sample vs. population, 22
 - special case of
 - covariance, 30
- Venn diagram, 54
- XlispStat, 11, 83, 84, 172, 175, 227, 256, 276