EXTRACT: UNIVARIATE ANALYSIS

Chapters 1-4 and Appendix B

Analysis of Economics Data: An Introduction to Econometrics

Department of Economics, University of California - Davis.
Version 1.1

Contents

List of Figures

List of Tables

Preface

Motivation

Data analysis and data literacy are important and valuable skills in today's data age. Students taking economics and related majors are well placed to take advantage of the demands for these skills, as their major places greater weight on mathematical and statistical training than many other majors.

This book is suited to the following uses.

1. A true first course in regression analysis, the statistical method most used in analysis of economics data and the field of econometrics. The book's main goal is to reach this audience.

2. A helpful inexpensive adjunct to "Introductory Econometrics" courses that use a more advanced text.

3. A stand-alone reference for any more advanced data analysis course or economics field course that presumes a basic knowledge of linear regression.

How to Use this Book

The book takes a learning-by-doing approach. The key requirement is use of an econometrics or statistical package. For the particular statistical package that is chosen to use with this book, the instructor and student can easily work through each chapter using the datasets and computer code that are all available at the book website. **This is by far the best way to learn the material.** The book itself is limited to presenting key summary tables; few specific commands and consequent computer output are provided as these vary with the package used in instruction.

The website `cameron.econ.ucdavis.edu/aed` provides the datasets as a Stata version 11 dataset, readable by most other packages, and as a comma-separated values text file, readable by all packages. Datasets are referred to in the book using capital letters without any prefix or file extension. For example, the dataset called HOUSE in the text is available as file AED_HOUSE.DTA, a Stata version 11 dataset, and as file AED_HOUSE.DTA.csv, a comma-separated values text file.

The book is written as much as possible to be usable with any statistical package. An appendix provides key details on using **Stata**, the free packages **R** and **Gretl,** and the commercial econometrics package **Eviews**. The datasets can be read into these packages. Selected parts of the main text provide additional details on use of these packages. The spreadsheet programs **Excel** and **Google Sheets** can also be used, but are more limited. Appendix A summarizes key commands for the various statistical packages. The book website provides computer code for repeating the analysis in the book using **Stata, R** and **Gretl.**

The book includes over three hundred end-of-chapter exercises that are mainly learning-by-doing empirical exercises. Many use a wide range of datasets that can be obtained from the book website. Overhead slides for each chapter are also available at the book website.

Table 1: Book Chapters.

PART	Ch.	Title	Essentials
I: UNIVARIATE	1	Analysis of Economics Data	x
(Single Series)	2	Univariate Data Summary	x
	3	The Sample Mean	
	4	Statistical Inference for the Mean	x
BIVARIATE	5	Bivariate Data Summary	x
(Two series)	6	The Least Squares Estimator	
	7	Statistical Inference for Bivariate Regression	x
	8	Case Studies for Bivariate Regression	x
	9	Models with Natural Logarithms	
MULTIVARIATE	10	Data Summary with Multiple Regression	x
(Several series)	11	Statistical Inference for Multiple Regression	x
	12	Further Topics in Multiple Regression	x
	13	Case Studies for Multiple Regression	x
	14	Regression with Indicator Variables	x
	15	Regression with Transformed Variables	
FURTHER TOPICS	16	Checking the Model and Data	x
	17	Special Topics	
APPENDICES	A	Using Statistical Packages	x
	B	Some Essentials of Probability Theory	
	C	Properties of OLS and IV Estimators	
	D	Solutions to Selected Exercises	x
	E	Tables for Key Distributions	
	F	Further Reading	

Book Outline

Table 1 provides a summary of the book, which is divided into four parts.

1. Analysis of a single variable that covers the key parts of material presented in an introductory probability and statistics class.

2. Analysis of the relationship between two variables, y and x, with emphasis on bivariate linear regression.

3. Analysis of the relationship between y and several other variables, with emphasis on multiple linear regression.

4. Model and data checking and brief overviews of the most commonly-used methods beyond OLS: fixed effects and random effects for panel data and clustered data, logit and probit for binary dependent variable, several methods for causal inference, and time series regression.

Book appendices cover statistical packages, more advanced material on probability and estimation theory, solutions to odd-numbered exercises, and statistical tables.

Course Outline

The book is intended to be suitable for three different introductory courses.

1. An essentials course places less emphasis on motivating the distribution of the sample mean and regression coefficient estimates, skipping chapters 3 and 7, and places less emphasis on various extensions, skipping chapters 9, 12.2-12.8, 15 and 17.

2. A fast-paced ten-week quarter-long course with two lectures a week can cover the key material in Chapters 1-16, including one case study in each of the two case studies chapters.

3. A semester-long course could cover most of chapters 1-16 and selected parts of chapter 17.

This book is additionally written to be used as a supplement to courses that use regression. There is enormous heterogeneity in students knowledge of regression even after taking a first course in regression. This book, one available at low cost, can be used as a supplement to fill gaps.

For the Instructor

The book is written to be suitable for students with a wide range of mathematical and statistical backgrounds. The book presents methods in the main text with minimal use of mathematics - the optional Appendices B-C provide greater mathematical detail. In particular, to be accessible to a wide range of students the book is deliberately written at a lower level than the excellent leading texts by Wooldridge (2019) and by Stock and Watson (2018).

Some use of mathematics is nonetheless necessary. The main text presents formulas using summation notation. Changes in one variable with respect to another are generally presented using delta notation, though at times connections to derivatives are made for the benefit of those with a calculus background. Less-prepared students may find it possible to gloss over much of the mathematics. The emphasis of the book is on the interpretation of statistical output rather than dexterity with mathematical and statistical formulas.

The book does provide the essentials of probability, so that students understand the distinction between sample and population, the distinction between estimate and parameter, and the concepts of confidence intervals and hypothesis tests. Ideally students have taken a prior course on probability and statistical inference for the population mean based on the sample mean. My own experience is that even if students have taken such a course, many do not understand or do not recall statistical inference. Accordingly students benefit greatly from seeing the material a second time. Furthermore, some instructors may prefer to teach this course to students with no background in probability and statistics. The essentials of probability are covered briefly in Chapter 3, and are presented in more detail in Appendix B. Basic statistical inference on the sample mean is covered in detail in Chapters 3 and 4. Derivations of the properties of the OLS estimator are provided in Appendix C.

Regression results presented in this book are generally based on default standard errors in earlier chapters, as these are identical across statistical packages. Appropriate robust standard errors are used especially from chapter 12 on. Note that formulas for computing robust standard errors can vary across statistical packages due to different finite sample adjustments; see Chapter 12.1.9.

An individual chapter can be covered in one or two seventy-five minute lectures. To simplify the exposition of methods, my approach is to work with the one data example throughout a chapter, or even across several chapters in the case of summarizing individual earnings and modeling the sale price of a house. At the same time many additional datasets are introduced throughout the book, most notably in the case studies chapters and in the many exercises at the end of each chapter. Some data examples come from empirical research articles published in leading economics journals that were deliberately chosen in the belief that the associated articles would be intelligible to undergraduate students.

As already noted, the exercises at the end of each chapter are mainly learning-by-doing empirical exercises. Solutions to most odd-numbered exercises are given in Appendix D. It is easy for an instructor to make variations on these exercises that lead to different answers. Variations include using alternative datasets, using the same dataset with some observations dropped, or using the same dataset with different variables.

Version History

The book is available as a pdf and as a hard copy at a modest price through Amazon's Kindle Direct Publishing.

Version 1.0 is dated December 2021 and was released January 5, 2022.

Version 1.1 is dated February 2022 and was released late February, 2022. It corrects errors in Version 1.0 and in places provides some rewording for clarity. Section numbering is unchanged and pagination is essentially unchanged.

This Edition compared to Earlier Drafts

This edition is a revision of the unpublished 2015 version. The basic progression of topics is unchanged. Compared to that earlier version the initial chapters are simplified. More difficult concepts such as power of tests and many uses of natural logarithms are pushed to later chapters. And there is less repetition of material across chapters. The goal is to emphasize basic statistical analysis. For details see a document provided at the book website `cameron.econ.ucdavis.edu/aed`.

Acknowledgements

The late Jack Repcheck, book editor at a commercial publisher, provided great encouragement for the writing of this ultimately self-published book.

I was fortunate to receive an unusually strong undergraduate education in econometrics at The Australian National University and subsequent graduate training at Stanford University. A list of teachers to whom I am indebted would be very long. This book is my attempt to in turn introduce econometrics to students.

I have benefitted greatly from research coauthors, most notably Pravin Trivedi and Doug Miller, and from the strong empirical economics community at U.C. Davis and my fellow econometricians Òscar Jordà, Shu Shen and Takuya Ura.

Finally I thank the Department of Economics in supporting the course on which this book is based, a course that is both a terminal course for some of our majors and a pre-requisite course to additional optional econometrics courses for other Economics majors.

Chapter 1

Analysis of Economics Data

Statisticians specialize in data analysis, and offer courses that cover many of the statistical techniques in this book. This chapter summarizes the main statistical methods used in analyzing economics data.

1.1 Statistical Methods

The starting point of statistical analysis is a dataset, a collection of measurements that most often come from a survey or from an experiment.

Statistical analysis begins with a **summary** or **description** of what can be a bewilderingly large set of numbers. There are several standard statistics that are used to summarize features of the data such as the central tendency of the data and the spread of the data. For example, given data on annual income of a number of individuals we may compute the average income for these individuals. This is relatively straightforward.

Most data analyses seek to go further and use such summary measures to extrapolate to the world beyond the particular dataset at hand. For example, if the average annual income in a sample of forty Californians is $60,000, what can we say about the average income of all Californians? Or if forty tosses of a coin lead to 18 heads and 22 tails, can we conclude whether or not the coin is fair coin?

This extrapolation entails the much more challenging methods of **statistical inference** - inferring details of a population from the sample at hand. The two main statistical tools used are **confidence intervals** and **hypothesis tests**; much of the book is focused on learning how to use these tools in a variety of settings.

Additionally, steps should be taken to ensure that a sample is representative and obtained in such a way that the phenomena of interest can be measured sufficiently precisely. Like other books at this level, these issues are only briefly considered; they are covered in detail in separate statistics courses on survey sampling and experimental design.

In some special cases the dataset may be large enough and precise enough that there is no need to control for randomness due to sampling. For example, this would be the case if we had a complete census of the population or if we could toss the coin a million times. While very large

datasets are increasingly available, such as those from internet transactions, in typical economics applications one needs to control for uncertainty.

1.2 Types of Data

The discipline of statistics covers a wide range of data types and associated methods that are summarized in this section.

Within this wide range economists, and hence this book, focus on observational data on continuously measured variables analyzed using regression methods. An example is sale price data from a sample of individual house sales.

1.2.1 Economics Data

There are several types of data that may demand different statistical methods:

- Numerical data that are continuous.
- Numerical data that are discrete.
- Categorical data.

Economics data are usually **numerical data** that are naturally recorded and interpreted as numbers. Furthermore, they often potentially take so many different values that they are viewed to be **continuous numerical data**. Examples are individual annual income or national GDP.

Less often the data are **discrete numerical data** that take only integer values. Examples are the number of jobs held at a point in time or the number of patents awarded to a firm in a year.

Categorical data are an alternative to numerical data where the data are recorded as belonging to one of several possible categories, such as whether or not a person is employed. Such data may be coded as numbers, e.g. 1 if employed and 0 if not employed, but are not intrinsically numerical.

This book emphasizes the study of **economics data** that are continuous numerical data. Many examples will be provided, including leading relationships that are discussed in introductory microeconomics and macroeconomics courses. In many cases **economic theory** is used to guide in model selection. And in some cases economic data are used to test economic theory or to distinguish between economic theories.

Before analysis begins, data are often transformed to a more suitable form. For example, suppose interest lies in modelling improvements in living standards over time. A standard measure to use is the annual growth rate in real per capita gross domestic product (GDP). This entails transformations of the original GDP data to first adjust for inflation and population and to subsequently calculate year-to-year proportionate changes.

1.2.2 Observational Data

Economics data are most often **observational data**, meaning they are based on observations of actual behavior in an uncontrolled environment. A particular challenge of using observational data

is that while it is easy to detect a relationship between two data series, it can be very difficult to determine cause and effect.

By contrast many physical and biological sciences in particular use **experimental data** that are observations on the results of experiments which can be controlled by the investigator. Experimental methods and quasi-experimental methods are increasingly used in econometrics. For pedagogical reasons these methods are deferred to Chapter 17.5 which presents various methods to determine **causal** relationship, the goal of many econometrics studies.

1.2.3 Types of Data Collection

Distinction is made between several types of data collection:

- Cross-section data on different individuals at a point in time.
- Time-series data on the same quantity at different points of time.
- Panel (or longitudinal) data on the same individuals at different points of time.
- Repeated cross-section data on different individuals at different points in time

Cross-section data are data on different entities, such as individuals, households, firms or countries, collected at a common point in time. Examples are earnings of individuals and output of firms. Such data are most often used in **microeconomics**. Standard notation is to use the subscript i to denote the typical observation. The sample of size n is denoted $x_1, ..., x_n$ with i^{th} observation x_i.

Time-series data are data on the same quantity collected at different points in time. Examples are gross domestic product and the interest rate on a 13-week Treasury bill. Such data are most often used in **macroeconomics** and **finance**. Standard notation is to use the subscript t to denote the typical observation. The sample of size T is denoted $x_1, ..., x_T$ with t^{th} observation x_t. In this book we use subscript i as much as possible, but revert occasionally to subscript t in some time series settings. In particular, the one-period change in a time series variable is $\Delta x_t = x_t - x_{t-1}$.

Panel data or **longitudinal data** are data on the same individuals, such as firms or people or countries, where each individual is observed at several points in time. Examples include analysis of individual income over several years, and analysis of GDP in several countries over time. Such data are used in both microeconomics and macroeconomics. The typical observation is x_{it}, data for the i^{th} individual at time t.

Repeated cross-section data or **pooled data** are cross-section data collected in more than one time period, but in each time period different individuals are observed. Many surveys conducted on a regular basis sample different individuals in each survey.

The same basic statistical principles apply for all these types of data collections. However, each type of data collection also adds its own special considerations for statistical inference, such as computing confidence intervals, and for model specification. We focus on the simplest case of cross-section data.

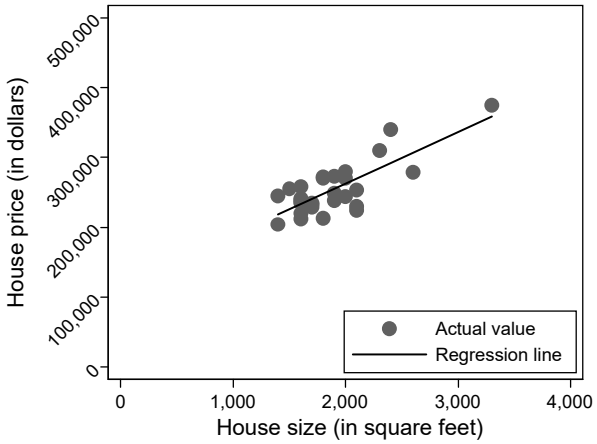

Figure 1.1: Linear regression example

1.3 Regression Analysis

Introductory statistics courses focus on data on a single variable considered in isolation, such as the individual income and coin toss examples. Some economic statistics such as the unemployment rate or the growth rate in real GDP or median earnings are also of interest on their own.

We first analyze **univariate data**, studying a single data series such as house price, with individual observations denoted x_i or denoted x_t. The treatment of univariate data is similar to that in an introductory statistics course.

Most economic data analysis, however, is focused on measuring the relationship between two or more variables. The statistical method used to measure such inter-relationships is called **regression analysis**. Most of this book studies regression analysis.

Bivariate data are data on two related data series, denoted y and x. For example, in Chapter 5 we consider the relationship between house price (in dollars) and house size (in square feet) for a sample of 29 house sales. Figure 1.1 presents a scatter plot of the data which suggests that, as expected, a higher price is associated with a higher price. Superimposed on this scatter plot is a line, called a **regression line**, that is the best fitting line for these data using a criterion given in Chapter 5. The regression line has slope coefficient equal to 74, approximately, so an increase in house size of one square foot is associated with an increase in house price of \$74.

Multivariate data methods consider three or more related series. Usually one of those variables, say y, is explained by several other variables, say x_1, x_2, ... using **multiple regression**. For example, we may consider the relationship between house price and several features of the house, such as size, number of bedrooms and lot size.

The term regression arises due to the phenomena of **regression towards the mean**. For example, consider the relationship between the height of a father (x) and the height of his son (y). If the father is of above average height, then the height of the son turns out to be on average also

of above average height, though not as high as that of the father. Similarly the son of a below average height father is on average below average, but not as below average as the father. More generally the term regression refers to fitting a model of y as a function of x.

Regressions may be used to measure how an outcome variable (y) **changes** as one of the regressors (x) changes, or may be used to **predict** the outcome variable (y) for a given level of the regressors (x).

1.4 Key Concepts

1. There are two aspects to statistical analysis of data: description and inferential statistics. The latter attempts to extrapolate from the sample to the population, often using confidence intervals and/or hypothesis tests.

2. The analysis of economics data uses a subset of statistical methods, most notably regression analysis for continuous numerical data, and emphasizes economic interpretation of economics-related data.

3. Economics data are usually observational rather than experimental. This makes it difficult to establish causal effects. For pedagogical reasons this complication is deferred to Chapter 17.5, though much econometrics research seeks to estimate causal relationships, even with observational data.

4. Cross-section data (denoted x_i) are data on different individuals at a point in time; time-series data (denoted x_t) are data on the same quantity at different points of time; panel data or longitudinal data (denoted x_{it}) are data on the same individuals at different points of time; repeated cross-section data are cross-section data collected in more than one time period, but in each time period different individuals are observed.

5. The book covers, in turn, univariate data (single series), bivariate data (two series), and multivariate data (several series).

6. The key method of this book is regression analysis.

7. Key Terms: Summary statistics; sample; population; statistical inference; continuous numerical data; discrete numerical data; categorical data; observational data; experimental data, cross-section data; time series data; panel data; longitudinal data; univariate data; bivariate data; multivariate data; regression analysis; bivariate regression; multiple regression.

1.5 Exercises

1. For each of the following examples state whether the data are numerical or categorical, and state whether the data are cross-section, time series, panel or repeated cross-section data.

   (a) Quarterly data on the level of U.S. new housing construction from 2000 to 2018.

   (b) Data on number of doctor visits in 2018 for a sample of 192 individuals.

   (c) Data on annual health expenditures for each U.S. state from 2000 to 2018.

   (d) Data on usual mode of transportation used to commute to work for a sample of 151 individuals.

   (e) Data on individual income from an annual survey from 2000 to 2018 that surveys different individuals each year.

2. For each of the following examples state whether the data are numerical or categorical, and state whether the data are cross-section, time series, panel or repeated cross-section data.

   (a) Data on annual health expenditures in 2018 for the U.S. by use of funds.

   (b) Data for several days on whether the Dow Jones Index at the close of trading was at a higher or lower value than at the close of trading the preceding trading day.

   (c) Data on sales this quarter by each of 23 sales representatives.

   (d) Data on the price of 1 gigabyte of computer disk storage each year from 1980 to 2018.

   (e) Annual earnings of 153 individuals in each of the years 2010 to 2018.

3. For each of the following state whether the data are observational or experimental.

   (a) Data on earnings for individuals some of whom chose to participate in a training program and some who did not.

   (b) Data on earnings for individuals some of whom were randomly assigned to a training program and some who were not.

   (c) Data on school outcomes for charter schools and for traditional schools.

4. For each of the following state whether or not statistical inference is being used.

   (a) Recording the number of heads in 40 coin tosses.

   (b) Determining whether a coin is likely to be fair on the basis of the number of heads in 40 coin tosses.

   (c) Recording the annual earnings of 125 randomly chosen people and calculating the average.

   (d) Recording the annual earnings of 125 randomly chosen people and then determining how likely it is that mean annual earnings in the population exceed $40,000.

Chapter 2

Univariate Data Summary

Univariate data are a single series of data that are observations on one variable. A numerical data example is annual earnings for each person in a sample of women. A categorical data example is expenditures in each of a number of categories.

The chapter begins with presentation of summary statistics for numerical data. These are useful both in their own right and as a tool for checking that there are no obvious errors in data entry, such as negative values for a variable that should be nonnegative.

The chapter then presents charts that can provide a very quick way to grasp the essential features of univariate data. The graphical methods used vary with the type of data. While the key charts are given, there are many possible variations. The graphs presented here are quite basic. Presentation quality graphics entail much more preparation and are beyond the scope of this book. Useful resources for graph styles are leading publications such as *The Economist*, *The New York Times* and *The Wall Street Journal* that frequently present charts for economics data.

Statistical inference, using data from a sample to make inferences about the population from which the data is sampled, is introduced in Chapter 3.

2.1 Summary Statistics for Numerical Data

Summary statistics or **descriptive statistics** provide a summary of data on a numerical variable.

Consider data on the annual earnings of a sample of 171 women who are 30 years of age in 2010, all of whom worked full-time (35 or more hours per week and 48 or more weeks per year). The data are in dataset EARNINGS.

Table 2.1: Summary statistics: Annual earnings of 30 year-old female full-time workers in 2010 (n=171).

Statistic	**Value**
Mean	41,413
Standard deviation	25,527
Minimum	1,050
Maximum	172,000
Number of Observations	171
Variance	651,630,282
Upper quartile (75th percentile)	50,000
Median (50th percentile)	36,000
Lower quartile (25th percentile)	25,000
Skewness	1.71
Kurtosis	7.32

Table 2.1 presents various summary statistics, rounded to the nearest dollar, that are explained in this section. A summary statistics command in a statistical package usually automatically reports at least the first five of these.

As a quick check of the data we note that there are 171 observations that range from \$1,050 to \$172,000. The minimum value is surprisingly low as it implies earnings of less than \$1 per hour for this sample of full-time workers. From a more detailed check of the original survey data, this individual was self-employed, so such a low value is possible. The second lowest sample value of annual earnings is \$9,000.

The observations for a **sample** of size n are denoted

$$x_1, x_2, ..., x_n.$$

Here x_1 is the first observation, x_2 is the second observation, ... and x_n is the n^{th} observation. For cross-section data the typical observation is the i^{th} observation, denoted x_i, while for time series data it is more customary to use the subscript t, in which case x_t is the t^{th} observation.

2.1.1 Central Tendency

A measure of **central tendency** or **central location** describes the center of the distribution of the data.

The most common measure is the **sample mean**, which is the arithmetic average of the data. For example, if the data take values 8, 3, 7 and 6, then the sample mean is $(8 + 3 + 7 + 6)/4 = 6$. More generally, for a sample of size n, the sample mean $\bar{x}$ is defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

A shorthand notation to present this formula, and many other formulas for summary statistics, uses **summation notation**. In general $\sum_{i=1}^{n} x_i$ denotes the sum of all the x_i from $i = 1$ to n, so

that

$$\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n.$$

In the current example, $x_1 = 8$, $x_2 = 3$, $x_3 = 7$ and $x_4 = 6$ so $\sum_{i=1}^{4} x_i = 8 + 3 + 7 + 6 = 24$. As a second example, if $x_i = 5+2i^2$ then $\sum_{i=1}^{n} x_i = \sum_{i=1}^{3}(5+2i^2) = (5+2\times1^2)+(5+2\times2^2)+(5+2\times3^2) = 7+13+23 = 43$. For constant c that does not vary with i, $\sum_{i=1}^{n} c = n\times c$ and $\sum_{i=1}^{n} cx_i = c\times\sum_{i=1}^{n} x_i$. And $\sum_{i=1}^{n}(cx_i + dy_i) = c\sum_{i=1}^{n} x_i + d\sum_{i=1}^{n} y_i$.

Using summation notation, the sample mean can be written as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The other leading estimate of central tendency is the **sample median**. The data are first ordered from the lowest value to the highest value, and the median is that value that divides the ordered data into two halves. This is directly obtained as the midpoint of the ordered data if there is an odd number of observations. For an even number of observations one chooses the average of the two observations in the middle. For sample 8, 3, 7 and 6, the ordered sample is 3, 6, 7 and 8, and the median equals $(6 + 7)/2 = 6.5$.

The median has the advantage of being more resistant to outliers than the mean. For example, mean income will change a lot if Bill Gates is included in the sample, whereas the median is essentially unchanged. And the median can be used if the highest values are top-coded, as is often the case for data on individual incomes. The mean is more often used, however, and this book focuses on statistical inference for the mean rather than the median.

A third measure, less commonly-used, is the **mid-range** which is the average of the smallest and largest values in the sample. This is very sensitive to outliers.

A fourth measure, the **mode**, is the most commonly occurring value. This is only useful when the data is discrete, or if the underlying data are intrinsically continuous but the observed data are greatly rounded, so that a given value can occur multiple times in the sample. Even then the mode is not necessarily a good measure of central tendency, especially if the distribution has more than one mode or if the distribution is asymmetric, defined below.

From Table 2.1, earnings are on average \$41,413. For these data with 171 observations the median is the 86^{th} of the ordered observations, and this equals \$36,000. So half the women in the sample earn less than \$36,000 and half earn more than this amount. Note that mean earnings in this example are substantially greater than median earnings. This is often the case for data on incomes, earnings and prices. The midrange is $(172000 + 1050)/2$ or \$86,525; this is much higher than the mean and is not particularly meaningful here. The mode, not given in Table 2.1, is \$25,000. In practice it is unlikely that any two women in this sample of size 171 would have exactly the same earnings. Here, due to rounding in reporting, ten women reported earnings of exactly \$25,000.

2.1.2 Quartiles, Deciles and Percentiles

The median is the point that equally divides an ordered sample. One can consider other divisions of the ordered sample.

The **lower quartile** is that point where one-quarter of the ordered sample lies below and three-quarters of the ordered sample lies above. The **upper quartile** is that point where three-quarters of the ordered sample lies below and one-quarter of the ordered sample lies above. For example, with 9 observations the upper quartile is the 3^{rd} highest, the median is the 5^{th} highest and the lower quartile is the 3^{rd} lowest. Adjustment, similar to that for the median with an even number of observations, is needed when more than one data point could be the quartile. The median is the middle quartile.

Even more detailed divisions of the sample are possible. **Percentiles** split the ordered sample into hundredths. The p^{th} percentile is the value for which p percent of the observed values are equal to or less than the value. The upper quartile, median, and lower quartile are, respectively, the 75^{th}, 50^{th}, and 25^{th} percentiles. **Deciles** split the ordered sample into tenths and are often used, for example, to summarize the distribution of individual income. A **quantile** is a percentile reported as a fraction of one rather than as a percentage. For example the .81 quantile is the 81^{st} percentile.

From Table 2.1 the lower and upper quartiles of earnings are, respectively, \$25,000 and \$50,000, so the middle half of 30 year-old female full-time workers earned between \$25,000 and \$50,000 per year.

2.1.3 Data Dispersion or Spread

A measure of **dispersion** describes the **spread** or **variability** of the data. The most commonly-used measure is the standard deviation.

An obvious measure to use is the average of the deviations $(x_i - \overline{x})$ of the data x_i from the sample mean $\overline{x}$. But this can be shown to always equal zero, because in sum the negative deviations exactly balance the positive deviations. Instead these deviations are squared, before averaging, to get the **sample variance** s^2 where

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The division by $(n-1)$ rather than the more obvious n is explained in Chapter 3.2. A simpler computational formula for the sample variance is $s^2 = \frac{1}{n-1}\{(\sum_{i=1}^{n} {x_i}^2) - n\bar{x}^2\}$; see exercise 26 which shows that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = (\sum_{i=1}^{n} x_i^2) - n\bar{x}^2$.

The sample variance is measured in units that differ from those in the original data, due to the squaring. For example, if the data were in units of dollars then the variance is in units of dollars squared. To return to the original units we take the square root. This yields the **sample standard deviation** s, defined as

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

If one sample has a larger sample standard deviation than another then we view the sample as having greater variability.

As an example consider the sample 8, 3, 7 and 6 which has $n = 4$ and $\bar{x} = 6$. Then the sample variance

$$s^2 = \frac{(8-6)^2 + (3-6)^2 + (7-6)^2 + (6-6)^2}{4-1} = \frac{14}{3} \simeq 4.667,$$

and the sample standard deviation is $s = \sqrt{14/3} \simeq 2.16$.

In some cases it is useful to measure the variability of the data relative to the mean, using the **coefficient of variation**

$$\text{CV} = \frac{s}{\bar{x}}.$$

This measure is useful for comparing the relative variability in a variable across groups. For example, the sample of 171 women had $\bar{x} = 41413$ and $s = 25527$, while a similar sample of 191 men had $\bar{x} = 52345$ and $s = 65035$. So men have greater variability in earnings than women, but this may potentially just be an artifact of men also having higher earnings on average. The coefficient of variation controls for the different means. Since CV= 65035/52345 = 1.24 for men exceeds CV= 0.62 for women, men have higher variability in earnings relative to mean earnings than do women.

Three other measures of variation in the data are the range, the interquartile range, and the average absolute deviation.

The **range** is the difference between the **maximum** and **minimum** values in the sample.

An **outlying observation**, or **outlier**, is an observation that is unusually large or small. The **interquartile range**, the difference between the upper quartile and the lower quartile, has the advantage of being more resistant to outliers than the standard deviation or the range.

The **average absolute deviation**, $\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$, is also more resistant to outliers than the standard deviation or the range.

From Table 2.1, the sample standard deviation of earnings is \$25,527. The coefficient of variation is 25,527/41,413 = 0.62, so the standard deviation of earnings is 62% of mean earnings. The range is (172000 − 1050) or \$170,950. The interquartile range is (50000 − 25000) or \$25,000.

For income and wealth data, interest lies in measuring relative shares and how these change over time. The **P90/P10 ratio** measures the ratio between the 90th percentile and the 10th percentile and is necessarily at least one. The P90/P10 ratio has the advantage that it does not require data on the richest individuals whose data may be top-coded for reasons of anonymity or unavailable due to survey nonresponse. If data on the entire distribution is available the **Gini coefficient** can be constructed. This measure ranges from zero with perfect equality to one if all goes to one individual. For wages and salaries of full-time full-year workers in the U.S. the P90/P10 ratio is in the range 5 to 6 while the Gini coefficient is around 0.35 to 0.40. Increases in these measures over time indicates rising **inequality**.

2.1.4 Interpretation of the Standard Deviation

The standard deviation is the commonly-used measure of variability, as is clear from subsequent chapters. It is not as easy to understand as the mean, which is simply the average.

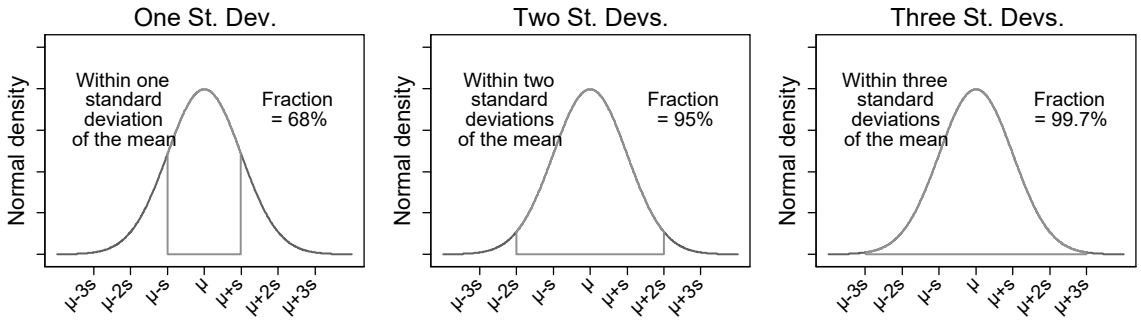

Figure 2.1: Normal distribution: Probability of being within one, two or three standard deviations of the mean.

A useful way to interpret the standard deviation is to use results for the normal distribution. For a random variable X that is normally distributed, the probability of being within one, two or three standard deviations of the mean is, respectively, 0.684, 0.955 and 0.997.

It follows that approximately two-thirds of the sample is within one standard deviation of the mean, 95% is within two standard deviations and 99.7% is within three standard deviations of the mean. These results can also provide an approximate guide for data that are not normally distributed.

This is illustrated in Figure 2.1, where the mean is denoted μ ("mu", the Greek letter for m), and the standard deviation is denoted σ ("sigma", the Greek letter for s).

Regardless of the actual distribution, a result called **Chebychev's inequality** implies that it always the case that at least three-quarters of a random sample is within two standard deviations of the mean, and at least eight-ninths is within three standard deviations of the mean.

As an example, consider the earnings data. These data have $\bar{x} = 41,413$ and $s = 25,527$, so the interval (15,886, 66,940) is within one standard deviation of the mean since, for example, $\bar{x} - s = 41,413 - 25,527 = 15,886$. For these data 77% of the observations are within this interval, compared to 68% predicted by the normal approximation. Similarly, for these data 96% of the observations are within two standard deviations of the mean, compared to 95% predicted by the normal approximation.

2.1.5 Box-and-Whisker Plot

A **box-and-whisker plot** or, more simply, a **box plot**, provides some of the key summary statistics for the data in a simple graphic.

All box-and-whisker plots give the lower quartile, median and upper quartile; these form the "box." Simple box-and-whisker plots additionally give the minimum and maximum; these form the "whiskers." More complicated box-and-whisker plots additionally plot outlying values. In that case

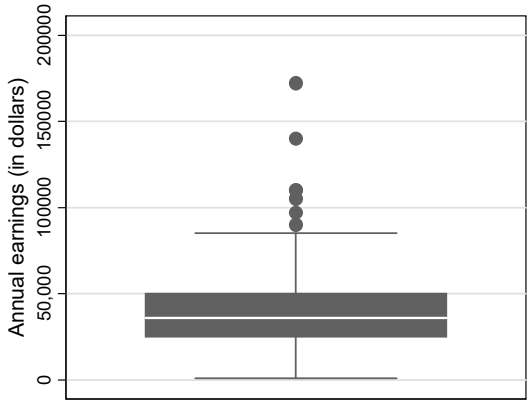

Figure 2.2: Box Plot: Annual earnings of 30 year-old female full-time workers in 2010.

the whiskers are data-determined lower and upper bounds, ones appropriate for data that are not too greatly dispersed, and outlying values are observations that exceed these bounds.

Figure 2.2 gives a box-and-whisker plot, of the more complicated form, for the earnings data. The solid shaded region ranges from the lower quartile of \$25,000 to the upper quartile of \$50,000. The solid white line within the shaded region is the median of \$36,000. The upper bar equals the upper quartile plus 1.5 times the inter-quartile range. Here this equals $50,000 + 1.5 \times 25,000$ or $\$87,500$. The six dots represent the six distinct values of earnings above \$87,500 in the sample. (In fact due to one duplicate there are seven observations in excess of \$87,500). The lower bar is the minimum sample value of \$1,050, as in this example the minimum exceeds the lower quartile minus 1.5 times the inter-quartile range.

The plot clearly shows the right-skewness of the data. The difference between the upper quartile and the median is much greater than the difference between the median and the lower quartile. And there are quite a few outlying sample points that take large values.

2.1.6 Symmetry

A **symmetric distribution** is one whose shape is the same when reflected around the median. The normal distribution is an example.

Positive skewed or **right-skewed** data have a much longer tail on the right. Most of the data are bunched on the left, but there is a continued presence of high values on the right. **Negative skewed** or **left-skewed** data have a much longer left tail.

Skewness can sometimes be visually detected. Figure 2.3 presents histograms for symmetric, right-skewed and left-skewed data.

A formal measure of asymmetry is the **skewness statistic**, calculated as a scale-free measure by normalizing by the standard deviation. Different statistical packages can use slightly different formulae in computing the skewness statistic. The simplest measure, used by most econometrics

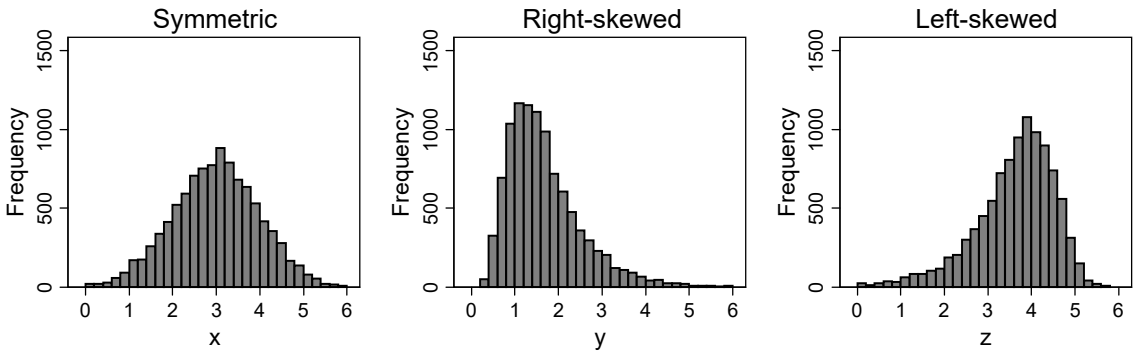

Figure 2.3: Histograms for symmetric, right-skewed and left-skewed data.

packages, is

$$\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}.$$

Some statistical packages, including Excel, multiply this measure by $\sqrt{n(n-1)}/(n-2)$, so Skew$=\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$. This adjustment is felt to lead to a better measure in small samples. In large samples the difference between the two measures disappears and both approximately equal $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$, where s is the sample standard deviation.

A zero value indicates symmetry since there is then no skewness. A positive value indicates positive or right-skewness and a negative value indicates negative skewness. There is no clear-cut rule for when data are highly skewed; a skewness measure in excess of one in absolute value indicates at least mild skewness. Note also that in small samples the skewness statistic is a less precise estimate of data skewness. For the three examples in Figure 2.3 the skewness measure equals, respectively, -0.04, 1.92, and -2.31.

Appreciable difference between the sample mean and sample median is also a sign of skewness. For right-skewed data the sample mean usually exceeds the sample median. For left-skewed data the sample mean usually is less than the sample median.

If economics data are skewed then they are usually right-skewed. For the earnings data, the histogram given below in Figure 2.4 clearly displays right skewness with a long right tail. For example, 94% of observations lie below the midrange of $\$85,475$ and only 6% lie above the midpoint. And, from Table 2.1, the mean of \$41,413 exceeds the median of \$36,000 and the skewness measure is 1.71.

Much economic analysis centers on modelling central tendencies. If skewness leads to an appreciable difference between the mean and the median, then both may be reported or, depending on the purpose, only one of the mean or median may be reported. For example, household income is right-skewed and government statistical reports emphasize median household income rather than mean household income. This reports the income of the household in the middle of the household

income distribution. At the same time, other government reports emphasize real per capita GDP which is a mean. This is in part because in that case the median cannot be computed, as data are not collected on GDP at the individual level. But it is also because in measuring the resources available to the economy interest lies in how much is available per person rather than how much is available to the median person.

2.1.7 Kurtosis

The **kurtosis statistic** measures the relative importance of observations in the tail of the distribution.

Different statistical packages can use slightly different formulae in computing the **kurtosis statistic**. The simplest measure, used by most econometrics packages, is

$$\text{Kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}.$$

Some statistical packages use an alternative measure of excess kurtosis that is felt to be better in small samples. One such measure, used by Excel, multiplies the kurtosis measure given above by $\frac{(n+1)(n-1)}{(n-2)(n-3)}$ and then computes excess kurtosis by subtracting $3\frac{(n-1)^2}{(n-2)(n-3)}$ rather than 3. In large samples the difference between different measures disappears and they approximately equal $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4$, where s is the sample standard deviation.

The normal distribution, with Kurt = 3, is often used as a benchmark, especially if the distribution is reasonably symmetric. **Excess kurtosis** measures kurtosis relative to the normal distribution, yielding

$$\text{ExcessKurt} \simeq \text{Kurt} - 3.$$

Positive excess kurtosis means that there is greater area in the tails than for the normal distribution with the same mean and variance, since $x_i - \bar{x}$ is raised to the fourth power. Some references state that the kurtosis statistic additionally measures the peakedness of the distribution, but this need not be the case especially if the distribution is asymmetric or bimodal.

The kurtosis measure is most often used for financial data. Fat tails are a feature of data on investment returns, and the greatest interest may lie in the tails since unusual events can provide the greatest opportunity to make a profit (or a loss).

From Table 2.1 the earnings data has kurtosis statistic (Kurt) of 7.32, substantially greater than 3, suggesting that the sample distribution has fatter tails than the normal distribution. For the three examples in Figure 2.3 the kurtosis measure equals, respectively, 3.04, 11.68, and 16.57.

2.2 Charts for Numerical Data

Histograms are the leading method for graphical inspection of cross-section numerical data. Histograms can also be useful for time series numerical data, provided that the data have been transformed to have little overall trend. As an example, histograms may be useful for summarizing real GDP growth rates or price inflation rates over time, but are of very limited use for describing GDP and price levels which trend upward over time.

Table 2.2: Frequencies: Individual annual earnings in bins of width 15,000.

Range (or bin)	Frequency	Relative frequency (%)
0-14,999	12	7.0
15,000-29,999	53	31.0
30,000-44,999	52	30.4
45,000-59,999	20	11.7
60,000-74,999	11	6.4
75,000-89,999	16	9.4
90,000-104,999	2	1.2
105,000-119,999	3	1.8
120,000-134,999	0	0.0
135,000-149,999	1	0.6
150,000-164,999	0	0.0
165,000-180,000	1	0.6

2.2.1 Histograms

Table 2.2 summarizes the earnings data grouped into intervals of width \$15,000. Each interval is called a **bin**; here there are 13 bins, each of equal **bin width** of \$15,000. The **frequency** is the number of observations that fall into a given bin, and the **relative frequency** is the proportion (or percentage) that fall into a given bin. For example, 53 observations or 31.0% of the sample have earnings between \$15,000 and \$29,999.

A **histogram** is a graph of the frequency distribution of the data where, for continuous data, the data are first grouped into bins. The horizontal axis has the values of the variable, while there are two variations for the vertical axis. One variation has the frequencies in each bin on the vertical axis. A second variation has the density (the relative frequency divided by the bin width) on the vertical axis – then the shaded area of the histogram has area one.

The first panel of Figure 2.4 presents the histogram corresponding to Table 2.2, with frequencies on the vertical axis. The second panel of Figure 2.4 provides a more detailed histogram that groups the data over a narrower range, with bin width \$7,500.

The histogram varies with the number of bins, with a trade-off between few bins providing not enough detail and too many bins yielding a histogram that is very choppy. Given n observations, a common default choice for the number of bins is $\sqrt{n}$. The class intervals are then of width approximately equal to the highest value minus the lowest value divided by the number of bins, with possible modification for unusually small or large observations. For $n = 171$ this yields 13 bins of equal width $(172000 - 1050)/13 = 13,150$. Table 2.2 and the first panel of Figure 2.4 instead round these defaults to 12 bins of equal width \$15,000 with a start value of \$0. The second panel of Figure 2.4 doubles the number of bins, by halving the bin width to \$7,500.

A variation on a histogram, one that gives more detail on the actual values taken by the data, is a **stem and leaf display**. This splits each data point into leading digits, called a stem, and remaining digits, called a leaf. For example for the earnings data the ten thousands may be the stem and the remaining digits the leaf. The data are then presented in tabular form where each

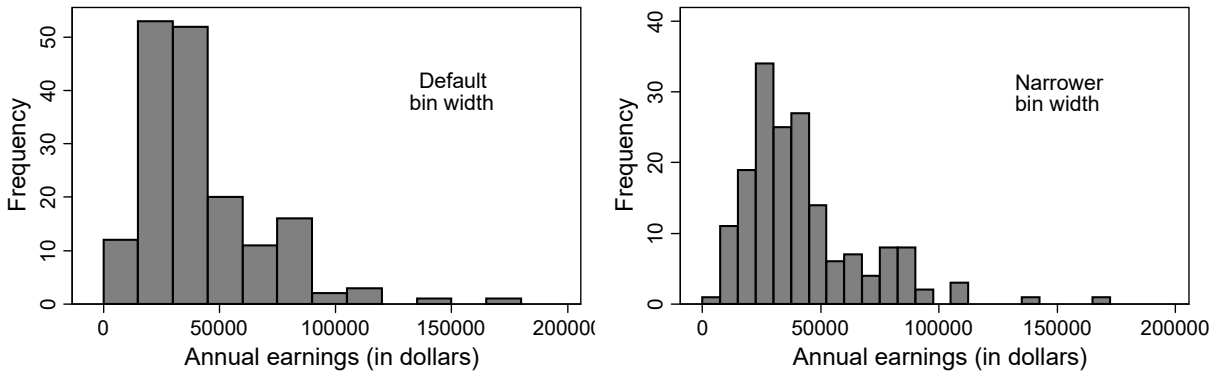

Figure 2.4: Histogram: Individual annual earnings with two different bin widths

row corresponds to a stem value and has first column the stem value and the second column the leaf values for that column.

Histograms can be used for numerical data that is either discrete or continuous. For **discrete data** that take low values, such as the number of different jobs held by a person during the year, each distinct value forms a bin so the bin width is one.

2.2.2 Smoothed Histograms (Kernel density estimate)

Data that take many different values, such as earnings data, have an underlying continuous probability density function rather than a discrete probability mass function. A classic example is the normal distribution which has a bell-shaped density. Probabilities are determined by areas under the curve and the total area under a density is one. It is then natural to directly estimate the density, using a smoothed histogram.

A **smoothed histogram** smooths the histogram in two ways. First, it uses rolling bins (or **windows**) that are overlapping rather than distinct. Second, in counting the fraction of the sample within each bin it gives more weight to observations that are closest to the center of the window and less to those near the ends of the window.

The smoothed histogram varies greatly with choice of **window width**, just as the histogram varies with the bin width. It varies less with the weights that are used. Different statistical packages may have different rules for choosing the default window width, and use different weights, called **kernel weights**, leading to different smoothed histograms.

The most commonly-used smoothed histogram is a **kernel density estimate**. Two kernel density estimates for the earnings data are presented in Figure 2.5. The first panel uses a window width close to the statistical package's default width, while the second is smoother as it uses a window width that is twice as large. The kernel density estimate is not bell-shaped, implying that the data are not normally distributed, and appears to be right-skewed. The vertical axis is scaled so that the area under the curve equals one.

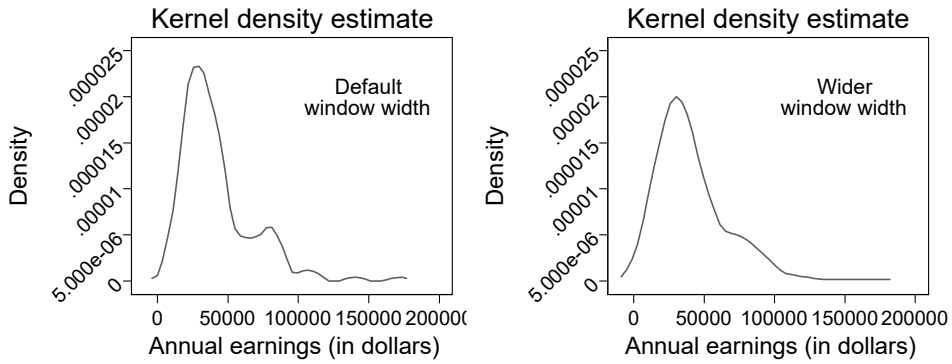

Figure 2.5: Smoothed histogram: kernel density estimate for individual annual earnings with two different window widths

2.2.3 Histograms and Smoothed Histograms using a Statistical Package

Histograms can be obtained using the `histogram` command in Stata, `hist` function in R, `hist` function in Gretl, and `distplot hist` command in Eviews. The default number of bins depends on the number of observations; as an alternative the number of bins can be specified.

Kernel density estimates can be obtained using the `kdensity` command in Stata, `density` function in R, `kdensity` function in Gretl and `distplot kernel` command in Eviews. The key option to consider changing from the default is the window width. More specialized is to change the kernel weight function from the default.

2.2.4 Line Charts for Ordered Data

A **line chart** plots the successive values $x_1, x_2, ...$ of the data against the successive index values 1,2,...

The leading application is to time series data that are ordered by time. This leads to graphs that plot the variable of interest against time.

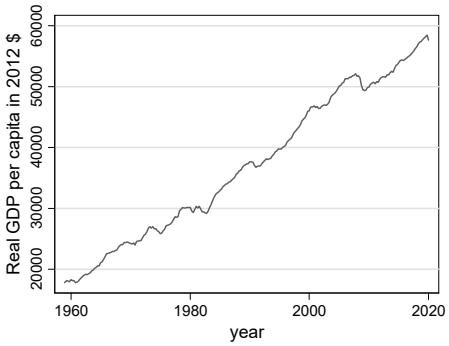

Figure 2.6: Line Chart: Real GDP per capita in U.S. (in 2012 dollars).

Figure 2.6 presents a line chart of quarterly data from 1959 to 2019 for U.S. real gross domestic product (GDP) per capita in constant 2012 dollars. The data are in dataset REALGDPPC. The line chart clearly indicates great improvement in living standards, with per capita real GDP tripling over the sixty years.

More generally, line charts can be useful whenever there is a natural ordering of the observations. For example, given data on test scores for 31 students it may be helpful to arrange the scores in descending order and produce a line chart of test score against student rank.

2.3 Charts for Numerical Data by Category

Standard charts for numerical data by category include bar charts, pie charts and, for geographic location categories, spatial maps.

2.3.1 Bar Charts

Consider U.S. health expenditures in 2018 of $3,653 billion (18% of GDP), broken into its main subcomponents. The data in dataset HEALTHCATEGORIES are completely listed in Table 2.3.

A **bar chart** provides a bar for each category where the length of the bar is determined by the category value, here expenditures on the category of health.

A **column chart** or **vertical bar chart** puts the values on the vertical axis and the category on the horizontal axis. A **horizontal bar chart** instead puts the category on the vertical axis and the value on the horizontal axis. The choice of which to use is determined in part by whether one wants a short and wide chart, in which case a column chart is most often used, or a tall and narrow chart, in which case a horizontal bar chart is most often used.

Table 2.3: Numerical data by category: U.S. health expenditures in 2018 in billions of dollars

Category	Amount ($ billions)
Hospital Care	1192
Physician and Clinical Services	726
Dental	136
Other Professional	104
Other Health and Personal	192
Home Health Care	102
Nursing Care	169
Drugs and Supplies (Retail Sales)	456
Government Administration	48
Net Cost of Health Insurance	259
Government Public Health	94
Noncommercial Research	53
Structures and Equipment	122

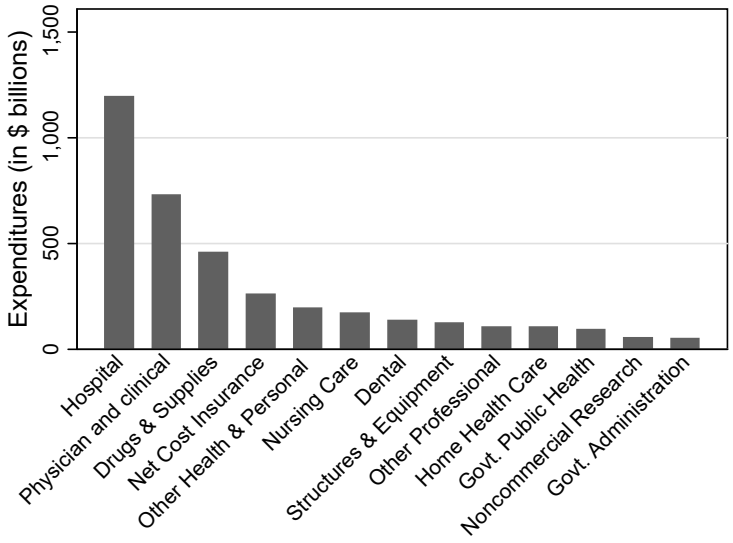

Figure 2.7: Column chart: U.S. health expenditures in 2018 (in billions of dollars)

Figure 2.7 presents a column chart for U.S. health expenditure data in 2018. This chart, ordered by size of category, makes it clear that hospital and physician expenditures are by far the largest components of total health expenditures.

Bar charts can also be used for larger datasets by first forming different categories according to what range of values the numerical data falls into. For example, one might group years of completed schooling into 0-11 (less than high school), 12 (high school graduate), 13-15 (some college), 16 (4-

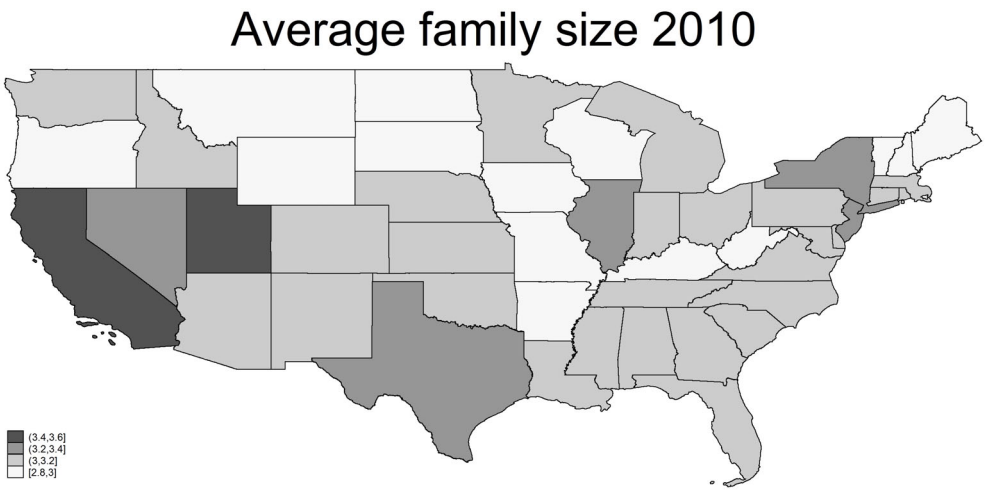

Figure 2.8: Spatial Map: Average family size in each U.S. state in 2010.

year college graduate) and 17 and above (postgraduate). Then give a bar chart of average income by schooling category. A histogram is just a column chart of frequencies plotted against the class boundaries.

2.3.2 Spatial Map

A **spatial map** for data that varies by geographic location plots the data against a geographic map.

As an example, consider average family size in each U.S. state in 2010 which ranges from 2.83 in Maine to 3.56. Figure 2.8 shows the average family in size in each state with darker shades corresponding to larger family size. The figure shows that southwest states tend to have larger families while north central states have smaller families. Spatial maps require more specialized software than that needed for the other graphs presented in this chapter.

2.4 Summary and Charts for Categorical Data

As an example of intrinsically categorical data consider choice of fishing site for a sample of 1,182 fishers given in dataset FISHING. There are four possibilities – fishing may be from a beach, pier, charter boat or private boat.

2.4.1 Data Summary using Tabulation

The fishing site data may be recorded as text, such as "beach", "pier", "charter" and "private". Or they may be recorded as numbers, such as 1, 2, 3 and 4. But even in the latter case the possibilities are intrinsically categorical. Furthermore there is no natural ordering of the categories.

For such data it is meaningless to compute summary statistics such as the sample mean. Instead the data are summarized using a **tabulation** of the **frequencies** for each category. For the fishing site data this is given in Table 2.4. It is clear that more people fished from a boat (private or charter) than from the shore (beach or pier).

Table 2.4: Categorical data: Frequencies at each fishing site

Category	Frequency	Relative frequency (%)
Beach	134	11.34
Pier	178	15.06
Private Boat	418	35.36
Charter Boat	452	38.24

2.4.2 Pie Charts

A **pie chart** splits a circle into slices, where the area of each slice corresponds to the relative frequency of observations in each category. Pie charts are most useful for visually representing each categories' share of the total, provided there are not too many categories.

Figure 2.9 presents a pie chart using the fishing site data. Again this makes clear that the largest categories are charter boat and private boat fishing.

The health expenditure data of Chapter 2.3 could be presented using a pie chart, if one was interested in the shares of each category of health spending. But this would be difficult to read as there are too many categories. Instead it would be best to aggregate the smallest categories. For example, one might use hospital, physician, drugs and supplies, and all other categories combined.

2.5 Data Transformation

Continuous numerical data are often transformed before analysis.

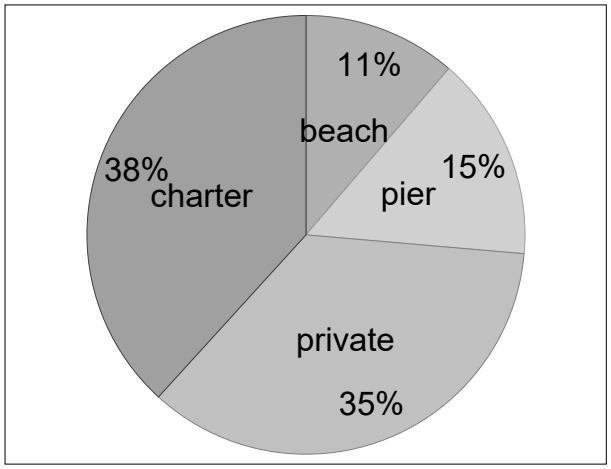

Figure 2.9: Pie Chart: Fishing site used

2.5.1 Natural Logarithms

Many cross-section data sets can be right-skewed. For example, data on income or wages of a sample of individuals are often right-skewed. The **natural logarithm** transformation can lead to a transformed data series that is more symmetrically distributed. It reduces especially large outlying values. Chapter 9.5 presents many uses of the natural logarithm.

The left panel of Figure 2.10 presents a histogram of earnings of female full-time workers aged 30 in 2010, along with a kernel density estimate, using data from dataset EARNINGS introduced at the start of this chapter. The data are clearly right-skewed. The second panel of Figure 2.10 shows the histogram after transformation to natural logarithms. The second panel histogram is close to symmetric, aside from one very small value (the sample included an observation with unusually low annual earnings of $1,050 and corresponding low natural logarithm of 6.96), and is approximately normally distributed. In both cases the vertical axes are scaled so that the areas under the histograms and the kernel density estimates equals one.

If a variable x is such that $\ln x$ is normally distributed, then x itself is said to follow the **lognormal distribution**.

2.5.2 Standardized Scores (z-scores)

A **standardized score** is obtained by subtracting the mean and dividing by the sample standard deviation. Thus

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, ..., n,$$

where x_i is the original value, $\bar{x}$ is the sample mean and s is the sample standard deviation. The resulting score has sample mean zero and sample standard deviation one.

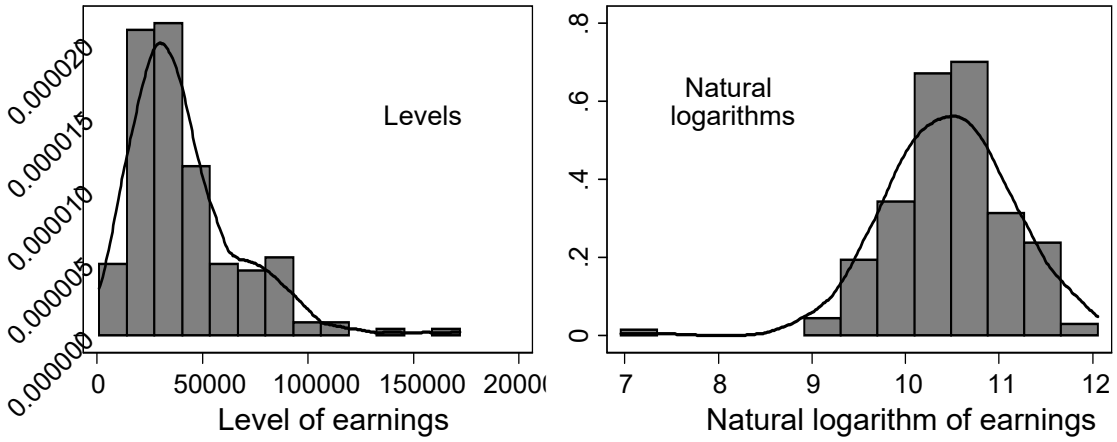

Figure 2.10: Levels and natural logarithms: Histograms and kernel density estimates for individual annual earnings

A standardized score is often called a **z-score** as its distribution may be well approximated by a standard normal distribution, which also has mean 0 and variance 1. Note also that the symmetry and kurtosis statistics approximately equal the sample averages of the standardized scores for each observation raised to, respectively, the third and fourth power.

A standardized score is immediately interpretable – a one unit increase in z_i equals a one standard deviation increase in the original score x_i.

Standardized scores are useful for comparing data series that are scaled differently. For example, suppose we wish to compare student performance on two tests with different total points or of different difficulty, so that the sample means and standard deviations differ across the tests. Then we compare the sample values of the standardized scores $z_{1i} = (x_{1i} - \bar{x}_1)/s_1$ and $z_{2i} = (x_{2i} - \bar{x}_2)/s_2$, where the subscripts 1 and 2 denote the first and second tests.

2.6 Data Transformations for Time Series Data

In this section we present some commonly-used transformations for time series data.

2.6.1 Moving Averages

A **moving average** or **rolling average** smooths data by taking the average of observations in several successive periods. This is especially useful for data that bounce around from period to period; averaging can smooth the data. Visual analysis of long-term trends in the data are easier to see, since period-to-period variation is reduced.

A **simple moving average** averages the current and immediate past observations. For exam-

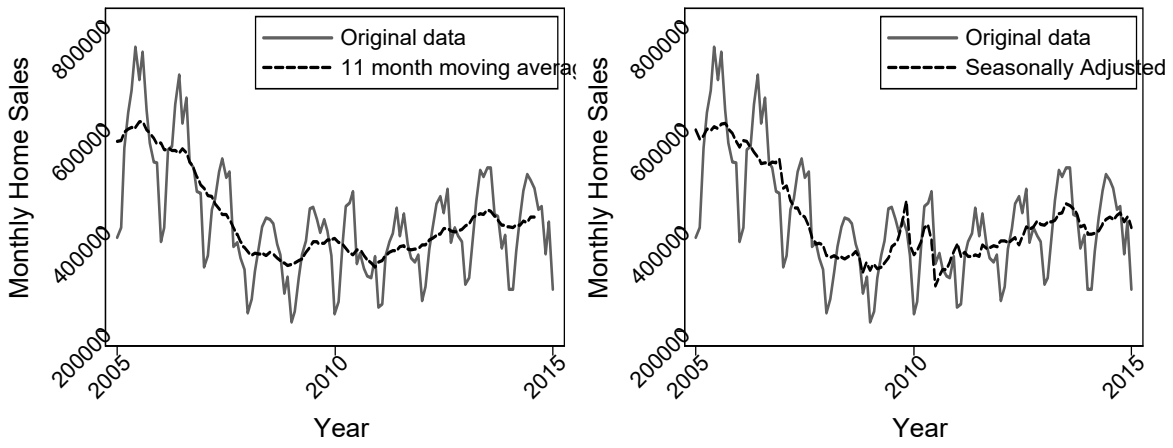

Figure 2.11: Moving average and seasonal adjustment smoothing: Monthly sales of existing homes

ple, a five-period moving average takes the average of the data over the current and preceding four periods, or $(x_t + x_{t-1} + x_{t-2} + x_{t-3} + x_{t-4})/5$.

If instead the current observation appears in the middle, then the moving average is a **centered moving average.** For example, a centered five-period moving average takes the average of the data two periods ago, one period ago, this period, next period, and the period after that, or $(x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2})/5$. The centered moving average at time t has the disadvantage that it is not immediately available at time t as its computation also uses data from some future time periods.

A moving average can be used for several reasons, including reducing random noise in the data, smoothing out business cycle variation, and smoothing out seasonal variation.

As an example we consider smoothing out seasonal variation in U.S. monthly data from 2005 to 2014 on sales of existing homes, compiled by the National Association of Realtors. The data are in dataset MONTHLYHOMESALES.

The first panel of Figure 2.11 plots the original data along with an eleven-month centered moving average. The original data are relatively variable within a year, with a low point in January and February and a peak in summer. The data also indicate the large decrease accompanying the global financial crisis, with a thirty percent decline compared to 2005. The moving average smooths the data considerably. Note that centering the moving average comes at the expense of it not being computable for the most recent months as it requires data in future months.

2.6.2 Seasonal Adjustment

For data that fluctuate within the year due to seasonal influences specific methods have been developed to smooth out the seasonal variation that hopefully are better than simply using a moving average.

Seasonal adjustment smooths data to control for seasonal variation in the data. For example, monthly data are decreased in months that have relatively high values every year and are increased in months that have unusually low values every year.

The second panel of Figure 2.11 presents as a dashed line the published seasonally adjusted data for the existing homes sales series, using the widely-used X-11-ARIMA seasonal adjustment program developed by the U.S. Census Bureau. The seasonal adjusted series is much smoother than the original, and essentially eliminates the seasonal variation.

Many macroeconomics series are released as **seasonally adjusted data**. Analysts interpreting these data should be aware that there is no indisputable best way to seasonally adjust.

2.6.3 Real and Nominal Data

Economics data are often measured in dollars. Any meaningful interpretation of these data over time requires conversion to the purchasing power of a dollar in a benchmark year. The original data are called **nominal** data, measured in **current dollars**. Thus 1990 data are measured in 1990 dollars, 1991 data are measured in 1991 dollars, and so on. The data after conversion are called **real data**, measured in **constant dollars**. Then data in various years are reported in dollars of a given year, say 2012 dollars for example. Similar conversion using exchange rates or purchasing power parity indexes is needed to compare nominal data across countries with different currencies.

There is no perfect way to create a price index (or a quantity index) when both prices and quantities of the various goods and services that are components of the index change over time. The leading published indexes use methods that control partially for this problem.

As an example of use of real data rather than nominal data, consider U.S. Gross Domestic Product (GDP), the standard measure of the economy's total output. The solid line in the first panel of Figure 2.12 plots quarterly data on nominal GDP from 1959 to the first quarter of 2020. The data in dataset REALGDPPC are seasonally adjusted quarterly data, annualized by multiplying by four. Nominal GDP has increased 42 times, from \$510 billion to \$21,500 billion. The fall in GDP in the recession of 2007-2009 is most clearly visible.

Part of this large increase in nominal GDP reflects price inflation – a dollar in 1959 had much more purchasing power than a dollar in 2020. The conversion from nominal to real data is done by using a **price index,** which measures prices relative to a value of 100 in a base year. Here we use the GDP chain-type price index, normalized to equal 100 in 2012. The index in the first quarter of 1959 was 16.347, so a 1959 dollar was worth $100/16.347 = \$6.12$ in 2012 dollars, and 1959 first quarter nominal GDP of \$510.33 billion was worth \$3,121 billion ($510.33 \times 100/16.347$) in 2012 dollars. Similarly in the first quarter of 2020 the index was 113.502 and nominal GDP of \$21,539 billion was worth \$18,977 billion in 2012 dollars. Table 2.5 summarizes these calculations.

The dashed line in the first panel of Figure 2.12 plots real GDP from 1959 to 2020, measured in 2012 dollars. Real GDP increased 6.1 times, from \$3,121 billion to \$18,977 billion. This is still a substantial increase, but it is much less than the 42 times increase in nominal GDP. The difference is due to a 6.9 times (113.502/16.347) increase in prices over this period, leading to real GDP rising $42/6.9 = 6.1$ times.

The recessions in 1973-74, 1980, 1982 and 1991 become more pronounced using real GDP data, with more pronounced dips due to eliminating increases in nominal GDP that occur due to price

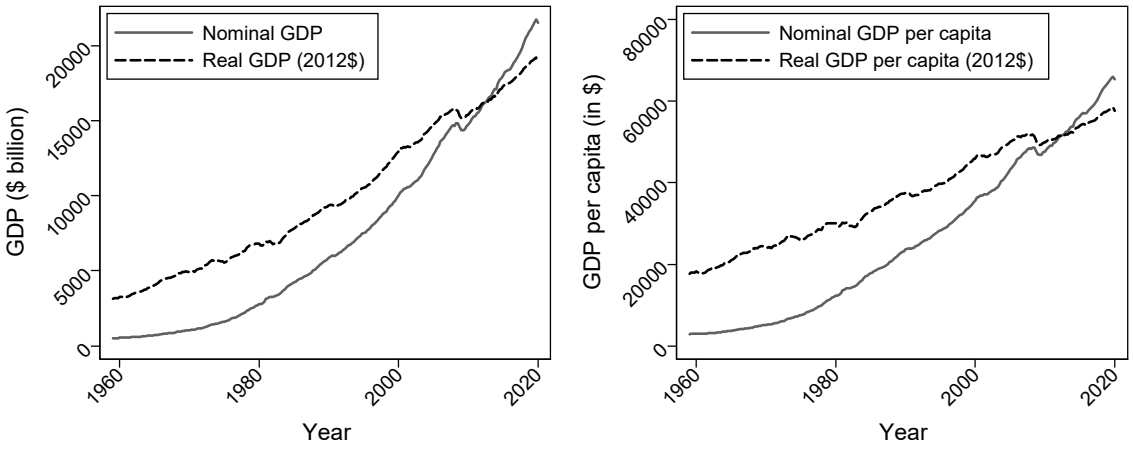

Figure 2.12: Nominal and real data: U.S. GDP and per capita GDP in current dollars and in 2012 dollars.

Table 2.5: Nominal and real data: U.S. GDP example

Time Period	**Nominal Value**	**Index**	**Real Value**
	Current $ billions	2009=100	2009 $ billions
1959 Q1	510.33	16.347	510.33×100/16.347=3,121
2020 Q1	21,539	113.502	21,539×100/113.502=18,977

inflation.

What if we used a year different from 2012 as the base year? Then the price index will differ and real GDP will differ as it is no longer measured in 2012 dollars. But the resultant proportionate changes will be unchanged, with a 6.9 times rise in prices and real GDP becoming 6.1 times larger over the 60 years.

2.6.4 Per Capita Data

Per capita data are data formed from an original series by dividing by the size of the population.

In some cases interest lies in aggregate data and in some cases per capita data. For example, to compare the size of the economy over time use real aggregate GDP, but to compare living standards over time use real per capita GDP.

As already noted, real GDP increased 6.1 times from 1959 to 2020. But the U.S. population is 1.87 times larger, with increase from 176 million to 329 million. Thus real per capita GDP has grown about 3.3 times, as $6.1/1.87 = 3.3$. This is illustrated in the second panel of Figure 2.12, with increase in real GDP per capita from $17,700 to $57,600. This is still an appreciable improvement

over time, and is about 2.0 percent per annum, since $1.020^{60} \simeq 3.3$. But it is nowhere near as large as the initial starting point of a 42 times increase in nominal GDP.

Often U.S. real GDP growth is compared to that in western Europe or Japan. U.S. growth in real GDP is higher, but so too is its population growth. In fact the growth rate for per capita real GDP in the U.S. is similar to that in western Europe and Japan.

2.6.5 Growth Rates and Percentage Changes

If interest lies in changes over time it can be convenient to transform to **percentage changes** or **growth rates**. For example, to analyze changes in living standards we consider percentage changes in real per capita GDP over time.

The **one-period percentage change** in x_t is calculated as

$$\text{Percentage change in } x_t \ = 100 \times \frac{x_t - x_{t-1}}{x_{t-1}}.$$

In many cases this is converted to an **annualized rate**. For example, for quarterly data the quarterly change multiplied by four gives the annualized quarterly change. Alternatively, for quarterly data one can instead compute a four-period percentage change, $100 \times (x_t - x_{t-4})/x_{t-4}$, which also expresses the change as an annual rate. For data that are not already seasonally adjusted, this latter method can smooth out quarterly seasonal fluctuations. Similarly for monthly data we may use $100 \times (x_t - x_{t-12})/x_{t-12}$.

Potential confusion can arise when statements are made about changes in growth rates or interest rates. For example, suppose the growth rate increases from 3 percent in one year to 5 percent the next year. It is misleading to call this a 2 percent increase in the growth rate, since this literally means that an increase in the growth rate from 3.0 percent to $3.0 \times 1.02 = 3.06$ percent. Instead the correct term to use is that the growth rate increased by two **percentage points**. Very small changes are described in **basis points**, where a basis point is one-hundredth of a percentage point. For example, an increase from 3.0 percent to 3.15 percent is an increase of fifteen basis points.

An alternative calculation method for computing approximate percentage changes is to use

$$\text{Percentage change in } x_t \ \simeq 100 \times (\ln x_t - \ln x_{t-1}).$$

This result uses the calculus result that $d \ln x/dx = 1/x$, so $d \ln x = dx/x$. Thus $\Delta \ln x \simeq \Delta x/x$ or the change in $\ln x$ approximately equals the proportionate change in x. See Chapter 9.1 for further details.

2.7 Key Concepts

1. Commonly-used statistics for numerical data include the mean and median (for central tendency), the standard deviation, inter-quartile range and range (for dispersion), quartiles and percentiles, and symmetry and kurtosis statistics.

2. An outlying observation, or outlier, is an observation that is unusually large or small.

3. A box plot provides a visual summary of key sample statistics. Some box plots also plot outlying observations.

4. Commonly-used charts that can provide a useful visual presentation of the data are histograms, kernel density graphs, column charts, line charts, bar charts and pie charts. Which is best to use depends on whether the data is numerical (continuous or discrete) or categorical, and whether the data are cross-section or time series data.

5. Common transformations of economic data include the natural logarithm, standardized scores and (for time series data) moving averages, seasonal adjustment, real data, growth rates and percentage changes.

6. Key Terms: sample; summary statistics; central tendency; central location; summation notation; sample mean; median; mid-range; mode; quartile; decile; percentile; quantile; dispersion; sample variance; standard deviation; coefficient of variation; range; outlying observation; outlier; inter-quartile range; symmetry; skewness; right-skewed; positive skewed; kurtosis; normal distribution; box plot; histogram; frequency; relative frequency; stem and leaf display; smoothed histogram; kernel density estimate; line chart; horizontal bar chart; vertical bar chart; column chart; pie chart; standardized score; moving average; seasonal adjustment; nominal data; real data; growth rates; percentage changes; percentage points; basis points.

2.8 Exercises

1. Obtain $\sum_{i=1}^{n} z_i$ for the following cases with $n = 5$:

   (a) $z_i = 1$. (b) $z_i = i$. (c) $z_i = 2i^2$. (d) $z_i = 1/i$. (e) $z_i = (2 + 3i)$.

2. Calculate the following

   (a) $\sum_{i=1}^{6} 2$. (b) $\sum_{i=1}^{4} 2/i$. (c) $\sum_{i=1}^{3} 3i^3$. (d) $\sum_{i=4}^{6} i$. (e) $\sum_{i=1}^{4}(5 + 2i)$.

3. For the panel variable x_{it} that takes values $x_{11} = 5$, $x_{12} = 3$, $x_{13} = 7$, $x_{21} = 8$, $x_{22} = 6$, and $x_{23} = 4$:

   (a) Calculate $\sum_{t=1}^{3} x_{it}$ for $i = 1$ and for $i = 2$.
   (b) Calculate $\sum_{i=1}^{2} x_{it}$ for $t = 1$, for $t = 2$ and for $t = 3$.

4. For the panel variable x_{it} that takes values $x_{11} = 2$, $x_{12} = 5$, $x_{21} = 8$, $x_{22} = 4$, $x_{31} = 6$, and $x_{32} = 7$:

   (a) Calculate $\sum_{t=1}^{2} x_{it}$ for $i = 1$, for $i = 2$ and for $i = 3$.
   (b) Calculate $\sum_{i=1}^{3} x_{it}$ for $t = 1$ and for $t = 2$.

5. Compute from first principles (i.e. using the formula and a calculator) the mean, standard deviation, coefficient of variation, skewness statistic and kurtosis statistic for the sample 4, 2, 0, 2. Show all calculations.

6. Repeat the previous exercise when the observations are ten times larger, so the sample is now 40, 20, 0, and 20. Which of the measures are scale-free measures?

7. Repeat exercise 5 when the observations are translated by 2, so the sample is now 6, 4, 2, and 4. Which of the measures are unchanged by translation?

8. A sample of size 200 has mean of 20 and standard deviation of 5. If the data are normally distributed, what range of values do you expect 95% of the sample to lie in?

9. IQ scores have a mean of 100, standard deviation of 14 and are approximately normally distributed. What range of IQ scores do you expect 99.7% of the population lie in?

10. For a sample of size 1,000 the central two-thirds of the observations lie between 60 and 100. If these data are normally distributed, provide an estimate of the mean and standard deviation.

11. For each of the following situations state whether the median price or the mean price of cars sold is a more useful measure of central tendency.

    (a) You want to know the typical price of a car.
    (b) You also know the number of cars sold and want to calculate sales tax receipts when car sales are subject to a 5% tax.

12. In each of the following situations state whether or not the data are likely to be positively skewed, or whether there is not enough information to know.

    (a) The mean is 50 and the median is 20.
    (b) The skewness statistic is 0.1.
    (c) The excess kurtosis statistic is 5.
    (d) The 10^{th} percentile is 20, the median is 50 and the 90^{th} percentile is 200.

13. The dataset HOUSE has data on the price and size of houses sold in a small homogeneous community.

    (a) Read the data into your statistical package.
    (b) Obtain detailed summary statistics for price. Do the data appear to be skewed? Explain.
    (c) Obtain a histogram. Do the data appear to be normally distributed? Explain.
    (d) Obtain a kernel density estimate. Do the data appear to be normally distributed? Explain.

14. Repeat the previous exercise for house size.

15. A sample of 30 people had the following years of completed schooling: 12, 12, 14, 12, 12, 12, 12, 12, 16, 12, 14, 12, 12, 13, 14, 12, 17, 12, 12, 16, 12, 12, 8, 14, 16, 12, 12, 17, 12, 16.

    (a) Read the data into your statistical package.

    (b) Obtain summary statistics. Give the inter-quartile range. List the first five observations.

    (c) Obtain a table of frequencies for these data.

    (d) Give a histogram, with a bin width of one for these discrete data. Do the data appear to be normally distributed?

    (e) Provide a pie chart - what is the most common value of the variable?

16. Repeat the previous exercise for the following samples:

    (a) 20 people age 30 with the following number of annual doctor visits: 0, 0, 3, 4, 2, 5, 5, 2, 11, 2, 2, 2, 3, 0, 8, 0, 8, 1, 2, 4.

    (b) 25 families with the following number of family members: 3, 3, 4, 7, 4, 3, 5, 2, 2, 4, 7, 3, 4, 3, 3, 5, 3, 4, 4, 1, 6, 5, 4, 5, 5.

17. The unemployment rate for college graduates (bachelor's degree or higher) aged 25 to 34 years in April in each of the years 2000 to 2019 was 1.3, 2.0, 2.7, 2.9, 2.6, 2.1, 2.2, 1.9, 2.2, 4.3, 4.7, 3.9, 3.6, 3.6, 3.0, 2.5, 2.1, 2.3, 1.9, 2.2.

    (a) Read the data into your statistical package.

    (b) Obtain key summary statistics. Give the inter-quartile range. List the first five observations.

    (c) Order by increasing unemployment rate and give a line chart.

18. Repeat the previous exercise for high school graduates (no college) aged 25 to 34 years with April unemployment rates of 4.4, 5.2, 8.7, 7.9, 7.1, 6.4, 6.1, 5.5, 6.8, 13.6, 13.9, 13.7, 10.1, 9.9, 8.8, 7.8, 7.8, 6.0, 6.3, 4.8.

19. Obtain data from the website https://fred.stlouisfed.org/ (FRED - Federal Reserve Economic Data) on the unemployment rate for those 25 years and over with some college or an associate degree in April in each of the years 2000 to the present. Answer the same questions as in exercise.17

20. Table 2.3 gives U.S. health expenditures by category for 2018. For 2013 the corresponding amounts were, respectively, 937, 587, 111, 80, 148, 80, 156, 370, 37, 174, 75, 47, 118.

    (a) Give a column chart, ordered by the amount of expenditure.

    (b) Give a pie chart. Is this more or less useful than a column chart? Explain.

21. The dataset PRICEEARNINGSRATIO has annual data on the Shiller cyclically-adjusted price-earnings ratio (variable *cape*) in January for S&P500 firms from 1881 to 2020.

    (a) Obtain the summary statistics for *cape*. Do the data appear to be skewed? Do the data appear to have greater kurtosis than the normal distribution? Explain.

    (b) Plot the histogram. Do the data appear to be skewed?

    (c) Provide a time series plot of the data. Comment on any unusual features.

    (d) Do the data to be unusually high or low in 2020? Explain.

22. The dataset AUSREGWEALTH has data on average net worth of households in thousands of dollars in 517 regions in Australia in 2003-04.

    (a) Obtain the summary statistics. Do the data appear to be skewed? Do the data appear to have greater kurtosis than the normal distribution? Explain.

    (b) Plot the histogram. Do the data appear to be skewed?

    (c) If your software does this, plot the kernel density estimate. Do the data appear to be skewed?

    (d) Now take the natural logarithm of average net worth and repeat parts a-c.

23. Use quarterly data in dataset STOCKINDEX from January 1957 to November 2012.

    (a) Calculate the z-score for each of the Dow Jones, Nasdaq and S&P 500 indexes.

    (b) Do these z-scores have mean zero and standard deviation one?

    (c) Give histograms (or kernel density estimates) for each of the three z-scores. Do they appear to be normally distributed?

    (d) On the same graph give line plots of each of the three z-scores against time. Do the three series appear to move together?

24. Use data in dataset GDPAUSTRALIA from January 1960 to September 2013. The data are quarterly data on nominal GDP (at an annual rate in millions of Australian dollars), a price index (=100 in 2011) and population (in millions).

    (a) Plot nominal GDP and real GDP (which you need to create) against time. Comment.

    (b) Compute nominal GDP per capita and real GDP per capita and plot these against time. Comment.

25. Use data in dataset GDPAUSTRALIA, described in the previous exercise.

    (a) Compute a four period moving average for nominal GDP. Has this reduced seasonal variation?

    (b) Compute annual growth rate in real GDP as four times the proportionate change from one quarter to the next.

    (c) Compute annual growth rate in real GDP as the proportionate change over the last four quarters.

    (d) Compare the two growth rate measures and comment.

26. Derivation of the alternative computational formula for the sample variance.

    (a) Show that $(x_i - \bar{x})^2 = x_i^2 - 2x_i\bar{x} + \bar{x}^2$.

    (b) Hence show that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2$.

    (c) Use the definition of $\bar{x}$ to show that $\sum_{i=1}^{n} x_i = n\bar{x}$.

    (d) Hence show that $\sum_{i=1}^{n} 2x_i\bar{x} = 2n\bar{x}^2$. (Hint: $\sum_{i=1}^{n} az_i = a\sum_{i=1}^{n} z_i$).

    (e) Substitute this result into part (b) and simplify to show that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = (\sum_{i=1}^{n} x_i^2) - n\bar{x}^2$.

Chapter 3

The Sample Mean

Obtaining the sample mean $\bar{x}$, and other summary statistics, is straightforward. But different samples will yield different values of these sample statistics, due to the inherent randomness in the data. How can this randomness be controlled for if we want to make statements about the unchanging features of the distribution for the entire population? More simply, how can we extrapolate from the sample to the population?

For example, dataset EARNINGS introduced in Chapter 2 has data on individual annual earnings for a sample of 30-year-old female full-time workers. The sample mean from a random sample of size 171 was \$41,413. What can be said about the likely range of values of mean earnings for all 30-year-old female full-time workers in the country? Are average earnings in this population really as high or as low as \$41,413? Or is the observed sample mean of \$41,413 just an artifact of this particular sample?

The chapter is relatively dense. While selecting only the essential material, it introduces a considerable amount of the probability theory covered in an introductory probability and statistics course. The focus is on the concepts of mean and variance of a single random variable, and the consequent distribution of the average of n random variables. The simplest material is presented in the text, with additional background material on probability and derivations for the sample mean presented in Appendix B.

For readers who skip this chapter, the essential properties of the sample mean are restated in Chapter 4.2.

3.1 Random Variables

Different samples take different values due to randomness. To account for this randomness we need to introduce random variables and define key properties of random variables, notably their mean, standard deviation and variance.

3.1.1 Random Variables

A **random variable** is a variable whose value is determined by the outcome of an experiment, where an **experiment** is an operation whose outcome cannot be predicted with certainty.

For example, the experiment may be tossing a coin and the random variable may take value 1 if heads and 0 if tails. As a second example, the experiment may be randomly selecting a person from the population and the associated random variable takes value equal to their annual earnings.

Standard notation is to denote the random variable in upper case, say X (or Y or Z), and to denote the values that the random variable can take in lower case, say x (or y or z).

3.1.2 Example: Coin Toss

The simplest example of a random variable is one that has only two possible values. We consider a coin toss with a fair coin and define the random variable X to take value 1 if heads and value 0 if tails. Because the coin is fair, there is equal probability of heads or tails.

The random variable is

$$X = \begin{cases} 0 & \text{with probability } 0.5 \\ 1 & \text{with probability } 0.5. \end{cases}$$

3.1.3 Mean of a Random Variable

Interest lies in summarizing the distribution of the random variable. Key measures used are the mean, to describe the average value that the random variable may take, and the variance and standard deviation, to measure the variability of the random variable.

The **expected value of the random variable** X is the long-run average value that we expect if we draw a value of X at random, draw a second value and so on, and then obtain the average of these values. Equivalently, calculate the probability-weighted average by weighting each value x that X may take by the probability of that value x occurring.

This expected value, denoted $\mathrm{E}[X]$, is called the **mean** of X.

Definition 1 *The mean $\mu = E[X]$ of the random variable X is the probability-weighted average of all values that the random variable X may take. The notation μ (or mu) is used to denote the man as μ is the Greek letter for m.*

Suppose our random variable may take values x_1, x_2, ... with potentially different probabilities $\Pr[X = x_1]$, $\Pr[X = x_2]$, ... These probabilities necessarily sum to one. Then **the mean of X is the probability-weighted average**

$$\begin{aligned} \mu \equiv \mathrm{E}[X] &= x_1 \times \Pr[X = x_1] + x_2 \times \Pr[X = x_2] + \cdots \\ &= \textstyle\sum_x x \times \Pr[X = x], \end{aligned}$$

where $\sum_x$ denotes summation over all the possible distinct values that X may take.

As an example, for a fair coin toss where X can take values 0 or 1 with equal probabilities of 0.5 and 0.5 we have

$$\begin{aligned}\mu &= \textstyle\sum_x x \times \Pr[X = x] \\ &= \Pr[X = 0] \times 0 + \Pr[X = 1] \times 1 \\ &= 0.5 \times 0 + 0.5 \times 1 \\ &= 0.5.\end{aligned}$$

As a second example, suppose the coin is not fair and X can take value 1 with probability 0.6 and value 0 with probability 0.4. Then $\mu = 0 \times 0.4 + 1 \times 0.6 = 0.6$.

3.1.4 Variance and Standard Deviation

The variance is the long-run average value that we expect if we draw a value of X at random, say x_1, and compute its squared deviation from the mean $(x_1 - \mu)^2$, draw a second value and compute $(x_2 - \mu)^2$, and so on, and then obtain the average of these values.

This **expected value** of $(X - \mu)^2$, denoted $\mathrm{E}[(X - \mu)^2]$, is called the **variance** of X and is also denoted σ^2 or σ_X^2.

Definition 2 *The variance $\sigma^2 = E[(X - \mu)^2]$ of the random variable X is the probability-weighted average of all values that $(X - \mu)^2$ may take. The standard deviation is $\sigma = \sqrt{\sigma^2}$. The notation σ (or sigma) is used to denote the standard deviation as σ is the Greek letter for s.*

For random variable X taking values x_1, $x_2, \ldots$ the **variance of X is the probability-weighted average**

$$\begin{aligned}\sigma^2 \equiv \mathrm{E}[(X - \mu)^2] &= (x_1 - \mu)^2 \times \Pr[X = x_1] + (x_2 - \mu)^2 \times \Pr[X = x_2] + \cdots \\ &= \textstyle\sum_x (x - \mu)^2 \times \Pr[X = x].\end{aligned}$$

The **standard deviation** σ is obtained by taking the square root of the variance.

Continuing the earlier fair coin toss example with $\mu = 0.5$

$$\begin{aligned}\sigma^2 &= \textstyle\sum_x (x - \mu)^2 \times \Pr[X = x] \\ &= (0 - 0.5)^2 \times \Pr[X = 0] + (1 - 0.5)^2 \times \Pr[X = 1] \\ &= 0.25 \times 0.5 + 0.25 \times 0.5 \\ &= 0.25.\end{aligned}$$

The variance of X is 0.25 and the standard deviation of X is $\sqrt{0.25} = 0.5$.

3.1.5 Example: Best Three of Five

Suppose two evenly matched teams, with equal probabilities of winning any game, play in a series of up to five games, with the winner the first to win three games. How many games do we expect on average?

Let X denote the number of games, in which case X can take values 3, 4 or 5. It can be shown that $\Pr[X = 3] = \frac{1}{4}$, $\Pr[X = 4] = \frac{3}{8}$, and $\Pr[X = 5] = \frac{3}{8}$. Then

$$\mathrm{E}[X] = 3 \times \tfrac{1}{4} + 4 \times \tfrac{3}{8} + 5 \times \tfrac{3}{8} = 4\tfrac{1}{8},$$

so on average we expect 4.125 games. Additionally

$$\mathrm{Var}[X] = (3 - 4\tfrac{1}{8})^2 \times \tfrac{1}{4} + (4 - 4\tfrac{1}{8})^2 \times \tfrac{3}{8} + (5 - 4\tfrac{1}{8})^2 \times \tfrac{3}{8} = \tfrac{39}{64}.$$

3.1.6 Some Key Properties of Random Variables

Further details on random variables are given in Appendix B. The mean of a constant a is that constant a. If we add a fixed amount a to a random variable then the mean is changed by the amount a. And if we multiply a random variable by a fixed multiple b then the mean is multiplied by b. Combining these results we have

Remark 1 *$E[a + bX] = a + b \times E[X]$, for constants a and b.*

The variance of a constant a is zero. If we add a fixed amount a to a random variable then the variance is unchanged. And if we multiply a random variable by a fixed multiple b then the variance is multiplied by b^2. Combining these results we have

Remark 2 *$Var[a + bX] = b^2 \times Var[X]$, for constants a and b.*

For example, if X has mean μ and variance σ^2, then $a + X$ has mean $a + \mu$ and variance σ^2 and bX has mean $b\mu$ and variance $b^2\sigma^2$. It follows that $a + bX$ has mean $a + b\mu$ and variance $b^2\sigma^2$. Applying these rules it follows that $Y = (X - \mu)/\sigma$ has mean 0 and variance 1; Y is called a **standardized random variable**.

3.2 Random Samples

For statistical inference we view our data as being a random sample with each observation being a random outcome.

3.2.1 Random Samples

A sample of size n takes values denoted $x_1, ..., x_n$. In Chapter 2 we focused on using various descriptive statistics and graphs to summarize these values. Now we recognize that each value is a random outcome: x_1 is the observed or realized or outcome value of the random variable X_1, x_2 is the observed or realized or outcome value of the random variable X_2, and so on.

For example, suppose we have a sequence of four coin tosses with consecutive results tails, heads, heads and heads. Then random variable X_1 has realized value $x_1 = 0$, X_2 takes value $x_2 = 1$, X_3 takes value $x_3 = 1$ and X_4 takes value $x_4 = 1$.

Definition 3 *A sample of size n has observed values $x_1, x_2, ..., x_n$ that are realizations of the random variables $X_1, X_2, ..., X_n$.*

3.2.2 The Sample Mean is a Random Outcome

The **sample mean** is the average of the n sample values $x_1, ..., x_n$, or

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

For example, for four coin tosses that yield values 0, 1, 1 and 1, the sample mean is $\bar{x} = (0 + 1 + 1 + 1)/4 \simeq 0.75$.

The sample values $x_1, ..., x_n$ are realized outcomes of the random variables $X_1, X_2, ..., X_n$. It follows that **the sample mean $\bar{x}$ is a realization of the random variable**

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The random variable $\bar{X}$ is also called the **sample mean**. It should be clear from the context whether sample mean refers to the random variable $\bar{X}$ or its observed value $\bar{x}$.

Definition 4 *The observed sample mean $\bar{x}$ is the realized value of the random variable $\bar{X}$; $\bar{X}$ is also called the sample mean.*

3.2.3 Sample Variance and Standard Deviation

The **sample variance** is the average of the squared deviations of x around $\bar{x}$, rather than around μ, since μ is unknown. From Chapter 2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

The divisor $(n-1)$ is called the **degrees of freedom** because only $(n-1)$ terms in the sum are free to vary since they are linked by the relationship $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Taking the square root of s^2 yields the **sample standard deviation** s.

Like the sample mean, the sample variance is the realization of a random variable, namely

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

Similarly, the sample standard deviation s is a realization of the random variable S.

3.3 Sample Generated by an Experiment: Coin Tosses

We consider an example of a sample generated by an experiment where the values for the mean and standard deviation of the underlying random variable X are known and are specified.

We then take a series of samples, by running the experiment many times, and for each sample obtain the sample mean $\bar{x}$. We are interested in comparing the distribution of the many sample means to the distribution of X.

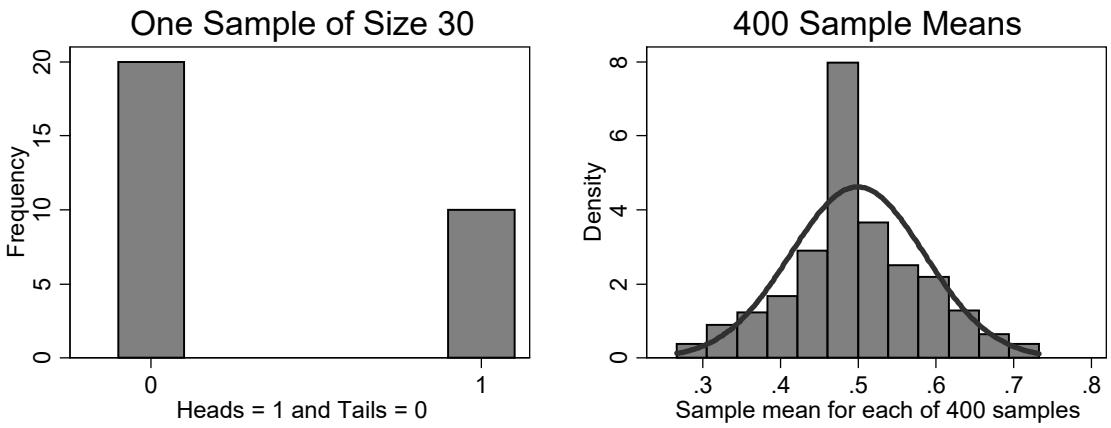

Figure 3.1: Coin tosses histograms: for x in one sample (n = 30) and for mean of x in 400 samples (n = 30).

3.3.1 Example: Coin Tosses

We consider the fraction of times that a fair coin lands heads in 30 tosses.

The random variable $X = 1$ if heads and $X = 0$ if tails. Given equal probabilities of heads and tails, X has mean $\mu = 0.5$ and standard deviation $\sigma = 0.5$.

The left panel of Figure 3.1 shows a histogram for one sample of 30 tosses. In this sample there were 10 heads and 20 tails, so $\bar{x} = 10/30 = 0.333$, and $s = 0.479$, values that due to randomness differ from $\mu = 0.5$ and $\sigma = 0.5$.

3.3.2 Many Samples

Now randomly draw 400 samples, each of 30 coin tosses. In this example the first three such samples have means $\bar{x}_1 = .333$, $\bar{x}_2 = .500$ and $\bar{x}_3 = .533$. Dataset COINTOSSMEANS has all 400 sample means.

The right panel of Figure 3.1 presents a histogram for the 400 sample means. The histogram is roughly centered on the individual mean; the average of the 400 means is 0.499 which is close to $\mu = 0.5$. The standard deviation of the 400 means equals 0.086. So there is much less variability in these 400 means than in the individual observations. Here the standard deviation of the 400 means is between one-fifth and one-sixth of $\sigma = 0.5$, the standard deviation of X. Finally, we superimpose the density for the normal distribution with mean 0.499 and standard deviation 0.086. It is clear that the histogram of the 400 means is roughly that of a normally distributed random variable.

In the preceding example we did not actually toss a coin 12,000 times to obtain the results for 400 samples, each with 30 coin tosses. Instead a computer was used to simulate the coin tosses. The method to do so is explained in Chapter 3.8.

3.4 Properties of the Sample Mean

The coin toss example yielded (1) sample mean that is on average close to the mean μ of the individual observations; (2) variability of the sample mean that is much less than that of the underlying individual observations; and (3) sample mean that is approximately normally distributed.

In this section these results are formalized in a general setting. The statistical properties of the random variable $\bar{X}$, the sample mean, are determined by the process generating the underlying individual random variables $X_1, X_2, ..., X_n$.

3.4.1 Assumptions

Standard basic assumptions about the individual random variables X_i are that

A. X_i has common mean μ: $\mathrm{E}[X_i] = \mu$ for all i.

B. X_i has common variance σ^2: $\mathrm{Var}[X_i] = \sigma^2$ for all i.

C. Different observations are statistically independent: X_i is statistically independent of $X_j, i \neq j$.

Here statistical independence implies, for example, that the value taken by X_2 is not influenced by the value taken by X_1; see Appendix B.2 for a more formal definition. For example, for a fair coin toss the probability of heads on the second coin toss is 0.5 regardless of whether the first coin toss yielded heads or tails.

Short-hand notation for assumptions A-B is that $X_i \sim (\mu, \sigma^2)$ for all i or, even more simply, that

$$X \sim (\mu, \sigma^2),$$

where $\sim$ means "**is distributed as**", and the terms in parentheses are, respectively, the mean and the variance of X_i.

Assumptions A-C are met when data are obtained from a **simple random sample**, often called more simply a **random sample**, where we make independent draws $X_1, ..., X_n$ from the same distribution. Chapter 3.4.6 discusses relaxing these assumptions.

3.4.2 Mean of the Sample Mean

The **mean of the sample mean** $\bar{X}$ is

$$\mu_{\bar{X}} \equiv \mathrm{E}[\bar{X}] = \mu.$$

In words, the expected value of the sample mean equals the mean for each individual, so the average of $\bar{X}$ from many samples is expected to equal μ.

This result means that if we were able to obtain many random samples and for each sample obtain the sample mean, then on average the sample means equal the mean of a single variable X.

Only assumption A (common mean of X_i) is needed to obtain this result. The proof uses $\mathrm{E}[aX] = a\mathrm{E}[X]$ and $\mathrm{E}[X+Y]$ =E[X]+E[Y]. Then

$$\begin{aligned}\mathrm{E}[\bar{X}] &= \mathrm{E}[\tfrac{1}{n}(X_1 + X_2 + \cdots + X_n)] \\ &= \tfrac{1}{n}\mathrm{E}\left[X_1 + X_2 + \cdots + X_n\right] \\ &= \tfrac{1}{n}\{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n]\} \\ &= \tfrac{1}{n}\{\mu + \mu + \cdots + \mu\} = \mu.\end{aligned}$$

3.4.3 Standard Deviation of the Sample Mean

The variability of $\bar{X}$ around its mean of μ is measured using the variance and standard deviation of $\bar{X}$.

The **variance of the sample mean** $\bar{X}$ is

$$\sigma^2_{\bar{X}} = \mathrm{Var}[\bar{X}] \equiv \mathrm{E}[(\bar{X} - \mu_{\bar{X}})^2] = \frac{\sigma^2}{n},$$

where σ^2 is the variance of X. The proof requires all of assumptions A-C (same mean, same variance and independence of X_i) and uses $\mathrm{Var}[aX] = a^2\mathrm{E}[X]$ and that for independent variables $\mathrm{Var}[X+Y]$ =Var[X]+Var[Y]; see Appendix B.2 for complete details.

The **standard deviation of the sample mean** $\bar{X}$ is then

$$\sigma_{\bar{X}} \equiv \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

The variance result that $\sigma^2_{\bar{X}} = \sigma^2/n$ implies that the sample mean is less variable than the underlying data, as demonstrated in Figure 3.1 for the coin toss example.

Furthermore the variability of the sample mean as an estimate of the mean of an individual variable X decreases greatly as the sample size increases, at rate n for the variance and at rate $\sqrt{n}$ for the standard deviation. Thus for the coin toss example the standard deviation of the 400 means was 0.086, close to the true standard deviation $\sigma/\sqrt{n} = 0.5/\sqrt{30} \simeq 0.091$.

As expected, **larger samples lead to greater precision** in estimating μ. Furthermore, $\sigma^2_{\bar{X}} = \sigma^2/n \to 0$ as $n \to \infty$, so the sample mean will be very close to μ as the sample size $n \to \infty$.

Remark 3 *Under simple random sampling the sample mean $\bar{x}$ is the realization of a random variable $\bar{X}$ that has mean equal to the mean μ and standard deviation $\sigma/\sqrt{n}$ that gets smaller as the sample size increases.*

3.4.4 Normal Distribution and the Central Limit Theorem

From the right panel of Figure 3.1 the sample means appear to be approximately normally distributed, even though each observation is clearly not from the normal distribution. Remarkably this is the case in quite general settings, provided the sample size is sufficiently large.

The preceding results imply that the random variable

$$\bar{X} \sim (\mu, \sigma^2/n).$$

Subtracting the mean and dividing by the standard deviation leads to a standardized random variable that by construction has mean 0 and variance 1. Here we denote this standardized random variable by Z, so

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim (0, 1).$$

In general the distributions of $\bar{X}$ and of Z vary with the distribution of X and there is no simple formula for these distributions. One notable extension is that if X, the underlying variable for a single observation, is normally distributed then $\bar{X}$ is normally distributed and Z is standard normal distributed. Remarkably even if X is not normally distributed we obtain these results, provided the sample size is large.

In particular, if the sample satisfies assumptions A-C and, additionally, the sample size $n \to \infty$, then a result from statistics called the **central limit theorem**, states that Z has the **standard normal distribution**, so then

$$Z \sim N(0, 1) \text{ as } n \to \infty.$$

This remarkable result result is proved using advanced mathematical methods; there is no intuition for the result. The central limit theorem, first derived in 1733, gets its name because it is for the limit (as $n \to \infty$) of a measure of the center of the distribution.

It follows that for large n a good approximation to the distribution of $\bar{X}$ is

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

Often $n > 30$ is sufficient for this to be a reasonable approximation. $\bar{X}$ is said to be **asymptotically normal** distributed, where the term **asymptotic** means as the sample size goes to infinity.

The wide applicability and usefulness of the central limit theorem cannot be understated. Regardless of the distribution of the underlying random variable X, if assumptions A-C hold then averaging leads to a standardized random variable that is standard normally distributed in large samples. It can also be extended to cases where not all of assumptions A–C hold; see Appendix B.2.

Remark 4 *Under assumptions A-C the central limit theorem implies that the standardized random variable $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is standard normal distributed as the sample size goes to infinity. For large n a good approximation is that $\bar{X} \sim N(\mu, \sigma^2/n)$.*

3.4.5 Standard Error of the Sample Mean

The variance and standard deviation of $\bar{X}$ depend on the variance σ^2 which is unknown. Replacing σ^2 by its estimate s^2, leads to the following estimates.

The **estimated variance** of $\bar{X}$ is

$$s^2_{\bar{X}} = \frac{s^2}{n} = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}{n}.$$

Taking the square root, the **estimated standard deviation** of $\bar{X}$, called the **standard error of the sample mean**, is

$$se(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}}{\sqrt{n}},$$

where s is the sample standard deviation, the sample estimate of the standard deviation of X.

Note that in general the term "**standard error**" means **estimated standard deviation**. The various estimators considered in this book each have a distinct standard error. In many situations computer output will include a reported "standard error", but this is not necessarily the standard error of the sample mean $\bar{X}$.

Remark 5 *Under simple random sampling the standard error (the estimated standard deviation) of the sample mean $\bar{X}$ equals $s/\sqrt{n}$ where s is the sample standard deviation for a single observation.*

It can be shown that, under assumptions A-C, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ has the desirable property that $\mathrm{E}[S^2] = \sigma^2$. For this reason the formula for s^2 divides by $n-1$ rather than the more obvious n.

3.4.6 Relaxing assumptions A-C

The starting point is to assume simple random sampling, but methods can be adjusted to relax assumptions A-C.

Assumption A requires a common mean μ. Regression analysis generalizes this by allowing the mean to differ with individual characteristics. For example, expected earnings for an individual may vary with education.

Assumption B requires a common variance and assumption C requires independence of observations. If either of these assumptions fail, then $\bar{X}$ still has mean μ, provided assumption A holds, but the variance of $\bar{X}$ is no longer σ^2/n. In particular, if observations are correlated (assumption C fails) then alternative formulas to $s/\sqrt{n}$ are used to compute estimate $se(\bar{x})$, the standard error of the sample mean. The correct $se(\bar{x})$ is most easily obtained by least squares regression on just an intercept and using appropriate robust standard errors; see Chapter 12.1.

A greater complication arises if the sample is not representative of the population. This is discussed in Chapter 3.7.

3.4.7 Summary for the Sample Mean

The distinction between variability in X_i, the random variable leading to the i^{th} sample value x_i, and the variability in $\bar{X}$, the random variable with observed value the sample mean $\bar{x}$, can cause confusion. A summary given simple random sampling is the following:

1. Sample values $x_1, ..., x_n$ are realized or observed values of the random variables $X_1, ..., X_n$.
2. Individual X_i are assumed to be independent have common mean μ and variance σ^2.
3. The average $\bar{X}$ of the n draws of X_i has mean μ and variance σ^2/n.

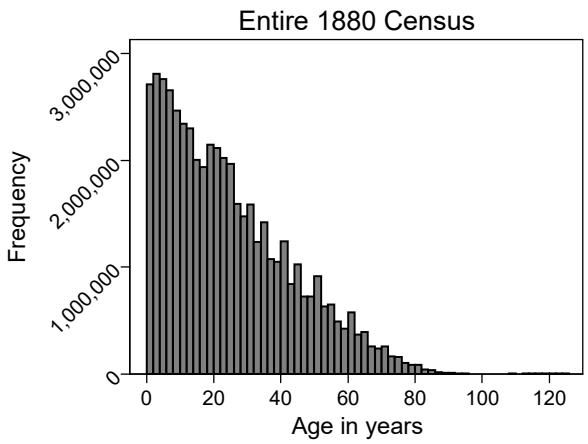

Figure 3.2: 1880 Census histogram: Age for the entire U.S. population.

4. The standardized statistic $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has mean 0 and variance 1.

5. Under assumptions A-C, Z is standard normal distributed as sample size $n \to \infty$, by the central limit theorem.

6. For large n a good approximation is that $\bar{X} \sim N(\mu, \sigma^2/n)$.

7. The standard error of $\bar{X}$ equals $s/\sqrt{n}$, where "standard error" is general terminology for "estimated standard deviation".

3.5 Sampling from a Finite Population: 1880 Census

As a second example of sampling we consider obtaining a sample from a finite population.

3.5.1 Example: 1880 U.S. Census

The 1880 Census provides a complete enumeration of the U.S. population in 1880. We consider one of the variables that was recorded, that on age in years.

3.5.2 Population

Figure 3.2 provides a histogram of age for all 50,169,452 people recorded as living in the U.S. in 1880. The distribution is basically declining in age. The blips are due to individuals rounding their age to the nearest five years or ten years.

For a complete census such as this, the observed distribution is actually the distribution of X, with the age of each person occurring with probability $1/N$, where $N = 50{,}169{,}452$. The population

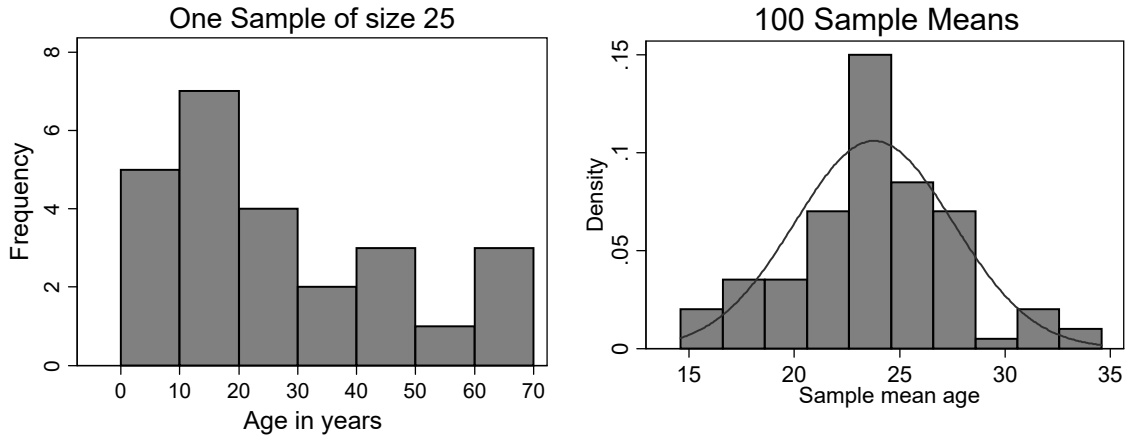

Figure 3.3: 1880 Census histograms: Age in one sample (n = 25) and Mean age in 100 samples (n = 25).

average age is 24.13 years, so $\mu = 24.13$, since $E[X] = \sum_x x \times \Pr[X = x] = \sum_{i=1}^{N} x_i \times \frac{1}{N} = 24.13$. Similarly, the population standard deviation of age is 18.61, so $\sigma = 18.61$.

3.5.3 Samples

Now consider taking one randomly-drawn **sample** of size $n = 25$ drawn from this population of size $N = 50{,}169{,}452$. The left panel of Figure 3.3 presents the histogram for this single sample of size $n = 25$. For this sample, the average age was 27.84 years, so $\bar{x} = 27.84$, and the standard deviation of age is 20.71, so $s = 20.71$. Due to the randomness of sampling, these are similar to, but not exactly equal to, μ and σ.

Now randomly draw 100 distinct samples of size 25, leading to 100 different sample means. The first three such samples turned out to have means $\bar{x}_1 = 27.84$, $\bar{x}_2 = 19.40$ and $\bar{x}_3 = 23.28$ years. The right panel of Figure 3.3 presents a histogram for these 100 sample means that are stored in dataset CENSUSAGEMEANS. Several things are apparent.

First, the histogram is roughly centered on the mean μ. In fact the average of the 100 means is 23.78, close to $\mu = 24.13$.

Second, there is much less variability in these 100 means than in the original population. Here the standard deviation of the 100 means is 3.76, roughly one-fifth of the standard deviation of $\sigma = 18.61$. In fact from theory already presented $\bar{X}$ has standard deviation $\sigma/\sqrt{n} = 18.61/\sqrt{25} = 3.72$.

Third, the histogram is roughly that of a normally distributed random variable. This is apparent by superimposing the density for the normal distribution with mean 23.78 and standard deviation 3.76.

3.6 Estimation of the Population Mean

In the examples given so far the distribution of X has been fully specified, so that we know the exact value of the **population mean** μ. In practice μ is unknown and we wish to estimate μ.

For example, if a coin is known to be fair then for a single coin toss $\mu = 0.5$. But suppose we do not know that the coin is fair. More generally let $\Pr[X = 1] = p$ in which case $\Pr[X = 0] = 1 - p$ and some simple algebra shows that $\mu = p$. Now we need to estimate p, which in this example is the same as estimating μ. The obvious estimator is $\bar{X}$, but in what sense is $\bar{X}$ a good estimator of μ?

Due to randomness, an estimator of μ will not exactly equal μ. Two desirable properties of an estimator of μ is that its distribution be centered on μ and that it has as little variability as possible around μ.

3.6.1 Parameter, Estimator and Estimate

The goal in estimation is to estimate one or more parameters, where a **parameter** is a constant that determines in part the distribution of X. Examples of parameters are the mean μ and the variance σ^2.

Definition 5 *A **parameter** is a constant that determines in part the distribution of X. An **estimator** is a method for estimating a parameter. An **estimate** is the particular value of the estimator obtained from the sample.*

For estimation of the mean of X using the sample mean, the parameter is μ, the estimator is the random variable $\bar{X}$, and the estimate is the sample value $\overline{x}$.

3.6.2 Unbiased estimators

The first goal of estimation is to have an estimator that is centered on the parameter we wish to estimate. One standard criteria used is **unbiasedness**.

Definition 6 *An **unbiased estimator** of a parameter has expected value equal to the parameter.*

The sample mean $\bar{X}$ is unbiased for μ under assumption A (a common mean) since, as already shown, $\mathrm{E}[\bar{X}] = \mu$.

Remark 6 *Under simple random sampling the sample mean is unbiased for μ, meaning that in repeated samples it will on average equal μ.*

3.6.3 Minimum Variance Estimators

Restricting attention to unbiased estimators still allows many potential estimators. For example, the sample median is an alternative estimator to the sample mean that is unbiased for μ if X is symmetrically distributed. In that case we discriminate between such estimators on the basis of the size of their variance.

Definition 7 *A **best estimator** or **efficient estimator** in a class of estimators has **minimum variance** among the class.*

Smaller variance is desired as then there will be less variability in the estimator from sample to sample. As an example of a poor choice for an unbiased estimator, suppose we just used the first observation in each sample of size n to estimate μ. Then this estimator is unbiased from sample to sample as $\mathrm{E}[X_1] = \mu$. But it has variance σ^2 which is high relative to other possible unbiased estimators.

For simple random samples the sample mean $\bar{X}$ has variance σ^2/n. Whether alternative unbiased estimators for μ can have smaller variance than this depends on the distribution for X. For some common distributions of X, notably the normal, Bernoulli, binomial, Poisson and exponential, it can be shown that, given data from a simple random sample, no other unbiased estimator of μ has smaller variance.

Remark 7 *Under simple random sampling the sample mean has the smallest variance among unbiased estimators for some common distributions of X (including the normal, Bernoulli, binomial, Poisson and exponential) though not for all distributions of X.*

The sample mean is generally used, for simplicity and because even in situations where the sample mean is not the most efficient estimator, its variance is usually not much greater than the minimum possible variance, so the efficiency loss in using the sample mean is not great. The next chapter presents confidence intervals for μ and tests of hypotheses on μ that use the sample mean as the estimate of μ.

3.6.4 Consistent estimators

A more advanced concept considers **asymptotic properties** of an estimator, i.e. behavior as the sample size goes to infinity.

Definition 8 *A **consistent estimator** of a parameter is one that is almost certainly arbitrarily close to the parameter, as the sample size gets very large.*

A sufficient condition for **consistency** is that (1) any bias disappears as the sample size gets large, and (2) the variance of the estimator goes to zero as the sample size gets large. A more precise definition of **consistency** is given in Chapter 6.4.

The sample mean $\bar{X}$ is consistent for μ under simple random sampling (assumptions A-C) as it is unbiased and has variance σ^2/n which goes to zero as $n \to \infty$. This convergence of $\bar{X}$ to μ as the sample size gets large is an example of a so-called **law of large numbers**.

3.7 Nonrepresentative Samples

The standard assumption is that data are generated from a simple random sample. As already noted, for unbiased estimation of the mean the key assumption is assumptions A, that the sample has common mean μ; assumptions B-C can be relaxed.

Serious complications arise, however, if the sample is not representative of the population of interest. Then assumption A in Chapter 3.4 does not hold, and $\bar{X}$ may be biased and inconsistent for the population mean μ.

This issue is particularly relevant for samples based on a survey. It has become relatively inexpensive to conduct a survey by means such as telephone or the internet. Due to nonrepresentativeness of the grouped surveyed, or high nonresponse rates even if the group surveyed is representative, the sample may be a very skewed sample. If a sample reveals a surprising result, it may be an artifact of being nonrepresentative.

3.7.1 Examples of Nonrepresentative Samples

For example, a survey of readers of Golf Digest will provide an inconsistent estimate of the golfing habits of all Americans, since it oversamples active golfers. (The survey might, however, provide a consistent estimate of the golfing habits of all readers of Golf Digest. This would be of use to the advertising department of Golf Digest. In this latter case the population of interest is viewed to be readers of Golf Digest rather than all Americans.)

A famous example of a nonrepresentative sample is the incorrect prediction of the winner of the 1948 U.S. presidential election. Opinion polls predicted that the Republican candidate John Dewey would defeat the incumbent Democrat, President Harry Truman. Yet Truman won convincingly. The opinion polls were not representative for two reasons. First, the last opinion polls were taken well before the election, so a late surge to Truman meant that they were not representative of opinions on election day itself. Second, the opinion polls were not based on random sampling - the interviewers were given too much discretion as to who they interviewed.

3.7.2 Weighted Mean for Survey Data

Samples obtained from government surveys and from political polling surveys are often not representative of the population. Yet the leading national surveys can nonetheless be adjusted to provide valid estimates of the population mean.

For example, the unemployment rate in the United States is obtained from the Current Population Survey (CPS), a monthly survey of 60,000 households. This survey is not a simple random sample. For example, households in smaller states are oversampled to provide more reliable state-level data. Similarly, minority and disadvantaged populations are oversampled. And the surveyed households are clustered geographically to reduce interview costs.

To overcome these complications, surveys such as the CPS, provide sampling weights that make possible unbiased estimation of the population mean using a **weighted mean**

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} = \frac{\frac{1}{n}\sum_{i=1}^n w_i x_i}{\frac{1}{n}\sum_{i=1}^n w_i},$$

where the **sample weights** w_i are the reciprocal of the **probability** that the i^{th} individual in the population is included in the sample. This method requires correct specification of the weights w_i.

Most standard statistical software enables computation of the weighted mean and the standard error of the weighted mean, provided the sample weights are known. An example is given in

exercise 19 of Chapter 12. Statistical software for **survey data** allows for additional complications of surveys.

3.8 Computer Generation of Random Samples

How are random samples generated on a computer?

The starting point is a **uniform random number generator** that creates values between 0 and 1 such that any value between 0 and 1 is equally likely and successive values appear to be independent of each other.

These random numbers are more properly called **pseudo random numbers**, as a deterministic rule is used to create the sequence of numbers $u_1, u_2, ...$ For example, one method given value j^{th} value u_j specifies the next value to be $u_{j+1} = (69069u_j + 1234567) \bmod 2^{32}$, where $a \bmod b$ is the remainder when a is divided by b. Remarkably this rule leads to u_{j+1} appearing to be unrelated to u_j and to the different possible values of u_j between 0 and 1 being equally likely.

The sequence depends on the starting value u_0, called the **seed**. For example, we might set the seed equal to 10101. When using random numbers it is always good practice to **set the seed**, as then results can be replicated exactly in future simulations.

3.8.1 Generating a Single Random Sample in a Statistical Package

In the coin toss example in Chapter 3.3 we did not actually toss a coin many times. Instead to simulate 30 coin tosses, say, we draw 30 uniform random numbers and let the result be heads if the uniform random number exceeds 0.5, and tails if the uniform random number is less than 0.5.

Similarly for the Census example, if the uniform random number is between 0 and $1/N$, where $N = 50{,}169{,}452$, we choose the first person. If the uniform random number is between $1/N$ and $2/N$ we choose the second person, and so on.

The uniform random numbers are also the basis for making draws from commonly-used distributions such as the binomial, Poisson and normal distributions. The following algorithm is used.

Remark 8 *To make n draws of the random variable X do the following: (1) set the sample size to n; (2) set the seed; and (3) make n draws of X from its specified distribution.*

For example, suppose we want to make 500 draws of variables x and y. Variable x is a draw from the uniform distribution on $(3, 9)$, so any value between 3 and 9 is equally likely. Equivalently variable x equals $3 + 6u$ where u is a draw from the uniform distribution on $(0, 1)$. Variable y is a draw from the $N(5, 2^2)$ distribution. Equivalently variable y equals $5 + 2z$ where z is a draw from the standard normal distribution.

In Stata give commands (1) `set obs 100`; (2) `set seed 10101`; (3) `generate x=runiform(3,9);` and (4) `generate x=rnormal(5,2)`.

In R give the commands (1) `set.seed(10101)`; (2) `x=runif(100,min=3,max=9);` and (3) `y=rnorm(100,5,2)`.

In Gretl give the commands (1) `nulldata 100`; (2) `set seed 10101`; (3) `genr x=uniform(3,9);` and (4) `genr y=normal(5,2)`.

In Eviews give commands (1) `wfcreate mywf 100`; (2) `rndseed 10101`; (3) `series x=3+6*rnd`; and (4) `series y=5+2*rnd`.

Note that different packages will yield different results as they use different algorithms.

3.8.2 Computer Generation of Many Samples

The examples in Chapters 3.3 and 3.5 obtain the sample mean from each of many samples. This requires commands that allow repeated operations and saving the results of these repeated operations for subsequent analysis. These more advanced commands vary with statistical package.

The following Stata code obtains 400 sample means in the coin toss example of Chapter 3.3, as well as standard deviations and the sample size.

```
program onesample, rclass
    drop _all
    set obs 30
    generate u = runiform()
    generate x = u > 0.5
    summarize x
    return scalar xbar = r(mean)
    return scalar sd = r(sd)
    return scalar nobs = r(N)
end
simulate xbar=r(xbar) stdev=r(sd) nobs=r(nobs), seed(10101) reps(400): onesample
summarize
```

The program `onesample` simulates each of 30 fair coin tosses by first drawing a uniform number `u` between 0 and 1 and setting variable random `x` to 1 if `u>0.5` and to 0 if `u<=0.5`. Some key results from command `summarize` are stored in `r()`, including $\bar{x}$ in `r(mean)`, s in `r(sd)`, s^2 in `r(Var)`, and N in `r(N)`. The `return scalar` commands generate variables that will be returned to command `simulate`. For example, the sample mean of the 30 `x`'s will be returned in variable `xbar`. The command `simulate` runs the program `onesample` 400 times (the value in `reps()`), leading to 400 observations on variables `xbar`, `stdev` and `nobs`. The option `seed(10101)` provides a starting value for the initial draw of `u`, leading to the same results each time this code is run.

The following R code obtains 400 sample means in the coin toss example of Chapter 3.3, as well as standard deviations.

```
set.seed(10101)
result.mean=array(dim=400)
result.stdev=array(dim=400)
for(i in 1:400){
    x=rbinom(30,1,0.5)
    result.mean[i]=mean(x)
    result.stdev[i]=sd(x)
}
```

```
mean(result.mean)
sd(result.mean)
summary(result.mean)
```

The R code obtains 30 fair coin tosses by using the command `rbinom(30,1,0.5)` that 30 times makes 1 draw of a random variable equal to one with probability 0.5 (and hence equal to zero with probability 0.5). The `for` loop repeats this 400 times, and the resulting means for each sample are stored in the array `result.mean` and the standard deviations in the array `result.stdev`. The `set.seed(10101)` command provides a starting value for the initial draw of `x`, leading to the same results each time this code is run.

The following Gretl code obtains 400 sample means in the coin toss example of Chapter 3.3, as well as standard deviations.

```
# Mean of 400 coin toss samples each of size 30
nulldata 30
set seed 10101
loop 400 --progressive
    genr u = uniform(0,1)
    genr x = (u > 0.5)
    scalar tosses = nobs(x)
    scalar mean = mean(x)
    scalar stdev = sd(x)
    # print out results
    print tosses mean stdev
    # and save results in gretl dataset
    store aed03simresults.gdt tosses mean stdev
endloop
# Summarize the 400 means
clear
open aed03simresults.gdt
summary --simple
```

3.9 Key Concepts

1. A random variable is a variable whose value is determined by the outcome of an experiment.

2. Random variables X are denoted in upper case and realized values x are denoted in lower case.

3. The mean μ is the probability-weighted average of all values that the random variable X may take.

4. The variance σ^2 is the probability-weighted average of all values that $(X - \mu)^2$ may take.

5. The population standard deviation is σ.
6. If X has mean μ and variance σ^2 then $a + bX$ has mean $a + b\mu$ and variance $b^2\sigma^2$.
7. $Y = (X - \mu)/\sigma$ has mean 0 and variance 1.
8. Statistical inference seeks to infer properties of the distribution of X from the sample at hand.
9. A sample of size n has observed values $x_1, x_2, ..., x_n$ that are realizations of the random variables $X_1, X_2, ..., X_n$.
10. The sample statistics, such as the sample mean, are random variables whose statistical properties are determined by those of the random variables whose realizations produced the sample.
11. In particular, the sample mean $\bar{x}$ is a realization of the random variable $\bar{X}$.
12. We assume that (A) $\mathrm{E}[X_i] = \mu$, (B) $\mathrm{Var}[X_i] = \sigma^2$, and (C) X_i is statistically independent of X_j, $i \neq j$.
13. A simple random sample is one whose observations are independent draws from the same distribution with $X_i \sim (\mu, \sigma^2)$. Then assumptions A-C are satisfied.
14. Under assumptions A-C the sample mean $\bar{x}$ is the realization of a random variable $\bar{X}$ that has mean equal to the population mean μ and standard deviation $\sigma/\sqrt{n}$ that gets smaller as the sample size increases.
15. The estimated standard deviation of $\bar{X}$, called the standard error of $\bar{X}$, is denoted $se(\bar{x})$.
16. Under assumptions A-C, $se(\bar{x}) = s/\sqrt{n}$.
17. Under assumptions A-C the standardized random variable $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is standard normal distributed as the sample size goes to infinity. For large n a good approximation is that $\bar{X} \sim N(\mu, \sigma^2/n)$.
18. A parameter is a constant that determines in part the distribution of X. An estimator is a method for estimating a parameter. An estimate is the particular value obtained from the sample.
19. An unbiased estimator of a parameter is a statistic whose expected value equals the parameter.
20. A consistent estimator of a parameter is a statistic that is almost certainly arbitrarily close to the parameter, as the sample size gets very large.
21. A best estimator or efficient estimator has minimum variance among the class of consistent estimators (or the class of unbiased estimators).
22. Under assumptions A-C, the sample mean is unbiased and consistent. Furthermore it is the best estimator in the special cases that the distribution of X is normal, Bernoulli, binomial, Poisson or exponential.

23. Adjustment to methods may be needed if the sample is not a simple random sample.

24. Key Terms: random variable; mean; variance; standard deviation; sample; assumptions A-C; simple random sample; sample mean; sample standard deviation; standard error of the sample mean; central limit theorem; normal distribution; standard normal distribution; asymptotically normal; parameter; estimate; estimator; unbiased; consistent; best estimator; minimum variance; nonrandom samples; weighted mean; random number generator.

3.10 Exercises

1. Let X denote annual health costs for an individual and suppose $X = 1000$ with probability 0.8 and $X = 5000$ with probability 0.2.

   (a) Obtain $\mu = \mathrm{E}[X]$ from first principles.
   (b) Obtain $\sigma^2 = \mathrm{E}[(X - \mu)^2]$ from first principles.
   (c) Hence find the standard deviation of X.

2. Repeat the previous exercise if $X = 0$ with probability 0.5, $X = 2000$ with probability 0.3 and $X = 12000$ with probability 0.2.

3. Suppose X has mean 5 and variance 4. For each of the following give the mean and variance.

   (a) $X + 3$. (b) $2X$. (c) $2X + 3$. (d) $(X - 5)/2$.

4. Suppose X has mean μ and variance σ^2. For each of the following give the mean and variance.

   (a) $Y = (X - \mu)$. (b) $Y = X/\sigma$. (c) $Y = (X - \mu)/\sigma$.

5. Let $\bar{X}$ be the mean of a simple random sample of size 100 from a random variable that is distributed with mean 200, variance 400, and a distribution that is not the normal distribution.

   (a) Give the mean of $\bar{X}$.
   (b) Give the variance and standard deviation of $\bar{X}$.
   (c) Is $\bar{X}$ likely to be approximately normally distributed? Explain.

6. Repeat the previous exercise for a simple random sample of size 400 from a random variable that is distributed with mean 400, variance 200.

7. Use a computer and a random number generator to obtain 1000 random numbers between 0 and 1, setting the seed to 10101. These are generated in such a way that they can be viewed as independent draws of the uniform random variable X with mean $\mu = 0.5$ and variance $\sigma^2 = 1/12$.

   (a) Are the sample mean and sample variance approximately equal to μ and σ^2?

   (b) How many of the 1,000 random numbers do you expect to lie between 0.0 and 0.1, and between 0.1 and 0.2, etc.? Hint: Any value between 0 and 1 is equally likely.

   (c) Plot a histogram of the random numbers drawn, with starting value 0, 10 bins, and frequency on the vertical axis. Do you (approximately) get what you expected from part (b).

   (d) Give a scatter plot of the random numbers against the observation number. Do they appear to be randomly draws between 0 and 1?

   (e) Give a line plot of the random number against observation number for the first 50 observations. Do consecutive random numbers appear to be related to each other, or do they appear to be independent?

8. For random sampling from $X \sim (\mu, \sigma^2)$ state which of the following statements are true

   (a) $\bar{X} = \mu$. (b) $\bar{X}$ has population mean μ. (c) $\bar{X}$ has population variance σ^2.

9. Consider simple random sampling from $X \sim (\mu, \sigma^2)$. State what happens to the size of $\mathrm{E}[\bar{X}]$, $\mathrm{Var}[\bar{X}]$ and the standard deviation of $\bar{X}$ when the sample size is made four times as large.

10. For simple random sampling from $X \sim (\mu, \sigma^2)$ state which of the following statements are true

    (a) $\mathrm{E}[\bar{X}] = \mathrm{E}[X]$. (b) $\mathrm{Var}[\bar{X}] = \mathrm{Var}[X]$. (c) $\bar{X}$ has standard deviation σ/N.

11. Let $X = 1$ with $\Pr[X = 1] = 1/6$ and $X = 0$ with $\Pr[X = 0] = 5/6$. (One way this would arise is if we tossed a six-sided die and set $X = 1$ if a five, say is obtained, and let $X = 0$ otherwise.)

    (a) Obtain $\mu = \mathrm{E}[X]$ from first principles.

    (b) Obtain $\sigma^2 = \mathrm{E}[(X - \mu)^2]$ from first principles.

    (c) Now use a computer and a random number generator to obtain a sample of size 100 for this example. Hint: A random number is less than 1/6 with probability 1/6.

    (d) Compare the mean $\bar{x}$ and variance s^2 of this sample to your answers in parts a-b.

    (e) Obtain the histogram. Does X appear to be normally distributed?

12. The preceding computer experiment was run 400 times, yielding 400 samples of size 100. The sample mean $\bar{x}$ for each sample is given in dataset DIETOSS.

    (a) Obtain the descriptive statistics for the 400 values of $\bar{x}$. Are the mean and standard deviation what you expect? Explain.

    (b) Obtain the histogram (or better still the kernel density estimate). Is this what you expect? Explain.

13. Suppose X takes values 1, 2 and 3 with probabilities of, respectively, 0.4, 0.2 and 0.4.

    (a) Obtain $\mu = [X]$ from first principles.

    (b) Obtain $\sigma^2 = \mathrm{E}[(X - \mu)^2]$ from first principles.

    (c) Now use a computer and a random number generator to obtain a sample of size 1,000 for this example. Hint: Let u be the random number. Then $x = 1$ if $u < 0.4$, $x = 2$ if $0.4 \leq u < 0.6$, and $x = 3$ if $u \geq 0.6$.

    (d) Compare the mean and standard deviation of this sample to your answers in parts a-b.

14. The preceding computer experiment was run 400 times, obtaining 400 samples of size 100. The sample mean $\bar{x}$ for each sample is given in dataset ONETWOTHREE.

    (a) Obtain the descriptive statistics for $\bar{x}$. Are the mean and standard deviation what you expect? Explain.

    (b) Obtain the histogram (or better still the kernel density estimate). Is this what you expect? Explain.

15. An insurance company offers insurance to 10,000 people with independent loss distributions that have mean \$5,000 and standard deviation \$20,000. Let $\bar{X} = \frac{1}{10000} \sum_{i=1}^{10000} X_i$ denote the average loss per individual.

    (a) Find the mean and standard deviation of $\bar{X}$.

    (b) Suppose the insurance company sells insurance that provides complete coverage for \$5,400. For simplicity suppose that the insurance company has no costs aside from paying out any insurance claims. Is the insurance company likely to make a loss? Explain your answer. Hint: By the central limit theorem $\bar{X}$ is normally distributed.

16. Repeat the previous exercise when the insurance pool is 2,500 people with independent loss distributions that have mean \$10,000 and standard deviation \$30,000 with $\bar{X} = \frac{1}{2500} \sum_{i=1}^{2500} X_i$.

17. The dataset TDIST4 has the sample means $\bar{x}$ and corresponding standard standard deviations s from 1000 random samples of size 4 where $X \sim N(100, 16^2)$.

    (a) Obtain the descriptive statistics for $\bar{x}$. Are the mean and standard deviation what you expect? Explain.

    (b) Obtain the descriptive statistics for $se(\bar{x}) = s/\sqrt{n}$ for these data. Is the mean what you expect? Explain.

    (c) Compute $z = (\bar{x} - 100)/8$. Explain why z is standard normal distributed.

    (d) Obtain summary statistics for z, a histogram and a kernel density estimate.

    (e) For these data does z appear to be standard normally distributed? Explain using results in (d).

18. The dataset TDIST25 has the sample means $\bar{x}$ and and corresponding standard standard deviations s from 1000 random samples of size 25 where $X \sim N(200, 50^2)$.

    (a) Obtain the descriptive statistics for $\bar{x}$. Are the mean and standard deviation what you expect? Explain.

    (b) Obtain the descriptive statistics for $se(\bar{x})$ for these data. Is the mean what you expect? Explain.

    (c) Compute $z = (\bar{x} - 200)/10$. Explain why z is standard normal distributed.

    (d) Obtain summary statistics for z, a histogram and a kernel density estimate.

    (e) For these data does z appear to be standard normally distributed? Explain using results in (d).

19. State whether the following samples are likely to be representative or nonrepresentative of the population.

    (a) Every twentieth person is sampled. All people respond.

    (b) Every twentieth person is sampled. But only ten percent of those sampled respond.

    (c) Every person is sampled. Only ten percent of those sampled respond. We question every twentieth person who did respond.

20. Suppose we take a simple random sample from $X \sim (\mu, \sigma^2)$.

    (a) We estimate μ by X_1, the first value of X in the sample. Is this estimator unbiased for μ? Is this estimator consistent for μ? Explain.

    (b) We estimate μ by $\bar{X} + \frac{1}{n}$. Is this estimator unbiased for μ? Is this estimator consistent for μ? Explain.

    (c) Now suppose $X \sim N(\mu, \sigma^2)$ and we estimate μ by an estimator $\widetilde{\mu}$ that has $\mathrm{E}[\widetilde{\mu}] = \mu$ and $\mathrm{Var}[\widetilde{\mu}] = 2\sigma^2/n$. Is this estimator a best unbiased estimator for μ? Explain.

21. The dataset AUSREGWEALTH has data on average net worth of households (x_i) in thousands of dollars in 517 regions in Australia in 2003-04. Calculate the weighted mean where weight by household size as follows. This is an example of weighting by frequency weights.

    (a) Let w_i equal number of households in each region. Compute $\sum_{i=1}^{n} w_i$.
    (b) Generate the variable $w_i x_i$ and hence the weighted mean $\bar{x}_w = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i$.
    (c) Compare the weighted mean to the unweighted mean.
    (d) Calculate the weighted variance as $\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2 / \sum_{i=1}^{n} w_i$.
    (e) Compare the weighted standard deviation to the unweighted standard deviation.
    (f) If your software computes weighted means and standard deviation, reproduce these results using your software.

22. Repeat the previous exercise for the annual income per taxpayer (in dollars).

23. Repeat exercise 17 except generate the 1,000 sample means yourself. For Stata, use the Stata code in Chapter 3.8 except replace `set obs 30` with `set obs 4`, replace commands `generate u=runiform()` and `generate x=u>0.5` with `generate x=rnormal(100,16)`, and replace `reps(400)` with `reps(1000)`. For R use the R code in Chapter 3.8 except replace `400` with `1000` and replace `x=rbinom(30,1,0.5)` with `x=rnorm(4,100,16)`.

24. Repeat exercise 18 except generate the 1,000 sample means yourself. For Stata, use the Stata code in Chapter 3.8 except replace `set obs 30` with `set obs 25` and replace commands `generate u=runiform()` and `generate x=u>0.5` with `generate x=rnormal(200,50)`, and replace `reps(400)` with `reps(1000)`. For R use the R code in Chapter 3.8 except replace `400` with `1000` and replace `x=rbinom(30,1,0.5)` with `x=rnorm(25,200,50)`.

Chapter 4

Statistical Inference for the Mean

The sample mean $\bar{x}$ is a random outcome – different samples lead to a different value of the sample mean. To deal with this randomness, the sample at hand is viewed as being one from sampling observations on a random variable X that has mean (or expected value) denoted by μ. The goal is to make inference on μ given the observed sample mean $\bar{x}$. For example, is the view that mean earnings in the population equal \$40,000, say, consistent with sample mean earnings equal to \$41,413?

This chapter analyzes inference based on the sample mean. It presents the fundamentals of statistical inference, notably confidence intervals and hypothesis tests. Confidence intervals give a range of plausible values of μ given the sample. Hypothesis tests are used to determine whether or not a specified value of μ or range of values of μ is plausible, given the sample.

While the focus is on statistical inference for the mean, these concepts carry over to other univariate statistics, such as the median, and to regression, the subject of most of this book. A good understanding of statistical inference is essential as it lies at the heart of analysis of economics data.

The chapter continues directly from the previous chapter. For readers who bypassed the details in Chapter 3, the key results for statistical inference on the mean are presented in Chapter 3.4.

4.1 Example: Mean Annual Earnings

The following example presents the methods of statistical inference that will be explained in this chapter.

Dataset EARNINGS introduced in Chapter 2 has data on individual annual earnings for a sample of 30 year-old full-time workers in 2010.

Table 4.1 presents several key sample statistics that are generated by a descriptive statistics command, such as the Stata `summarize` command.

The **sample mean** $\bar{x} = 41412.69$ and the **sample standard deviation** $s = 25527.05$.

The population considered is all 30 year-old female full-time workers in 2010 in the United States, with unknown population mean earnings denoted μ. We wish to make inference about the

Table 4.1: Summary statistics: Annual earnings of female full-time workers aged 30 in 2010 (n=171).

Variable	Obs	Mean	Std. Dev.	Min	Max
Earnings		41412.69	25527.05	1050	172000

mean μ, using data from the sample which is a simple random sample from the population. The standard tools of inference are confidence intervals and hypothesis tests.

Table 4.2 presents key results for inference on the mean produced by a command such as the Stata `mean` command.

Table 4.2: Confidence interval: Annual earnings.

Variable	Mean	Stand. Error	95% Conf.	Interval
Earnings	41412.69	1952.10	37559.21	45266.17

The entry Mean is the **sample mean** $\bar{x}$ and is the commonly-used estimate of μ. Here $\bar{x} = 41412.69$, so the estimate of mean earnings in the population of all 30 year-old female full-time workers in 2010 is \$41,413.

The entry Stand. Error is the **standard error of the sample mean**, where standard error is the statistical term for estimated standard deviation. This measures the precision of the sample mean $\bar{x}$ as an estimate of μ. A smaller standard error means greater precision of $\bar{x}$ as an estimate of μ. Here the standard error of the sample mean equals \$1,952. This is much smaller than the sample standard deviation of \$25,227 for just one observation, because averaging reduces the variability. In fact, under simple random sampling the standard error equals the sample standard deviation of a single observation divided by the square root of the sample size. Here $s/\sqrt{n} = 25527/\sqrt{171} = 1952$.

The entry 95% Conf. Interval gives a 95% **confidence interval** that provides a range of values that includes the true (unknown) population mean μ with 95% confidence. Here the 95% confidence interval for population mean earnings is (\$37,559, \$45,266).

An **hypothesis test** is a test of whether or not the data support a hypothesized value or range of values for the population mean μ. As an example we test the hypothesis that $\mu = 40000$ against the alternative that $\mu \neq 40000$. A command such as the Stata command `ttest earnings=40000` produces the following output.

Some of the output in Table 4.3 repeats that from the command `mean earnings`. Additionally it provides t = 0.7237, called the t statistic, `degrees of freedom` = 170, and the results of three related hypothesis tests. For test of $\mu = 40000$ against $\mu \neq 40000$ the middle output with $\Pr(|T| < |t|) = 0.4703$ is relevant. This value of 0.4703 is called the p-value of the test. It is common to test at significance level 0.05, in which case we would not reject the hypothesis that $\mu = 40000$ since the p-value $0.4703 > 0.05$.

This example presents the key methods for statistical inference on the population mean based on the sample mean. The remainder of this chapter provides complete explanation.

Table 4.3: Hypothesis test: Annual earnings have mean equal to 40000.

Variable	Obs	Mean	Stand. Error	Stand. Dev.	95% Conf.	Interval
Earnings	171	41412.69	1952.03	25527.05	37559.21	45266.17
mean = mean(earnings)						t = 0.7237
Ho: mean = 40000						degrees of freedom = 170
Ha: mean < 40000			Ha: mean != 40000			Ha: mean > 40000
Pr(T < t) = 0.7649			Pr(|T| < |t|) = 0.4703			Pr(T > t) = 0.2351

4.2 t Statistic and t Distribution

Interest lies in estimating the mean μ, and we use the sample mean as the estimator. Chapter 3 detailed properties of the sample mean and why it is a good estimator of μ.

Now we wish to construct confidence intervals on μ, and perform hypothesis tests on μ, which requires knowledge of the distribution of the sample mean. As detailed in Chapter 3 and repeated below, under certain assumptions the sample mean is normally distributed with mean μ and variance σ^2/n; see Chapter 3.4.7 for a summary.

However, to immediately use this result requires knowledge of σ^2/n. In practice this is not known, so we instead estimate it by s^2/n where s is the sample standard deviation of X. Since s is an estimate this adds noise that leads to inference based on the t distribution, a distribution that has fatter tails than the standard normal. In this section we focus on how to obtain probabilities for the t distribution.

4.2.1 Normal distribution

A sample of size n has observed values $x_1, x_2, ..., x_n$ that are realizations of the random variables $X_1, X_2, ..., X_n$. Then $\bar{x}$ is the realization of the random variable $\bar{X} = (X_1 + \cdots X_n)/n$. The properties of $\bar{X}$ depend on the properties of $X_1, X_2, ..., X_n$.

We assume a simple random sample so that the underlying random variables X_i

A. have common mean μ: $\mathrm{E}[X_i] = \mu$ for all i.

B. have common variance σ^2: $\mathrm{Var}[X_i] = \sigma^2$ for all i.

C. are statistically independent: X_i is statistically independent of $X_j, i \neq j$.

Under these assumptions, the central limit theorem states that if additionally the sample size is large then $\bar{X}$ is normally distributed, regardless of the actual distribution of X, and the standardized random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \text{ as } n \to \infty.$$

4.2.2 The t Statistic

In practice the sample standard deviation σ is unknown and we need to replace it by the standard deviation of X. Then the **distribution of the sample mean** $\bar{X}$ is defined in terms of the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

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

The distribution of the random variable T is in general complicated. The standard approximation is to suppose that

$$T \sim T(n-1),$$

where $T(n-1)$ denotes the t **distribution** with $(n-1)$ degrees of freedom.

Different degrees of freedom correspond to different t distributions just as, for example, different means μ would correspond to different normal distributions. The term **degrees of freedom** is used because the relationship $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ implies that only $(n-1)$ terms in the sum are free to vary.

The reason for using the $T(n-1)$ distribution is that T can be shown to be exactly $T(n-1)$ distributed under assumptions A-C and the additional assumption that X is normally distributed. When X is not normally distributed a common rule-of-thumb is that the $T(n-1)$ distribution is generally a good approximation if $n > 30$.

We observe a single sample with sample mean $\bar{x}$, sample standard deviation s, sample standard error $se(\bar{x}) = s/\sqrt{n}$, and corresponding sample value of the t statistic. So the sample t statistic is a single realization of a $T(n-1)$ distributed random variable. For simplicity we write

$$t = \frac{\bar{x} - \mu}{se(\bar{x})} \sim T(n-1).$$

A common rule-of-thumb is that the approximation will be a good one if $n > 30$.

Remark 9 *From a **simple random sample** $x_1, ..., x_n$ calculate the **sample mean** $\bar{x}$, the sample standard deviation s and the **standard error of** $\bar{x}$, $se(\bar{x}) = s/\sqrt{n}$. The **t statistic***

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

*is a realization of a random variable that is approximately $T(n-1)$ distributed, where $T(n-1)$ denotes the **t distribution** with $n-1$ degrees of freedom.*

To form the t statistic the only summary statistics of the sample needed are the sample mean $\bar{x}$ and the sample standard deviation s. Additionally the t statistic depends on the population mean μ, which is unknown. The knowledge that the t statistic is $T(n-1)$ distributed is used to make statistical inference on μ, as detailed in subsequent sections of this chapter.

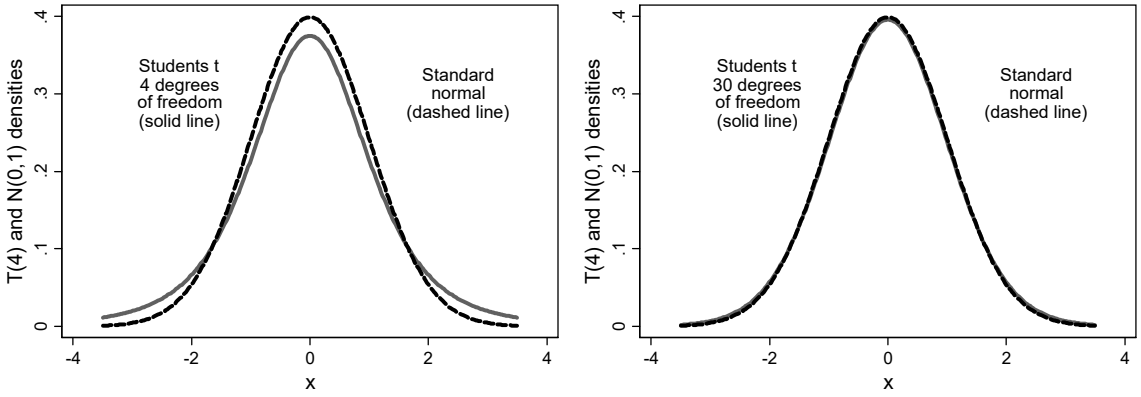

Figure 4.1: Student's t distribution: $T(4)$ and $T(30)$ compared to the standard normal.

4.2.3 The t Distribution

A t-distributed random variable is a continuous random variable. In that case probabilities are given by the area under the probability density function; see Appendix B.1. For example, $\Pr[a < T < b]$ is the area under the curve from a to b. The formula for the probability density function is complex and is not given here. Instead the properties of the t distribution are outlined.

The probability density function for the t **distribution**, or **Student's** t **distribution**, is a bell-shaped curve centered on zero, and symmetric about zero, that is a slightly squashed version of the standard normal. It has one parameter, denoted v here, called the **degrees of freedom**. The t distribution with v degrees of freedom, denoted $T(v)$, has mean 0, provided $v > 1$, and variance $v/(v-2)$, provided $v > 2$. The standard normal has the same mean of 0, but smaller variance of 1.

Figure 4.1 presents, respectively, the $T(4)$ and $T(30)$ probability density functions and compares them in each case to the standard normal. The T distribution has bell-shaped curve similar to the standard normal distribution, except it has fatter tails reflecting increased randomness due to replacing the constant σ by the random variable S. The t distribution approaches the standard normal as v gets larger and the difference disappears as $v \to \infty$.

4.2.4 Probabilities for the t Distribution

There is no simple formula for **probabilities** for the t distribution; instead computation requires advanced numerical methods. Until recently statisticians needed to refer to published tables such as those given in Appendix E. Now one can directly use a computer.

For example, to compute $\Pr[T_{30} > 2]$, the probability that a $T(30)$ random variable exceeds 2, one can use the Stata function `ttail(30,2)` which returns a probability of 0.0273.

From the second panel of Figure 4.1 there is seemingly little difference between the $T(30)$ and the standard normal distributions. But there is still an appreciable difference in the tails of these

distributions, with the t distribution having fatter tails. In fact $\Pr[|T_{30}| > 2] = 0.0546$. This is approximately 20% larger than $\Pr[|Z| > 2] = 0.0455$ for Z standard normal distributed.

Such differences can be large enough to matter for confidence intervals and hypothesis tests because they use tail probabilities. For this reason statistical packages and this book base inference on the t distribution rather than the standard normal distribution.

As the degrees of freedom $v \to \infty$, the difference disappears since the t distribution then collapses to the standard normal distribution. For example, for $T_{1000} \sim T(1000)$ we have $\Pr[|T_{1000}| > 2] = 0.0458$, very close to 0.0455 for the standard normal.

Remark 10 *The t distribution with v degrees of freedom, denoted $T(v)$, is like a squashed version of the standard normal distribution with fatter tails. As $v \to \infty$ the t distribution goes to the standard normal.*

4.2.5 Inverse Probabilities for the t Distribution

In some situations this computation needs to be inverted. The probability is set and we wish to calculate the associated value of t that gives this probability.

For example we may wish to find the value c such that the probability that a $T(170)$ distributed random variable exceeds c is equal to 0.05. Then, for example, one can use the Stata function `invttail(170,.05)` which returns a value of 1.6539. We have that $c = 1.6539$ solves $\Pr[T_{170} > c] = 0.05$. Appendix A of this book includes corresponding commands for various statistical packages other than Stata.

More generally the desired area is denoted α, the greek letter "alpha", and the **inverse probability**, called a **critical value**, $c = t_{v,\alpha}$ satisfies

$$\Pr[T_v > t_{v,\alpha}] = \alpha.$$

In words, the inverse probability or critical value $t_{v,\alpha}$ is that value such that a $T(v)$ distributed random variable exceeds $t_{v,\alpha}$ with probability α. Even more simply, **the area under the curve to the right of $t_{v,\alpha}$ equals α.**

The left panel of Figure 4.2 presents the example $\Pr[T_{170} > 1.654] = 0.05$. Then $\alpha = 0.05$ is the shaded area in the right tail, and the inverse probability $t_{v,\alpha} = t_{170,.05} = 1.654$ is given on the horizontal axis.

Definition 9 *The inverse probability or **critical value** $c = t_{v,\alpha}$ is that value for which a $T(v)$ distributed random variable exceeds $t_{v,\alpha}$ with probability α, i.e. $\Pr[T_v > t_{v,\alpha}] = \alpha$.*

Sometimes we want the **combined area** in left and right tails to equal α. Given the symmetry about 0 of the t distribution we have

$$\Pr[|T_v| > t_{v,\alpha/2}] = \Pr[T_v < -t_{v,\alpha/2}] + \Pr[T_v > t_{v,\alpha/2}] = \alpha/2 + \alpha/2 = \alpha.$$

The **combined area under the curve to the left of $-t_{v,\alpha/2}$ and to the right of $t_{v,\alpha/2}$ equals α.**

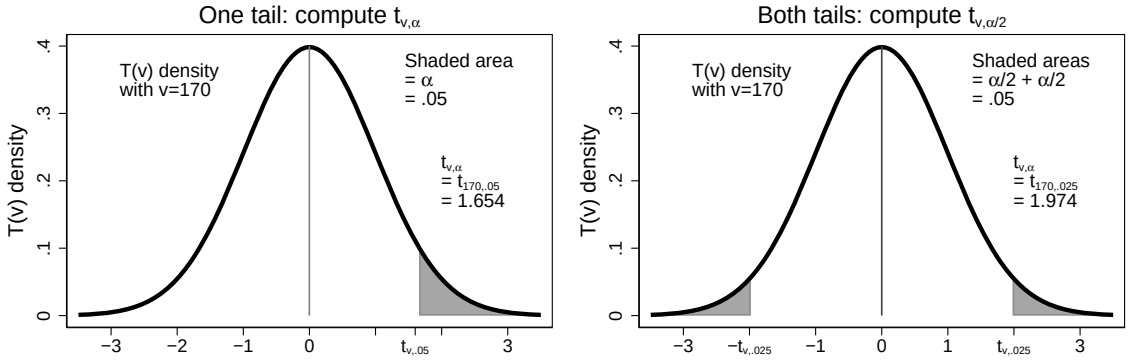

Figure 4.2: Student's t distribution: Critical values $t_{v,\alpha}$ and $t_{v,\alpha/2}$ for $v = 170$ and $\alpha = 0.05$.

For example, $\Pr[T_{170} > 1.974] = 0.025$, so $t_{170,.025} = 1.974$. Combining both tails of the t distribution it follows that $\Pr[|T_{170}| > 1.974] = 0.05$.

The right panel of Figure 4.2 presents the example $\Pr[|T_{170}| > 1.974] = 0.05$. Then the shaded area in each tail is 0.025, $\alpha = 0.05$ is the combined area in the two tails, and the critical value $t_{v,\alpha} = t_{170,.025} = 1.974$ is given on the horizontal axis.

Definition 10 *A t distributed random variable with v degrees of freedom exceeds* ***in absolute value*** *the* ***critical value*** *$t_{v,\alpha/2}$ with probability α, i.e.* $\Pr[|T_v| > t_{v,\alpha/2}] = \alpha$.

Note that some books define $t_{170,.05}$, for example, to be the .05 quantile or 5^{th} percentile, so the area in the **left tail** of the distribution is .05. Throughout this book, however, $t_{170,.05}$ is that value for which the area in the **right tail** of the distribution is .05. This makes no difference in practice due to the symmetry of the t distribution about zero, for example, $t_{170,.95} = -t_{170,.05}$.

4.3 Confidence Intervals

Different samples will lead to different estimates of the population mean. A confidence interval for an unknown parameter, such as the population mean, gives a range of values that the parameter lies in with a certain "confidence level", defined next.

4.3.1 95% Confidence Interval

A confidence interval for the unknown population mean μ is a range of values that might contain μ with a pre-specified frequency. For example, a 95% confidence interval for μ is a range of values that may contain μ with 95% frequency, i.e. if we had infinitely many samples and constructed infinitely many confidence intervals then 95% of these confidence intervals will include the true value of μ.

Under assumptions 1-3 we have

Definition 11 *A **95 percent confidence interval for the population mean** is*

$$\bar{x} \pm t_{n-1,.025} \times se(\bar{x}),$$

*where $\bar{x}$ is the sample mean; $t_{n-1,.025}$ is that value (called a **critical value**) such that a $T(n-1)$ distributed random variable exceeds it in absolute value with probability 0.025; and $se(\bar{x}) = s/\sqrt{n}$ is the standard error of the sample mean.*

For derivation see Chapter 4.3.3. The confidence interval is centered around $\bar{x}$, the estimate of μ, the sample mean $\bar{x}$, and is symmetric. The use of the $T(n-1)$ distribution is exact under the additional assumption that X is normally distributed; otherwise it is a commonly-used approximation. The specific value $t_{n-1,.025}$ is used since with area 0.025 in each tail the area in the center of the $T(n-1)$ distribution is 0.95, corresponding to 95% confidence.

Intuitively the confidence interval is narrower the more precise is our estimate of μ. This is indeed the case, as from the formula the confidence interval is narrower the smaller is the standard error of $\bar{x}$. In particular, we have the following result.

Remark 11 *The confidence interval narrows as the sample size gets larger, since larger samples lead to a smaller standard error.*

4.3.2 Example: Mean Annual Earnings

For the female annual earnings data in dataset EARNINGS, introduced in Chapter 2, $n = 171$, $\bar{x} = 41413$, $s = 25527$, and $se(\bar{x}) = s/\sqrt{n} = 1952$. From T_{170} tables, $t_{170,.025} = 1.974$.

It follows that a 95% confidence interval for population mean earnings of thirty year-old female full-time workers is

$$\bar{x} \pm t_{n-1,.025} \times se(\bar{x}) = 41413 \pm 1.974 \times 1952 = 41413 \pm 3853 = (37560,\ 45266).$$

This is the 95% confidence interval that was given in Chapter 4.1.

4.3.3 Derivation of a 95% Confidence Interval

We derive a 95% confidence interval from first principles. For simplicity consider a sample with $n = 61$, in which case $n - 1 = 60$ and $t_{60,.025} = 2.0003$. Thus

$$\Pr[-2.0003 < T_{60} < 2.0003] = 0.95,$$

which we round to $\Pr[-2 < T < 2] = 0.95$. Substituting $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ it follows that

$$\Pr\left[-2 < \frac{\bar{X} - \mu}{S/\sqrt{n}} < 2\right] = 0.95.$$

This interval can be converted to an interval that is centered on μ as follows

$$\begin{array}{rrll} & \Pr\left[-2 < \frac{\bar{X}-\mu}{S/\sqrt{n}} < 2\right] & = 0.95 & \\ \Rightarrow & \Pr\left[-2S/\sqrt{n} < \bar{X} - \mu < 2S/\sqrt{n}\right] & = 0.95 & \text{multiplying all terms by } S/\sqrt{n} \\ \Rightarrow & \Pr\left[-\bar{X} - 2S/\sqrt{n} < -\mu < -\bar{X} + 2S/\sqrt{n}\right] & = 0.95 & \text{subtracting } \bar{X} \text{ from all terms} \\ \Rightarrow & \Pr\left[\bar{X} + 2S/\sqrt{n} > \mu > \bar{X} - 2S/\sqrt{n}\right] & = 0.95 & \text{multiplying by } -1 \text{ reverses inequalities.} \end{array}$$

Re-ordering the final inequality yields

$$\Pr\left[\bar{X} - 2 \times S/\sqrt{n} < \mu < \bar{X} + 2 \times S/\sqrt{n}\right] = 0.95.$$

Replacing random variables by their observed values, the interval $(\bar{x} - 2 \times s/\sqrt{n},\ \bar{x} + 2 \times s/\sqrt{n})$ is called a 95% confidence interval for μ.

More generally with sample size n the critical value is $t_{n-1;.025}$. Then a 95% confidence interval is $(\bar{x} - t_{n-1,.025} \times se(\bar{x}),\ \bar{x} + t_{n-1,.025} \times se(\bar{x}))$.

4.3.4 What Level of Confidence?

Ideally there is both a high level of confidence and a narrow confidence interval. For example, having 95% confidence that μ lies between 20 and 40 is preferred to having only 90% confidence that μ lies between 20 and 40. And having 95% confidence that μ lies between 20 and 40 is preferred to having 95% confidence that μ lies in the broader range of 10 to 50.

Unfortunately there is a trade-off between these two considerations. In order to have greater confidence the confidence interval needs to widen. For example, to be 100% confident we can only say that μ lies in the range $(-\infty, \infty)$.

So what value of confidence should be used? There is no best value in general, but **it is most common to use a 95% confidence interval**.

More generally, we consider confidence intervals with **confidence level** $100(1-\alpha)\%$, in which case the critical value is $t_{n-1,\alpha/2}$ since the area in each tail is then $\alpha/2$ leaving area $1-\alpha$ in the center of the $T(n-1)$ distribution.

Definition 12 *A* $\mathbf{100(1-\alpha)}$ ***percent confidence interval for the population mean*** *is*

$$\bar{x} \pm t_{n-1,\alpha/2} \times se(\bar{x}).$$

The value $\alpha = 0.05$ corresponds to a 95% confidence interval since $100(1-.05) = 100 \times 0.95 = 95$. The other common choices are to use narrower 90% confidence intervals, with $\alpha = 0.10$, and wider 99% confidence intervals with $\alpha = 0.01$.

Table 4.4 presents the critical value $t_{v,\alpha/2}$ for various confidence levels, corresponding to different values of α, and for selected different numbers of observations, corresponding to different values of $v = n - 1$. The value $t_{v,\alpha/2}$ decreases as the sample size increases. For 95% confidence intervals, presented in bold, the t value is 2.042 for the t_{30} distribution falling to 1.960 for the t_∞ distribution which is equivalent to the standard normal distribution. A detailed table for the t distribution is provided in Table E.2 in Appendix E.

In typical econometrics applications the sample size $n > 30$, in which case from Table 4.4 the critical value $t_{n-1,.025}$ approximately equals 2. This leads to the following.

Table 4.4: Student's t distribution: Critical values for various degrees of freedom and confidence levels.

Confidence Level	$\mathbf{100(1-\alpha)}$	**90%**	**95%**	**99%**
Area in both tails	α	0.10	**0.05**	0.01
Area in single tail	$\alpha/2$	0.05	**0.025**	0.005
t value for $v = 10$	$t_{10,\alpha/2}$	1.812	**2.228**	3.169
t value for $v = 30$	$t_{30,\alpha/2}$	1.697	**2.042**	2.750
t value for $v = 100$	$t_{100,\alpha/2}$	1.660	**1.984**	2.626
t value for $v = \infty$	$t_{\infty,\alpha/2}$	1.645	**1.960**	2.576
standard normal value	$z_{\alpha/2}$	1.645	**1.960**	2.576

Remark 12 *It is most common, though arbitrary, to use a 95% confidence interval. An* ***approximate 95% confidence interval*** *for the population mean is a* ***two standard error interval****: the sample mean plus or minus two times the standard error.*

This is a useful guide. And it makes clear that if we are willing to tolerate an error range of plus or minus two standard errors, then a good choice of the confidence level is 95%. In any published work or in assignments, however, use the more precise interval $\bar{x} \pm t_{n-1,.025} \times se(\bar{x})$.

Confidence intervals at different levels of confidence are easily obtained using a statistical package. For example, a 90% confidence interval for earnings can be obtained using the Stata command `mean earnings, level(90).`

4.3.5 Interpretation of Confidence Intervals

Interpretation of confidence intervals is conceptually difficult. With a given sample we can only form one confidence interval, which will either correctly include the true unknown mean μ or not include μ. A 95% confidence interval is constructed so that it includes μ with probability 0.95.

To understand this interpretation it is necessary to imagine that there are many separate samples of the population, each of size $n = 171$ in this example. From each sample we form a 95% confidence interval. Then we expect that on average 95% of such confidence intervals will include the true (unknown) mean μ.

For the 1880 Census example in Chapter 3.5 we know $\mu = 24.13$. Further analysis of the 100 samples of size 25 summarized in dataset CENSUSAGEMEANS yields a 95% confidence interval (19.29, 36.39) for the first sample, (12.79, 26.00) for the second sample, and so on. In total 91 of the 100 samples had 95% confidence intervals that included $\mu = 24.13$. For example, the 20^{th} sample had 95% confidence interval (7.70, 22.38) that does not include $\mu = 24.13$. In theory we expect 95% of the 95% confidence intervals to include μ. The reason 91% rather than 95% included μ reflects randomness with just 100 confidence intervals, and that the $T(24)$ distribution for the t statistic is not exact for these right-skewed data. If we had obtained one million 95% confidence intervals, say, and the t statistic was exactly $T(24)$ distributed, then very close to 95% of these intervals would include μ. Similarly, for the coin toss example in Chapter 3.3, 388 of the 400 95% confidence intervals, or 97%, included the true parameter value $\mu = 0.5$.

As these examples demonstrate, a confidence interval will sometimes fail to include the population mean μ, due to the randomness inherent in sampling. A 95 percent confidence interval has the property that if we were able to obtain many separate random samples, then 95 percent of the resulting confidence intervals will include the population mean μ, and 5 percent will not.

In fact we have only one sample, and we say that the calculated 95 percent confidence interval from this sample includes the true population mean μ with probability 0.95. This probabilistic statement refers to the confidence interval, which is random, and not to μ which is fixed. It is **wrong** to instead interpret this confidence interval as meaning that with probability 0.95 the population mean μ lies inside (\$37,560, \$45,266) and with probability 0.05 it lies outside this range.

Remark 13 *A calculated 95 percent confidence interval for the population mean is an interval that if constructed for each of an infinite number of samples will include the true population mean μ 95% of the time (and will not include μ 5% of the time).*

4.4 Two-Sided Hypothesis Tests

4.4.1 Null and Alternative Hypotheses

The particular hypothesis under test is called the **null hypothesis** and is denoted H_0. The alternative to the hypothesis test is called the **alternative hypothesis** and is denoted H_a.

Here we consider test of whether μ takes a particular value. Let μ^* denote this value. Then the null hypothesis is $H_0 : \mu = \mu^*$, and the alternative hypothesis is $H_a : \mu \neq \mu^*$. Because the alternative hypothesis includes both $\mu < \mu^*$ and $\mu > \mu^*$ the test is called a two-sided test.

For example, consider the claim that population mean earnings equal \$40,000. To test this claim we test $H_0 : \mu = 40000$ against $H_a : \mu \neq 40000$.

Definition 13 *A **two-sided test** or **two-tailed test** for the mean μ is a test of the null hypothesis*

$$H_0 : \mu = \mu^*,$$

where μ^* is a specified value for μ, against the alternative hypothesis

$$H_a : \mu \neq \mu^*.$$

4.4.2 Significance Level of a Test

The result of a test is to either reject or not reject the null hypothesis.

This decision made may be in error. In particular, we may reject the null hypothesis when in fact it was true. This type of error is called a **type I error**.

For example, suppose the null hypothesis is that someone is innocent. Then a type I error is made if we reject the null and find the person guilty, when in fact the person was innocent. Similarly if the null hypothesis is that a person does not have a disease, then a type I error is to find disease when in fact none is present.

In the earnings example a type I error occurs if we reject $H_0 : \mu = 40000$ when in fact $\mu = 40000$.

Definition 14 *A **type I error** occurs if H_0 is rejected when H_0 is true.*

Ideally the probability of making a type I error is small. The following terminology is used.

Definition 15 *The **significance level** of a test or **test size**, denoted α, is the pre-specified maximum probability of a type I error that will be tolerated.*

The level of statistical significance to use is discussed in some length in a later section. It is most common to tolerate up to a 5% chance of making a type I error, in which case $\alpha = 0.05$.

Whatever the choice of α, we reject H_0 at significance level α if the probability of making a type I error is less than 0.05, and do not reject H_0 otherwise.

Note that if we do not reject the null hypothesis then we simply say that we "fail to reject the null hypothesis." We do not say that we "accept the null hypothesis." The reason for doing so is that there are other null hypotheses that we also fail to reject. For example, in the earnings example we show below that we do not reject $H_0 : \mu = 40000$. But for these data other null hypotheses, such as $H_0 : \mu = 41000$, will also be not rejected at level 0.05.

4.4.3 The t Statistic

The obvious decision rule is to reject the hypothesis that the mean equals μ^* if the sample mean $\bar{x}$ is far from μ^*. For example, we are much more likely to reject $H_0 : \mu = 40000$ if $\bar{x} = 80000$ than if, say, $\bar{x} = 45000$. So we form a test based on the difference $(\bar{x} - \mu^*)$.

Furthermore, the more precise $\bar{x}$ is as an estimate of μ, the more likely we are to reject H_0 for a given size of $(\bar{x} - \mu^*)$. The test statistic we use therefore normalizes by the standard error of $\bar{x}$. We therefore use the t statistic

$$t = \frac{\bar{x} - \mu^*}{se(\bar{x})}.$$

This has the additional advantage that we know t is the realization of a random variable that is $T(n-1)$ distributed.

Remark 14 *The t **statistic** for test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$ is $t = (\bar{x} - \mu^*)/se(\bar{x})$. Under $H_0 : \mu = \mu^*$, and assuming simple random sampling, t is the realization of a random variable that is approximately $T(n-1)$ distributed.*

For the data on earnings of 30 year-old female full-time workers in 2010 in the United States, $n = 171$, $\bar{x} = 41413$ and $se(\bar{x}) = 1952$.

For test of $H_0 : \mu = 40000$, the t statistic is therefore

$$t = \frac{\bar{x} - 40000}{se(\bar{x})} = \frac{41413 - 40000}{1952} = 0.724.$$

This is the value `t=0.7237` obtained in Chapter 4.1. Under the null hypothesis, that $\mu = 40000$, the t statistic is a draw from the T_{170} distribution, since $n - 1 = 170$.

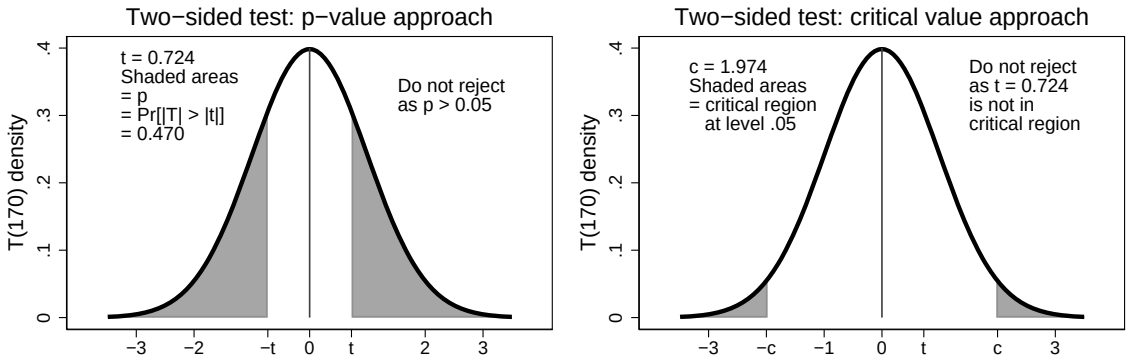

Figure 4.3: Two-sided hypothesis test: p-value approach and critical value approach

4.4.4 Rejection using p-values

We reject the null hypothesis if the t statistic is unusually large in absolute value. How unusual the value is can be determined since, under the null hypothesis, t is the realization of a $T(n-1)$ distributed random variable. If the value is so large as to be very unlikely to arise then our hypothesis that $\mu = \mu^*$ is most likely wrong and we reject the null hypothesis.

Definition 16 *The p-**value** is the probability of observing a t statistic at least as large in absolute value as that obtained in the current sample. For a two-sided test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$ the p-value is $p = \Pr[|T_{n-1}| \geq |t|]$.*

Definition 17 *H_0 is rejected at significance level α if $p < \alpha$, and is not rejected otherwise.*

Continuing immediately with the earnings example, we found that $t = 0.724$. Then

$$p = \Pr[|T_{170}| \geq 0.724] = 0.470.$$

This is the value `Pr(|T|>|t|)=0.4703` obtained in Chapter 4.1. There is high probability ($p = 0.470$) of observing a t value of 0.724 or larger in absolute value, even if the mean really is the hypothesized value of \$40,000. Since $p > 0.05$ we do not reject the null hypothesis that $\mu = 40000$ at significance level 0.05.

The left panel of Figure 4.3 displays the probability that $|T_{170}|$ exceeds the observed t statistic of 0.724.

We are more likely to reject the null hypothesis the larger is the absolute value of the t statistic. So, other things equal, we are more likely to reject the null hypothesis the smaller is $se(\bar{x})$, since computation of t entails division by $se(\bar{x})$. So a larger sample is better, as then $se(\bar{x})$ is smaller.

Remark 15 *More precise estimation of μ, such as through a larger sample size, makes it more likely that the null hypothesis is rejected.*

4.4.5 Rejection using Critical Regions

The p-value method requires access to a computer, in order to precisely compute the p-value for any possible value of t. Before widespread access to computers, an alternative method was used that, for given significance level α, leads to the same conclusion.

This **alternative method** defines a **critical region** or **rejection region**, which is the range of values of t that would lead to rejection of H_0 at the specified significance level α. Then reject H_0 if the computed value of t falls in this range.

Definition 18 *For a two-sided test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$, and for specified significance level α, the **critical value** c is such that $c = t_{n-1,\alpha/2}$; equivalently $\Pr[|T_{n-1}| \geq c] = \alpha$.*

Definition 19 *H_0 is rejected at significance level α if $|t| > c$, and is not rejected otherwise.*

Return to the female earnings example. For significance level 0.05 and $n - 1 = 170$, the critical value

$$c = t_{170,.025} = 1.974.$$

H_0 is not rejected at significance level 0.05, since $t = 0.724$ does not exceed 1.974 in absolute value. This conclusion is the same as that using the p-value approach.

The shaded region in the right panel of Figure 4.3 shows the rejection region, the range of values for which $|T_{170}|$ exceeds 1.974 since this occurs with probability 0.05. The sample t statistic equals 0.724 which does not fall in the shaded region. So we do not reject H_0.

It is important to note that the p-value and critical region approaches lead to the same conclusion, since $|t| > t_{n-1,\alpha/2}$ is equivalent to $p < \alpha$.

The critical value approach has the advantage that it does not require computing the p-value for any possible value of t. Instead one can refer to a reasonable-sized printed table for the t distribution at just a few selected probability values. Typically t tables are given for area in the right tail equal to 0.25, 0.10, 0.05, 0.01, and 0.005, and for all degrees of freedom from 1 to 30, every fifth integer from 30 to 60, then 70, 80, 90, 100 and ∞. To alternatively calculate p-values for a wide range of values of t would require many, many pages of tables.

The p-value approach is preferred as given p, the reader can easily test using his or her own preferred value of α. The alternative critical value method was developed for an earlier time when reliance on published tables made it difficult to accurately calculate p-values.

4.4.6 Which Significance Level?

Decreasing the significance level α makes it less likely that the null hypothesis is rejected. This should be clear from, for example, the second panel of Figure 4.3, where the rejection region will get smaller as the error α in the two tails gets smaller. What significance level should be used?

Remark 16 *It is most common to use $\alpha = 0.05$, called a test at the 5% significance level. Then a type I error is made 1 in 20 times.*

This is a convention and in many applications other values of α may be warranted. For example, in testing the null hypothesis that there will be no nuclear war the significance level may be chosen to be much higher than 0.05, since the consequence of incorrectly failing to reject the null hypothesis is so high. Reporting p-values allows the reader to easily test using their own preferred value of α.

4.4.7 Relationship to Confidence Interval

Two-sided tests can be implemented using confidence intervals. If the null hypothesis value μ^* falls inside the $100(1-\alpha)$ percent confidence interval then do not reject H_0 at significance level α. Otherwise reject H_0 at significance level α.

For the female earnings data, from Table 4.2 the 95 percent confidence interval for population mean female earnings is $(37559, 45266)$. Since this interval includes 40000 we do not reject H_0 : $\mu = 40000$ at significance level 0.05.

4.4.8 Summary

A summary of the preceding earnings hypothesis test example is the following.

Hypotheses	$H_0 : \mu = 40000$, $H_a : \mu \neq 40000$
Significance level	$\alpha = 0.05$
Data	$\bar{x} = 41413$, $s = 25527$, $n = 171$
Test statistic	$t = (41413 - 40000)/(25527/\sqrt{171}) = 0.724$
(1) p-value approach	$p = \Pr[\lvert T_{170}\rvert \geq \lvert 0.724\rvert] = 0.470$
(2) Critical value approach	$c = t_{170,.025} = 1.974$
Conclusion	Do not reject H_0 at level .05 as (1) $p > .05$ or (2) $\lvert t\rvert < c$.

The p-value and critical value approaches are alternative methods that for test at the same significance level always lead to the same conclusion.

4.5 Two-sided Hypothesis Test Examples

We consider three examples of two-sided hypothesis tests. Additionally we discuss complications that can arise – survey data should be from a representative sample and, for time series data, the standard error of the mean, $se(\bar{x})$, may no longer equal $s/\sqrt{n}$. The tests are computed manually here; more simply one can use a command such as the Stata `ttest` command.

4.5.1 Example: Gasoline Prices

Test the claim that the mean price of regular gasoline in Yolo County is neither higher nor lower than the norm for California.

The dataset GASPRICE comes from a website that provides daily data on gas prices. Data are available for 32 Yolo County gas stations on a day when the average price for all California gas stations was \$3.81. Descriptive statistics are given in Table 4.5. The standard error of the sample mean $se(\bar{x}) = s/\sqrt{n} = 0.1510/\sqrt{32} = .0267$.

Table 4.5: Summary Statistics: Gasoline price per gallon at 32 gas stations.

Variable	Obs	Mean	St. Dev.	Min	Max
Earnings	32	3.6697	0.1510	3.49	4.09

The null hypothesis is $\mu = 3.81$, tested against the alternative $\mu \neq 3.81$. The t statistic is

$$t = \frac{3.6697 - 3.81}{.0267} = -5.256.$$

Large values of t in absolute value favor the alternative, as then $\bar{x}$ is very different from 3.81. Using the p-value method we have

$$p = \Pr[|T_{31}| > |-5.256|] = 0.000.$$

We reject H_0 at level .05 since $p < .05$. Using the critical value method

$$c = t_{31,.025} = 2.040.$$

We reject H_0 at level .05 since $|t| = 5.256 > c = 2.040$. Therefore the claim that population mean Yolo County gas prices equal the California state-average price is rejected at significance level 0.05.

4.5.2 Example: Male Earnings

Test the claim that population mean earnings of male full-time workers in 2010 are $50,000.

The dataset EARNINGSMALE is a small subsample from the very large American Community Survey (ACS). The subsample is selected in such a way that it is a simple random sample of the population of 30 year-old male full-time workers in 2010. Descriptive statistics are given in Table 4.6. The minimum value of 1,000 is possible as the person was self-employed. The next lowest value was 8,000. The standard error of the sample mean $se(\bar{x}) = s/\sqrt{n} = 65034.74/\sqrt{191} = 4705.748$.

Table 4.6: Summary Statistics: Annual earnings of male full-time workers aged 30 in 2010.

Variable	Obs	Mean	St. Dev.	Min	Max
Earnings	191	52353.93	65034.74	1010	498000

The test is of $H_0 : \mu = 50000$ against $H_a : \mu \neq 50000$. The t statistic is

$$t = \frac{52353.93 - 50000}{4705.748} = .5002.$$

Large absolute values of t favor the alternative, as then $\bar{x}$ is much greater than 50000. Here

$$p = \Pr[|T_{190}| > .500] = 0.618.$$

We do not reject H_0 at level .05 since $p = .618$ is not less than .05. Alternatively, the critical value

$$c = t_{190,.025} = 1.973.$$

We do not reject H_0 at level .05 since $|t| = .500$ is not less than $c = 1.973$. The data do not support the claim that population mean earnings are more than \$50,000 at significance level .05.

Note that it is important that the sample be a representative sample. National government-sponsored surveys are usually not representative of the U.S. population, as they tend to oversample low population segments of interest to policy-makers, such as racial minorities, people with low income, and people in low population states. This is likely to lead to over-sampling of low-earnings individuals.

For **nonrepresentative samples** with sampling weights we should base inference on the sample mean $\bar{x}_w$ and its standard error $se(\bar{x}_w)$ that are defined in Chapter 3.7. Then the $100(1-\alpha)\%$ confidence interval for μ is $\bar{x}_w \pm t_{n-1,\alpha/2} \times se(\bar{x}_w)$ and the t statistic becomes $t = (\bar{x}_w - \mu^*)/se(\bar{x}_w)$.

This issue was avoided in this illustrative example, however, by using the sampling weights to select a subset of the original ACS dataset in a way that ensured that the sample considered here is a representative sample of 30 year-old males. Similarly the female earnings data analyzed in Chapters 2-4 were selected in such a way as to be a representative sample.

4.5.3 Example: Growth in U.S. real GDP per capita

Test the claim that the annual growth rate in U.S. real GDP per capita averaged 2.0% over the period 1959 to 2020. Do the test at significance level $\alpha = .05$.

We use dataset REALGDPPC introduced in Chapter 2.6. Here we use the year-to-year percentage changes in real per capita GDP, calculated as $100 \times (y_t - y_{t-4})/y_{t-4}$ where y_{t-4} denotes the variable four periods earlier. Descriptive statistics are given in Table 4.7. Assuming observations are independent, the standard error $se(\bar{x}) = 2.1781/\sqrt{241} = 0.1403$.

Table 4.7: Summary Statistics: Annual growth rate in U.S. real GDP per capita using quarterly data from 1959 to 2020.

Variable	Obs	Mean	St. Dev.	Min	Max
Growth	241	1.9904	2.1781	-4.77	7.63

The null hypothesis is $\mu = 2.0$, tested against the alternative $\mu \neq 2.0$. The t statistic is

$$t = \frac{1.9904 - 2.0}{.1403} = -0.068.$$

Large absolute values of t favor the alternative hypothesis, as then $\bar{x}$ is very different from 2.0. Using the p-value method we have

$$p = \Pr[|T_{240}| > |-0.068|] = 0.946.$$

We do not reject H_0 at level .05 since $p > .05$. Using the critical value method

$$c = t_{240,.025} = 1.970.$$

We do not reject H_0 at level .05 since $|t| = 0.068 < c = 1.970$. Therefore we do not reject the claim that population mean growth rate was 2.0% at significance level 0.05.

An important caveat in this example is that the underlying theory assumes that observations in the sample are statistically independent or unrelated with each other. In fact for time series data there can be dependence as, for example, high growth in one quarter is likely to recur again the next quarter. Failure to control for this dependence can lead to an overestimate of the precision of estimation, i.e. the reported standard error is too small.

Inference for time series regression, introduced in Chapter 12.1, provide statistical methods that are valid even with such dependence. In this particular example there is very high dependence from one quarter to the next, and appropriate methods lead to much larger standard error of $\bar{x}$. From Chapter 12.1, allowing for this complication yields $se(\bar{x}) = 0.275$ which is about twice as large. Then $t = (1.9904 - 2.0)/.275 = -0.35$. With this adjustment H_0 is still not rejected at level .05.

4.6 One-Sided Directional Hypothesis Tests

A two-sided test is called two-sided as both $\mu < \mu^*$ or $\mu > \mu^*$ are included as alternatives to the null hypothesis. For a **one-sided hypothesis test** the alternative considered is only that $\mu < \mu^*$or only that $\mu > \mu^*$.

A two-sided hypothesis test is non-directional, as rejection may be due to concluding that either $\mu > \mu^*$ or $\mu < \mu^*$. A one-sided test, by contrast, is directional. For example, we may test against the alternative that $\mu > \mu^*$.

For a one-sided directional hypothesis test care needs to be used in specifying the null and alternative hypotheses as the conclusion can differ according to which hypothesis is set up as the null and which is the alternative. As justified below, the following rule is used.

Remark 17 *For one-sided tests the statement being tested is specified to be the alternative hypothesis. And if a new theory is put forward to supplant an old, the new theory is specified to be the alternative hypothesis.*

For example, if we wish to test the claim that the population mean earnings exceed $40,000, we should test $H_0 : \mu \leq 40000$ against $H_a : \mu > 40000$. By contrast, to test the claim that the population mean earnings are less than $40,000, we should test $H_0 : \mu \geq 40000$ against $H_a : \mu < 40000$.

Definition 20 *An **upper one-tailed alternative test** is a test of $H_0 : \mu \leq \mu^*$, where μ^* is a specified value for μ, against $H_a : \mu > \mu^*$. A **lower one-tailed alternative test** is a test of $H_0 : \mu \geq \mu^*$ against $H_a : \mu < \mu^*$.*

Some textbooks instead define the null hypothesis of a one-sided test to be $H_0 : \mu = \mu^*$. For example, an upper one-tailed test is a test of $H_0 : \mu = \mu^*$ against $H_0 : \mu > \mu^*$. This alternative notation makes no difference to the subsequent analysis.

4.6.1 P-values and Critical Regions

Inference for both types of one-sided test is based on the same calculated test statistic

$$t = \frac{\bar{x} - \mu^*}{se(\bar{x})},$$

as used for two-sided hypothesis tests. As usual this statistic is viewed as being the realization of a $T(n-1)$ distributed random variable. What differs in the one-sided case is calculation of the p-values and critical values.

For an **upper** one-tailed alternative test, large positive values of t are grounds for rejection of H_0, since then $\bar{x}$ (the estimate of μ) is much larger than μ^*. Thus the p-value is the probability of being in the upper tail of the $T(n-1)$ distribution, so $p = \Pr[T_{n-1} \geq t]$. And the critical region for a test at significance level α is $t > c$ where c is such that $\Pr[T_{n-1} > c] = \alpha$ and is denoted $c = t_{n-1,\alpha}$. H_0 is rejected at significance level α if $p < \alpha$ or, equivalently, if $t > c$.

For a **lower** one-tailed alternative test large negative values of t lead to rejection of H_0, since then $\bar{x}$ is much smaller than μ^*, and the test procedure is appropriately modified.

Definition 21 *Let t be the usual t statistic. For an upper one-tailed alternative test the p-value is $p = \Pr[T_{n-1} \geq t]$, the critical value at significance level α is $c = t_{n-1,\alpha}$, and we reject H_0 if $p < \alpha$ or, equivalently, if $t > c$. For a lower one-tailed alternative test the p-value is $p = \Pr[T_{n-1} \leq t]$, the critical value at significance level α is $c = -t_{n-1,\alpha}$, and we reject H_0 if $p < \alpha$ or, equivalently, if $t < c$.*

A one-sided test is a more focused test that, at given significance level, requires less evidence to reject the null hypothesis, provided the test statistic t is in the correct tail of the distribution.

For example, consider an upper one-tail alternative test and suppose $t = 1.8$ and $n = 31$. We reject H_0 at significance level 0.05 since $p = \Pr[T_{30} > 1.8] = 0.041 < 0.05$, whereas with a two-sided test we would not reject as $p = \Pr[|T_{30}| > |1.8|] = 2 \times 0.041 = 0.082 > 0.05$.

4.6.2 Example: Mean Annual Earnings

Suppose we wish to evaluate the claim that the population mean exceeds \$40,000. A test of this claim is implemented as a test of $H_0 : \mu \leq 40000$ against $H_a : \mu > 40000$, an example of an upper one-tailed alternative test.

The t statistic has already been calculated as $t = 0.724$. Large positive values of t support rejection of H_0 since then $\bar{x}$ is much greater than the hypothesized population mean value of \$40,000.

The p-value, the probability that a t distributed random variable exceeds the observed t value of 0.724, is

$$p = \Pr[T_{170} \geq .724] = 0.235.$$

Since p is larger than 0.05, we do not reject H_0 at significance level 0.05. The p-value is the shaded region in the left panel of Figure 4.4.

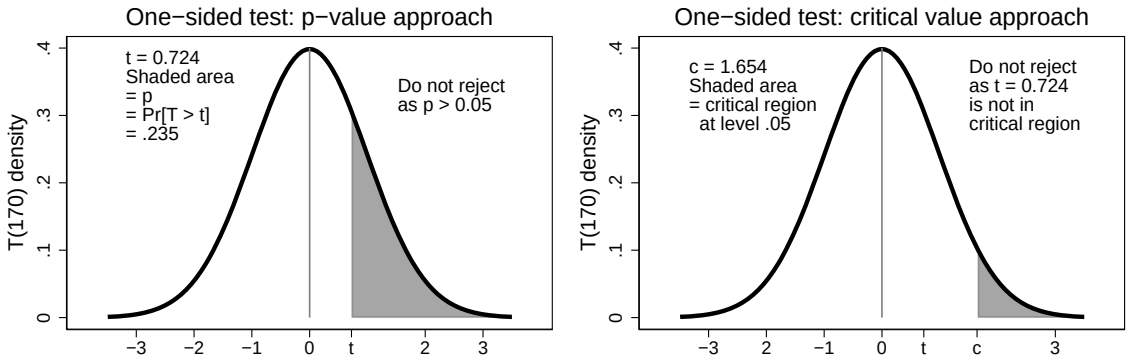

Figure 4.4: One-sided directional hypothesis test (upper one-tailed alternative): p-value approach and critical value apporach

Using the alternative equivalent critical value method instead, the critical value c solves $\Pr[T_{170} \geq c] = 0.05$ if testing is at the significance level 0.05. Then

$$c = t_{170,.05} = 1.654.$$

We do not reject H_0 at significance level 0.05, since $t = 0.724 \leq 1.654$. The critical region is the shaded area in the in the right panel of Figure 4.4.

Using either method we do not reject at significance level 0.05 the null hypothesis that the population mean earnings is less than or equal to \$40,000. There is not enough evidence to support the initial claim that mean earnings exceed \$40,000.

4.6.3 Specifying the Null Hypothesis for One-Sided Tests

This more difficult section explains why the statement being tested is specified as the alternative hypothesis.

Suppose the claim is made that population mean earnings are more than \$40,000. Should we perform an upper one-tailed alternative test or a lower one-tailed alternative test?

There are two potential ways to proceed, though only the first should be used as explained in what follows.

1. Test $H_0 : \mu \leq 40000$ against $H_a : \mu > 40000$.

2. Test $H_0 : \mu \geq 40000$ against $H_a : \mu < 40000$.

Suppose we take the first approach. The claim that $\mu > 40000$ is supported if H_0 is rejected. Rejection of H_0 requires a sample mean of considerably more than 40000, say $\bar{x} > 43000$. Suppose instead the second approach is taken. Then the claim is supported if H_0 is not rejected. Rejection

of H_0 requires a sample mean of considerably less than 40000, say $\bar{x} < 37000$ (by symmetry). Non-rejection of H_0 then occurs if $\bar{x} > 37000$. Thus the claim that $\mu > 40000$ is supported if $\bar{x} > 37000$.

To summarize, the claim is that mean earnings exceed \$40,000. The first specification of the null and alternative hypotheses leads to support of the claim if the sample average exceeds \$43,000, while the second specification leads to support of the claim if the sample average exceeds \$37,000. The philosophy of hypothesis testing is to require strong evidence to support a claim. The first specification is therefore used, with the claim made specified as the alternative hypothesis.

Remark 18 *For one-sided tests the claim being tested is specified to be the alternative hypothesis, as stronger evidence is then needed to support the claim than if the claim was set up as the null hypothesis.*

There can be considerable debate as to which hypothesis should be the null. For example, suppose we wish to determine whether women at a workplace have been discriminated against, a not unusual issue to be determined in court. One approach is to specify the alternative hypothesis to be that women are paid less than men (the claim made) while another approach is to specify this as the null hypothesis. Lawyers for the employer may favor the first approach, lawyers for the employee may favor the second approach, and the statistical methodology presented here selects the first approach.

4.7 Generalization of Confidence Intervals and Hypothesis Tests

Confidence intervals and hypothesis tests can be applied to parameters other than the population mean μ. Leading examples include the difference in two means $(\mu_2 - \mu_1)$, and the slope of a regression line, the subject of many later chapters. The approach for inference on μ extends easily to such settings.

4.7.1 Generalizations of Confidence Intervals

Inference on μ is based on $t = (\bar{x} - \mu)/se(\bar{x})$. In words, the t statistic equals the estimate minus the parameter divided by the standard error, where the standard error measures how precisely the parameter has been estimated. More generally we have the following result.

Remark 19 *For the estimators presented in this book, for sufficiently large sample size the statistic*

$$t = \frac{estimate - parameter}{standard\ error}$$

is a realization of a random variable that is approximately $T(v)$ distributed, where the degrees of freedom v for the t distribution varies with the application, and the standard error is the estimated standard deviation of the estimate.

A $100(1-\alpha)\%$ confidence interval for μ is $\bar{x} \pm t_{v,\alpha/2} \times se(\bar{x})$. This generalizes as follows.

Remark 20 *A* $100(1-\alpha)\%$ *confidence interval for the unknown parameter is*

$$\textit{estimate} \pm t_{v,\alpha/2} \times \textit{standard error.}$$

The most commonly-used confidence level is 95 percent and $t_{v,.025} \simeq 2$ for $v > 30$. This immediately leads to the following simple rule-of-thumb.

Remark 21 *An approximate 95% confidence interval for the unknown parameter is the two-standard error interval*

$$\textit{estimate} \pm 2 \times \textit{standard error.}$$

The term **margin of error** is used to describe the half-width of a confidence interval, or $t_{v,\alpha/2} \times se(\cdot)$. The term is most often used in the context of 95% confidence intervals, since these are the most commonly-used confidence intervals. Then since $t_{v,.025} \simeq 2$,

$$\text{Margin of error} = 2 \times \text{Standard error.}$$

As an example of the above, suppose the sample estimate of a parameter θ is 11 with standard error of the estimate equal to 3, and the sample size is large. Then an approximate 95% confidence interval for θ is $11 \pm 2 \times 3$ or $(5, 17)$. Since the standard error is 3, the margin of error is said to be $2 \times 3 = 6$, or ± 6.

4.7.2 Generalizations of Hypothesis Tests

For hypothesis testing we again use the more general form of the t statistic, assumed to be approximately $T(v)$ distributed.

Remark 22 *Consider a two-sided test at significance level* α *of the null hypothesis* (H_0) *that a parameter equals a hypothesized value against the alternative hypothesis* (H_a) *that it does not. Calculate the t statistic*

$$t = \frac{\textit{estimate} - \textit{hypothesized parameter value}}{\textit{standard error}}.$$

Under the null hypothesis t is the sample realization of a random variable that is approximately $T(v)$ *distributed. The p-value approach is to reject* H_0 *if* $p < \alpha$ *where* $p = \Pr[|T_v| > |t|]$. *The critical value approach is to reject* H_0 *if* $|t| > c$ *where* $c = t_{v,\alpha/2}$ *satisfies* $\Pr[T_v > t_{v,\alpha/2}] = \alpha$. *For given* α, *the two methods lead to the same conclusion.*

This generalizes inference for the mean, where the estimate is $\bar{x}$, the parameter is μ, the standard error is $se(\bar{x}) = s/\sqrt{n}$, and $v = n - 1$.

Continuing the earlier example with estimate of θ equal to 11 and standard error of 3, for a test of whether or not $\theta = 20$ the t statistic is $t = (11 - 20)/3 = -3$. We will reject $H_0 : \theta = 20$ at level 0.05 since (for all but very small v) the $T(v)$ critical value is approximately 2 and $|-3| = 3 > 2$.

For both confidence intervals and hypothesis tests, the standard normal distribution is sometimes used rather than the $T(v)$ distribution. The difference between $T(v)$ and standard normal disappears as $v \to \infty$.

4.8 Proportions Data

As an example of adaptation or extension of methods for the sample mean, consider analysis of **proportions data** that are data on the fraction of times that a given event occurs. Economic examples are unemployment rates and employment rates, and a common example in the media are political opinion polls on the fraction intending to vote for a given political candidate.

4.8.1 Analysis using General Results for the Sample Mean

Statistical inference on these data can be done using the methods of this chapter. The sample proportion is viewed as the sample mean $\bar{x}$ of data that take only the value 0 or 1, such as 1 if an individual in the sample intends to vote Democrat and 0 otherwise.

The computational formula for the sample variance is $s^2 = \frac{1}{n-1}\{\sum_{i=1}^n x_i^2 - n\bar{x}^2\}$; see Chapter 2.1 Since x_i takes only values 0 or 1, $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i = n\bar{x}$, so $\sum_{i=1}^n x_i^2 - n\bar{x}^2 = n\bar{x}(1-\bar{x})$ and the sample variance equals $s^2 = n\bar{x}(1-\bar{x})/(n-1)$. It follows that the standard error of the sample mean $se(\bar{x}) = s/\sqrt{n} = \sqrt{\bar{x}(1-\bar{x})/(n-1)}$.

For proportions data the mean parameter μ is denoted p (for proportion). Then applying the results of this chapter for the sample mean, a $100 \times (1-\alpha)\%$ confidence interval for p is $\bar{x} \pm t_{\alpha/2;n-1} \times \sqrt{\bar{x}(1-\bar{x})/(n-1)}$. And a hypothesis test of whether or not $p = p^*$ can be based on the statistic $t = (\bar{x} - p^*)/\sqrt{\bar{x}(1-\bar{x})/(n-1)}$ which is viewed as the realization of a $T(n-1)$ distributed random variable.

4.8.2 Analysis using Results Specific to Proportions Data

Statistical packages often have procedures specific to proportions data that yield slightly different results to the preceding analysis. So we detail these procedures.

For proportions data the underlying random variable X for each surveyed individual is viewed as taking value 1 with probability p and value 0 with probability $1-p$. This is the Bernoulli distribution with population mean $\mu = p$ and population variance $\sigma^2 = p(1-p)$; see Appendix B.1. For example, we may let $X = 1$ if the person intends to vote Democrat and $X = 0$ otherwise, so then p is the unknown population probability of voting for a Democrat candidate. The Bernoulli result that $\sigma^2 = p(1-p)$ implies that $\text{Var}[\bar{X}] = \sigma^2/n = p(1-p)/n$ which can be estimated by replacing p by an estimate.

For confidence intervals we estimate p by $\bar{x}$ leading to $se(\bar{x}) = \sqrt{\bar{x}(1-\bar{x})/n}$, rather than $\sqrt{\bar{x}(1-\bar{x})/(n-1)}$ given earlier. Furthermore it is most common to use critical values from the normal distribution. Then a $100 \times (1-\alpha)\%$ confidence interval for p is most often $\bar{x} \pm z_{\alpha/2} \times \sqrt{\bar{x}(1-\bar{x})/n}$.

For large n, often the case for proportions data such as political polling data, this confidence interval will be quite similar to that obtained using the earlier general result for the sample mean. First, the $z_{\alpha/2}$ and $t_{\alpha/2;n-1}$ critical values are very close for large n. Second, for large n the different estimates of $se(\bar{x})$ are very similar as division by n is then very similar to division by $n-1$.

For testing hypotheses on p it is most common to estimate $\text{Var}[\bar{X}] = p(1-p)/n$ by replacing p by the hypothesized value p^* (rather than by $\bar{x}$). Then $se(\bar{x}) = \sqrt{p^*(1-p^*)/n}$ and a two-sided

hypothesis test of $H_0 : p = p^*$ against $H_a : p \neq p^*$ is based on the statistic $z = \frac{\bar{x}-p^*}{\sqrt{p^*(1-p^*)/n}}$ which is $N(0,1)$ distributed under H_0.

The standard normal approximation is felt to be good if both $np > 10$ and $n(1-p) > 10$. Thus for low p or high p considerably more than 30 observations are needed. For polling data and for most economics applications with proportions data there are many observations, so the approximation will be good. (If instead n is too small then hypothesis tests can use exact statistical inference on p based on the binomial distribution; this is not presented here.)

4.8.3 Example: Voting Intentions

Suppose we have a random sample of 921 voters of whom 480 intend to vote Democrat and 441 who intend to vote Republican. For statistical inference we use the special approach for proportions data.

For these data $\bar{x} = [480 \times 1 + 441 \times 0]/921 = 0.5212$ and $se(\bar{x}) = \sqrt{\bar{x}(1-\bar{x})/n} = 0.01646$. For the standard normal, $z_{.025} = 1.960$. So a 95% confidence interval for the population mean proportion of Democrat voters is $0.5212 \pm 1.960 \times 0.01646 = (0.4889, 0.5535)$. Equivalently, a 95% confidence interval for the population percentage of voters who intend to vote Democrat is (48.9%, 55.3%).

Now suppose we wish to test the belief that the Democrat candidate will win the election. This is a one-sided test with the belief specified to be the alternative hypothesis. Thus we test $H_0 : p \leq 0.5$ against $H_a : p > 0.5$. For hypothesis testing we use $se(\bar{x}) = \sqrt{p^*(1-p^*)/n}$, where $p^* = 0.5$, so the test statistic $z = \frac{0.5212-0.5}{\sqrt{0.5\times(1-0.5)/921}} = 1.287$. Then $p = \Pr[|Z| > 1.287] = 0.198$ exceeds 0.05 so we do not reject the null hypothesis at significance level 0.05. Alternatively, the critical value $c = z_{.025} = 1.964$, and we do not reject H_0 at level .05 since $|z| = 1.287 < c$. We cannot conclude that the Democrat candidate will win the election at significance level 0.05.

4.8.4 Margin of Error of an Opinion Poll

The margin of error often reported alongside an opinion poll is two times the standard error, the approximate half-width of a 95% confidence interval. For proportions data $\bar{x}(1-\bar{x}) \leq 0.25$ since $\bar{x}(1-\bar{x})$ takes a maximum value of 0.25 when $\bar{x} = 0.5$. It follows that $se(\bar{x}) = \sqrt{\bar{x}(1-\bar{x})/n} \leq \sqrt{0.25/n} = 0.5/\sqrt{n}$, so the margin of error is at most $1/\sqrt{n}$.

Published opinion polls typically interview between 600 and 2,000 people. Then the corresponding maximum margin of error is, respectively, 4.1% and 2.2% since, for example, $100/\sqrt{600} \simeq 4.1$. For many purposes these margins of error are tolerable, but they will be too large for predicting the result of a close election.

4.9 Key Concepts

1. The key tools of statistical inference are confidence intervals and hypothesis tests.

2. For statistical inference on μ we use the t statistic $t = (\bar{x} - \mu)/se(\bar{x})$. This is the distance between $\bar{x}$ and μ, normalized by the standard error of the mean.

3. Under simple random sampling the t statistic t is the realization of a randomly variable that is $T(n-1)$ distributed, exactly if data are normally distributed and approximately for nonnormal data if n is sufficiently large.

4. In most cases for $n > 30$ it is reasonable to use the $T(n-1)$ distribution.

5. The $T(v)$ distribution, the t distribution with v degrees of freedom, is like a squashed version of the standard normal distribution with fatter tails. As $v \to \infty$ the t distribution goes to the standard normal.

6. T_v denotes a random variable that is $T(v)$ distributed.

7. The critical value $t_{v,\alpha}$ is that value such that a $T(v)$ distributed random variable exceeds $t_{v,\alpha}$ with probability α, i.e., $\Pr[T_v > t_{v,\alpha}] = \alpha$.

8. The critical value $t_{v,\alpha/2}$ is that value such that a $T(v)$ distributed random variable exceeds in absolute value $t_{v,\alpha/2}$ with probability α, i.e., $\Pr[|T_v| > t_{v,\alpha/2}] = \alpha$.

9. A $100(1-\alpha)\%$ confidence interval for μ is $\bar{x} \pm t_{n-1,\alpha/2} \times se(\bar{x})$. This interval will include μ with probability α.

10. A calculated 95 percent confidence interval for the population mean is an interval that if constructed for each of an infinite number of samples will include the true population mean μ 95% of the time (and will not include μ 5% of the time).

11. It is most common to use a 95% confidence interval, so $\alpha = 0.05$ and $\alpha/2 = 0.025$. A higher degree of confidence leads to a wider confidence interval.

12. An approximate 95% confidence interval for μ is the two standard error interval $\bar{x} \pm 2 \times se(\bar{x})$ where $se(\bar{x}) = s/\sqrt{n}$.

13. A two-sided hypothesis test is a test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$.

14. A type I error occurs if H_0 is rejected when H_0 is true.

15. The significance level or size of a test, denoted α, is the pre-specified maximum probability of a type I error that will be tolerated.

16. The t test statistic $t = (\bar{x} - \mu^*)/se(\bar{x})$ is the realization of a random variable that is approximately $T(n-1)$ distributed under $H_0 : \mu = \mu^*$.

17. The p-value is the probability of observing a t statistic at least as large in absolute value as that obtained in the current sample.

18. For a two-sided test $p = \Pr[|T_{n-1}| \geq |t|]$. H_0 is rejected at significance level α if $p < \alpha$,and is not rejected otherwise.

19. For a two-sided test the critical value $c = t_{n-1,\alpha/2}$. H_0 is rejected at significance level α if $|t| > c$, and is not rejected otherwise.

20. It is most common to test at significance level $\alpha = 0.05$.

21. For one-sided tests the statement being tested is specified to be the alternative hypothesis.

22. An upper one-tailed alternative test is a test of $H_0 : \mu \leq \mu^*$ against $H_a : \mu > \mu^*$. We reject H_0 if $p = \Pr[T_{n-1} \geq t] < \alpha$ or, equivalently, if $t > c = t_{n-1,\alpha}$.

23. A lower one-tailed alternative test is a test of $H_0 : \mu \geq \mu^*$ against $H_a : \mu < \mu^*$. We reject H_0 if $p = \Pr[T_{n-1} \leq t] < \alpha$ or, equivalently, if $t < c = -t_{n-1,\alpha}$.

24. For one-sided tests at significance level 0.05 a rough guide is to use as critical value 1.645 for an upper one-tail alternative and -1.645 for a lower one-tail alternative, as $t_{n-1,.025} = 1.645$ for large n.

25. In many settings the statistic $t = (\text{estimate}-\text{parameter})/(\text{standard error})$ can be viewed as the realization of a $T(v)$ distributed random variable where the degrees of freedom v varies with the application.

26. An approximate 95% confidence interval for a parameter is then the estimate plus or minus two times the standard error.

27. The half-width of a confidence interval is called the margin of error.

28. The margin of error for a 95% confidence interval is approximately two times the standard error.

29. Proportions data can be analyzed using the methods of this chapter, but are usually analyzed using a minor adaptation of these methods.

30. Key terms: t statistic; t distribution; degrees of freedom; confidence interval; two standard error interval; margin of error; hypothesis test; null hypothesis; alternative hypothesis; two-sided test; type I error; significance level; p-value; critical region; rejection region; critical value; one-sided test; upper one-sided alternative; lower one-sided alternative; margin of error.

4.10 Exercises

1. For a random variable T that is $T(22)$ distributed use a statistical package or a table of the $T(22)$ distribution to find

   (a) $\Pr[T > 2.0]$. (b) $\Pr[T < -2.0 \text{ or } T > 2.0]$. (c) t^* such that $\Pr[T > t^*] = .05$.
   (d) t^* such that $\Pr[T < -t^* \text{ or } T > t^*] = .05$.

2. Repeat the previous exercise for the $T(33)$ distribution.

3. For a standard normal distributed random variable Z, give the following (approximately) without using a computer or referring to a table. Hint: A normally distributed random variable lies within two standard deviations of its mean with approximate probability of 0.95.

   (a) $\Pr[Z > 2.0]$. (b) $\Pr[Z < -2.0 \text{ or } Z > 2.0]$. (c) z^* such that $\Pr[Z > z^*] = 0.025$.
   (d) z^* such that $\Pr[Z < -z^* \text{ or } Z > z^*] = 0.05$.

4. For a standard normal distributed random variable Z, give the following without using a computer or referring to a table. Hint: A normally distributed random variable lies within 1.645 standard deviations of its mean with approximate probability of 0.90.

   (a) $\Pr[Z > 1.645]$. (b) $\Pr[Z < -1.645 \text{ or } Z > 1.645]$. (c) z^* such that $\Pr[Z > z^*] = 0.05$.
   (d) z^* such that $\Pr[Z < -z^* \text{ or } Z > z^*] = 0.10$.

5. The dataset TDIST4 has the sample means $\bar{x}$ and corresponding standard standard deviations s from 1000 simple random samples of size 4 where $X \sim N(100, 16^2)$.

   (a) Compute $z = (\bar{x} - 100)/8$ for each of the 1,000 samples.
   (b) What mean, standard deviation and distribution do you expect for z? Explain.
   (c) Obtain detailed summary statistics for z. Compare these to your answers in (b).
   (d) Compute $t = (\bar{x} - 100)/se(\bar{x})$ where $se(\bar{x}) = s/\sqrt{N}$. Explain why t is t_3 distributed.
   (e) The t_3 distribution has mean 0 and variance 3 and for example, $\Pr[T_3 > 1.638] = 0.10$. Obtain detailed summary statistics for t and compare to these expected values.
   (f) On the same graph plot kernel density estimates for both z and t and comment on any differences.

6. The dataset TDIST25 has the sample means $\bar{x}$ and corresponding standard standard deviations s from 1000 simple random samples of size 25 where $X \sim N(200, 50^2)$.

   (a) Compute $z = (\bar{x} - 200)/10$ for each of the 1,000 samples.

   (b) What mean, standard deviation and distribution do you expect for z? Explain.

   (c) Obtain detailed summary statistics for z? Compare these to your answers in (b).

   (d) Compute $t = (\bar{x} - 200)/se(\bar{x})$ where $se(\bar{x}) = s/\sqrt{N}$. Explain why t is t_{24} distributed.

   (e) The t_{24} distribution has mean 0 and variance 1.091 and for example, $\Pr[T_{24} > 1.318] = 0.10$. Obtain detailed summary statistics for t and compare to these expected values.

   (f) On the same graph plot kernel density estimates for z and for t and comment on any differences.

7. For a random variable T that is $T(30)$ distributed, $\Pr[-1.3 < T < 1.3] = 0.80$. Using this result, obtain an 80% confidence interval for the population mean μ given a random sample with $n = 31$, $\bar{x} = 40$ and $s/\sqrt{n} = 10$.

8. Consider a random variable T that is $T(60)$ distributed.

   (a) Find $\Pr[|T| > 1.2]$. Show your answer on a hand-drawn graph similar to Figure 4.2.

   (b) Find $t_{60,.025}$. Show your answer on a hand-drawn graph similar to Figure 4.2.

9. Suppose we obtain a random sample with $\bar{x} = 10$, $s = 20$ and $N = 25$. Obtain a 95% confidence interval for μ using t critical values.

10. Repeat the previous exercise with $\bar{x} = 80$, $s = 60$ and $N = 100$.

11. The dataset HOUSE has data on the price and size of houses sold in a small homogeneous community.

    (a) Read the data into your statistical package.

    (b) Using a statistical package command obtain a 95% confidence interval for mean price.

    (c) Now manually calculate the same confidence interval using the sample mean, standard deviation, sample size and an appropriate t critical value.

12. Repeat the previous exercise for house size.

13. Suppose a sample yields a 95% confidence interval for μ of (20, 30). Do you expect a wider, narrower, or similar confidence interval in the following situations?

    (a) A 99% confidence interval is constructed; (b) the sample size is much larger;

    (c) the sample mean is much larger; (d) the standard deviation is much larger.

14. Suppose a sample yields a 95% confidence interval for μ of (20, 30). Which of the following is likely to lead to a narrower confidence interval?

    (a) A 90% confidence interval; (b) a smaller sample;

    (c) a sample that had the same mean but a larger standard deviation.

15. The dataset HOUSE has data on the price and size of houses sold in a small homogeneous community.

    (a) Read the data into your statistical package.

    (b) Using a statistical package command perform a two-sided test of whether or not for house price $\mu = 270000$ at significance level 0.05. State the null and alternative hypotheses and your conclusion.

    (c) Now manually perform the same test using the sample mean, standard deviation, sample size and an appropriate t critical value.

    (d) Repeat part (c) except compute the p-value. State your conclusion.

16. Repeat the previous exercise for house size and test of whether or not $\mu = 2000$.

17. Suppose a random sample of 25 economists forecast economic growth for the next year. The range of forecasts is from growth of -2.0 percent to growth of 3.5 percent with average 1.2 percent and with standard deviation of 2.0 percent.

    (a) Obtain the standard error of the sample mean forecast.

    (b) Give a 95 percent confidence interval for the population mean forecast.

    (c) Test at significance level 0.05 the claim that growth will be zero next year. State the null and alternative hypotheses and your conclusion.

    (d) Test at significance level 0.05 the claim that the next year will be a growth year, i.e. that growth will be positive. State the null and alternative hypotheses and your conclusion.

    (e) What distributional assumptions on the underlying forecasts are needed to justify the methods used in parts b to d?

    (f) You are told that the economists sampled are top advisors to the main opposition political party in the country. How, if at all, would the analysis in this exercise be affected?

18. The IQ score for a simple random sample of 88 people has sample mean 102 and sample standard deviation 14.

    (a) Give a 95% confidence interval for the population mean IQ. [Hint: Be sure to use the standard error of the sample mean and not the standard deviation].

    (b) Perform a test at significance level 0.05 of the null hypothesis that population mean IQ equals 100 against the alternative that it does not equal 100. Use the p-value approach.

    (c) Repeat part (b) using the critical value approach.

    (d) The claim is made that population mean IQ exceeds 100. Perform an appropriate hypothesis test at significance level 0.05 State clearly your conclusion.

19. Suppose we fail to reject H_0 at significance level 0.05. Do you expect it to be more likely that we reject H_0 in the following situations?

    (a) The test is at significance level 0.01.

    (b) The sample size is much larger.

20. For a random sample with $X_i \sim (\mu, \sigma^2)$ answer true or false to the following statements.

    (a) If we reject a hypothesis on μ at level 0.05 then we will necessarily also reject at level 0.01.

    (b) If a 95% confidence interval for μ includes zero then we will necessarily reject $H_0 : \mu = 0$ against $H_a : \mu \neq 0$ at level 0.05.

    (c) If $\bar{x} = 2$ leads to $p = 0.06$ in test of $H_0 : \mu = 0$ against $H_a : \mu \neq 0$ then we will reject $H_0 : \mu \leq 0$ against $H_a : \mu > 0$ at level 0.05.

21. The dataset TDIST4 has the sample means $\bar{x}$ and corresponding standard standard deviations s from 1000 random samples of size 4 where $X \sim N(100, 16^2)$.

    (a) For each of the 1,000 samples compute a 95% confidence interval for μ.

    (b) How many of these confidence intervals include 100. Is this what you expect? Explain.

    (c) For each of the 1,000 samples compute the t-statistic for test of $H_0 : \mu = 100$ against $H_a : \mu \neq 100$.

    (d) Count how many of these tests reject H_0 at level 0.05. Is this what you expect? Explain.

22. The dataset TDIST25 has the sample means $\bar{x}$ and corresponding standard standard deviations s from 1000 random samples of size 25 where $X \sim N(200, 50^2)$.

    (a) For each of the 1,000 samples compute separately the lower bound and upper bound of a 95% confidence interval for μ.

    (b) How many of these confidence intervals include 200. Is this what you expect? Explain.

    (c) For each of the 1,000 samples compute the t-statistic for test of $H_0 : \mu = 200$ against $H_a : \mu \neq 200$.

    (d) Count how many of these tests reject H_0 at level 0.05. Is this what you expect? Explain.

23. Repeat exercise 21 except generate the 1,000 sample means and standard deviations yourself. For example, use the Stata code in Chapter 3.7 except replace `set obs 30` with `set obs 4` and replace commands `generate u=runiform()` and `generate x=u>0.5` with `generate x=rnormal(100,16)`.

24. Repeat exercise 22 except generate the 1,000 sample means and standard deviations yourself. For example, use the Stata code in Chapter 3.7 except replace `set obs 30` with `set obs 25` and replace commands `generate u=runiform()` and `generate x=u>0.5` with `generate x=rnormal(200,50)`.

25. The summary statistics for usual hours worked per week (variable *hours*) for a simple random sample of women aged 30 years are the following

Variable	Obs	Mean	Std. Dev.	Min	Max
hours	109	32.81	19.73	0	90

    (a) Obtain the standard error of the sample mean.

    (b) Give a 95% confidence interval for population mean usual hours worked.

    (c) The claim is made that the population mean usual hours worked is 35 hours. Test this claim at significance level 0.01. State the null and alternative hypotheses and your conclusion.

    (d) The claim is made that the population mean usual hours worked is less than 35 hours. Test this claim at significance level 0.10. State the null and alternative hypotheses and your conclusion.

26. The summary statistics for educational level for a simple random sample of women aged 30 years who are full-time workers are the following

Variable	Obs	Mean	Std. Dev.	Min	Max
education	171	14.43	2.74	3	20

(a) Obtain the standard error of the sample mean.

(b) Give a 95% confidence interval for population mean usual hours worked.

(c) The claim is made that the population mean education is 14 years. Test this claim at significance level 0.05. State the null and alternative hypotheses and your conclusion.

(d) The claim is made that the population mean education is more than 14 year. Test this claim at significance level 0.05. State the null and alternative hypotheses and your conclusion.

27. A statistical package gives the following output. Use this output to answer the following questions in the easiest way.

```
One-sample t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	85	62.46812	10.07956	92.92892	42.42382	82.51242

```
    mean = mean(x)                                     t =   1.7330
Ho: mean = 45                     degrees of freedom =       84

    Ha: mean < 45                 Ha: mean != 45                 Ha: mean > 45
 Pr(T < t) = 0.9566         Pr(|T| > |t|) = 0.0868          Pr(T > t) = 0.0434
```

(a) Perform a test of whether or not $\mu = 0$ at significance level 0.05.

(b) The claim is made that the population mean is positive. State the null and alternative hypotheses and perform a test at significance level 0.05

(c) Calculate a 90% confidence interval for μ.

(d) The claim is made that the population mean exceeds 35. State the null and alternative hypotheses and perform an appropriate test at significance level 0.05.

28. The dataset NAEP has scores for 51 U.S. states (including the District of Columbia) on the National Assessment of Educational Progress (NAEP) for eighth-grade mathematics for the years 2003, 2005, 2007 and 2009. Consider the change in the score for each state from 2003 to 2005 (which you need to calculate).

(a) Generate data for the change in the score from 2003 to 2005.

(b) Calculate a 95% confidence interval for the mean score change.

(c) Test at significance level 0.95 the claim that there has been no change in the mean score. State the null and alternative hypotheses and your conclusion.

(d) Test at significance level 0.05 the claim that the mean score improved. State the null and alternative hypotheses and your conclusion.

29. Hospitals in the U.S. post very high charges, much higher than their costs. (Hospitals then discount off their posted prices with discount varying according to the power of the purchaser, such as an individual's health insurance company.) The dataset KNEEREPLACE has 2011 data for a number of New York hospitals on the average posted charge (*meancharge*) and average cost (*meancost*)of knee joint replacement for cases of moderate severity, where the average is over all such cases the hospital treated in 2011.

    (a) Calculate the ratio of average posted charge to mean cost for each hospital.
    (b) Calculate a 95% confidence interval for the mean of this ratio. Comment.
    (c) Test at significance level 0.05 the claim that the mean ratio is equal to 2.5. State the null and alternative hypotheses and your conclusion.
    (d) Test at significance level 0.05 the claim that the mean ratio is less than 2.5. State the null and alternative hypotheses and your conclusion.

30. The dataset SPOTFORWARD has data for the spot price (*niso*) and one-day ahead forward price (*npx*) in the California wholesale electricity market for the one hour period 5-6 p.m. for each day from April 1 1998 to February 12 2000. Restrict analysis to days in 1998.

    (a) Generate daily data (for 1998) on the difference between the spot price and the one-day ahead price.
    (b) Create a variable *dayofyear* and give a line plot of the difference against *dayofyear*. Comment.
    (c) Calculate a 95% confidence interval for the mean difference.
    (d) Test at significance level 0.95 the claim that there is no difference between the spot and one-day ahead price. State the null and alternative hypotheses and your conclusion.
    (e) Test at significance level 0.05 the claim that the one-day ahead price exceeds the spot price. State the null and alternative hypotheses and your conclusion.
    (f) If markets work fully efficiently then the day-ahead price should equal the spot price. Does this appear to be the case here?

31. Suppose the sample estimate of a parameter θ is 25 with standard error 4 and the sample size is large.

    (a) Obtain an approximate 95% confidence interval for θ.
    (b) Give the margin of error of the estimate.
    (c) Perform a test whether or not $\theta = 15$ at significance level 0.05.

32. Repeat the previous exercise if the estimate is 10 with standard error 3.

33. Suppose a random sample of 1,025 potential voters of whom 550 plan to vote for candidate A and the remainder vote for candidate B. Use the methods of the section 4.8.3 example in answering the following.

    (a) Provide an estimate of the proportion voting for candidate A and the standard error of this estimate.

    (b) Provide a 95% confidence interval for the proportion voting for candidate A.

    (c) Give the margin of error for this opinion poll.

    (d) Perform a test at level 0.05 of whether candidate A will win the election.

34. A government survey of 5,283 people finds that the unemployment rate is 6.7%. Use the methods of the section 4.8.3 example in answering the following.

    (a) Compute the margin of error for the estimate of the unemployment rate.

    (b) Provide a 95% confidence interval for the unemployment rate.

    (c) Test at significance level 0.05 whether the population mean unemployment rate is 7.0%.

Appendix B

Some Essentials of Probability Theory

Statistical inference extrapolates from sample estimates, notably the sample mean and regression coefficients, to their population analogs – the mean and the conditional mean. This extrapolation controls for the randomness of the sample using results from probability theory.

Appendix B.1 presents basic probability theory, Appendix B.2 presents results on the distribution of the sample mean, and Appendix B.3 presents material on conditional distributions that is the probability basis for the linear regression model.

B.1 Probability Theory for a Single Random Variable

A **random variable** is a variable whose value is determined by the outcome of an experiment, where an **experiment** is an operation whose outcome cannot be predicted with certainty. Standard notation is to denote the random variable in upper case, say X (or Y or Z), and to denote the values that the random variable can take in lower case, say x (or y or z).

The **probability distribution** of a random variable X describes the random behavior of X.

B.1.1 Discrete Random Variables

A **discrete random variable** is a random variable that can only take a finite number of values (or a countably infinite number of variables such as 0, 1, 2, 3, ...). As an example, X may measure whether or not a person is currently employed, so X may take values 1 (employed) or 0 (not employed). As a second example, X may be the number of consultations with a doctor over the past year; then X may take the values 0, 1, 2, ...

In general a discrete random variable takes values x_1, x_2, ... The **probability mass function** cumulative probability distribution function gives the probabilities for each value taken by the random variable:

$$\Pr[X = x], \quad x = x_1, x_2, ...$$

Probabilities lie between 0 and 1 and sum to one over all possible values of X, so

$$\sum\nolimits_x \Pr[X = x] = 1,$$

where $\sum_x$ denotes summation over all possible values taken by X.

The **cumulative distribution function** gives the probability that the random variable X is less than or equal to a particular value:

$$\Pr[X \leq x] \quad x = x_1, x_2, ...$$

The probability that X lies in a given range can be calculated using either the probability mass function or the cumulative distribution function. We have

$$\begin{aligned} \Pr[a \leq X \leq b] &= \Pr[X = a] + \cdots + \Pr[X = b] \\ &= \Pr[X \leq b] - \Pr[X < a]. \end{aligned}$$

The **expected value** of a function $g(X)$ of the random variable X is the long-run average value that we expect if we draw a value x_1 of X at random and compute $g(x_1)$, draw a second value and so on, and then obtain the average of these values. Equivalently, for each value that x might take, compute $g(x)$ and then calculate the probability-weighted average of these values by weighting this value by the probability of that value x occurring. Then the **expected value** of $g(X)$

$$\begin{aligned} \mathrm{E}[g(X)] &= g(x_1) \times \Pr[X = x_1] + g(x_2) \times \Pr[X = x_2] + \cdots \\ &= \sum\nolimits_x g(x) \times \Pr[X = x]. \end{aligned}$$

The two most commonly-used expected values are the **mean** $\mu \equiv \mathrm{E}[X]$ that sets $g(X) = X$ and the **variance** $\sigma^2 \equiv \mathrm{E}[(X - \mu)^2]$ that sets $g(X) = (X - \mu)^2$. Additionally the **standard deviation** σ is the square root of the variance. The expected value of a constant is the constant: $\mathrm{E}[a] = a$.

The discrete probability distributions analyzed in introductory probability courses are the **Bernoulli**, **binomial** and **Poisson** distributions. Basic analysis of economics data uses only the first of these, which is presented next.

B.1.2 Bernoulli Distribution

The **Bernoulli distribution** is the term used to describe the distribution of a random variable that takes just one of two values: 0 or 1. This is the simplest example of a discrete random variable.

Denote the probability that $X = 1$ by p. For example, $p = 0.5$ in the case of a coin toss if the the coin is fair. Then it must be the case that $X = 0$ with probability $1 - p$, since the probabilities over all possible outcomes must sum to one, and $p + (1 - p) = 1$. So

$$\Pr[X = x] = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0. \end{cases}$$

The quantity p that determines the probability distribution is called a **parameter**. It is unknown, but can be estimated given a sample.

A Bernoulli random variable has mean $\mu = p$ and variance $\sigma^2 = p(1 - p)$.

These properties can be obtained using the following algebra. For the mean, $\mathrm{E}[X] = 0 \times \Pr[X = 0] + 1 \times \Pr[X = 1] = 0 \times (1 - p) + 1 \times p = p$. And the variance $\sigma^2 = \mathrm{E}[(X - \mu)^2] = (0 - p)^2 \times \Pr[X = 0] + (1 - p)^2 \times \Pr[X = 1] = p^2 \times (1 - p) + (1 - p)^2 \times p = p(1 - p)\{p + (1 - p)\} = p(1 - p)$.

B.1.3 Linear Transformation of a Random Variable

If we **add a fixed amount** a **to a random variable** then the mean is changed by the amount a and the variance is unchanged. If we **multiply a random variable by a fixed multiple** then the mean is multiplied by b and the variance is multiplied by b^2 (hence the standard deviation is multiplied by b). Combining, if X has mean μ and variance σ^2 then the random variable $Y = a+bX$, a **linear transformation** of X, has mean $\mathrm{E}[Y] = a + b\mathrm{E}[X]$ and variance $\mathrm{Var}[Y] = b^2\mathrm{Var}[X]$.

It follows that if X has mean μ and variance σ^2 then $X - \mu$ has mean $\mathrm{E}[X] - \mu = \mu - \mu = 0$ and variance σ^2. Subsequent division by σ leads to mean $0/\sigma = 0$ and variance $\sigma^2/\sigma^2 = 1$. Thus the random variable $Y = (X - \mu)/\sigma$ is a **standardized random variable** with mean zero and variance 1. So if we subtract the mean and divide by the standard deviation we transform the random variable X to the new random variable Y that necessarily has mean 0 and variance 1.

B.1.4 Continuous Random Variables

Not all random variables take just discrete values.

A **continuous random variable** X can take an uncountably infinite number of values, such as any real value, or any positive real value, or any real value between zero and one. As an example, X may be annual income of an individual. Or X may be the length of time that the individual has been employed at their current job.

Since X can take any value, the probability that it equals any particular value is infinitesimally small. So it is meaningless to consider the probability of X taking a particular value. Instead we evaluate the probability that X lies in a range of values.

A **continuous probability distribution** is defined by the **probability density function** $f(x)$. This function has the property that the probability that X lies between two values, say a and b, is given by the area under the function $f(x)$ between a and b. Since $\Pr[X = a] = 0$ for a continuous random variable, the following expressions yield equivalent probabilities

$$\Pr[a < X < b] = \Pr[a \leq X \leq b] = \Pr[a \leq X < b] = \Pr[a < X \leq b].$$

The total area under the probability density function curve, from the minimum to maximum value of X, equals one since probabilities sum to one.

The associated **cumulative probability distribution function**, denoted $F(x)$, is defined as $\Pr[X \leq x]$ and is given by the area under the curve from the lowest value that X can take to x.

B.1.5 Standard Normal Distribution

The leading example of a continuous random variable is a standard normal random variable. The **standard normal distribution** is defined by its probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right), \quad -\infty < x < \infty.$$

The standard normal can be shown to have mean $\mu = 0$ and standard deviation $\sigma = 1$. The notation $N(0, 1)$ is used to denote the standard normal distribution. The standard normal density is the curve given in Figure B.1.

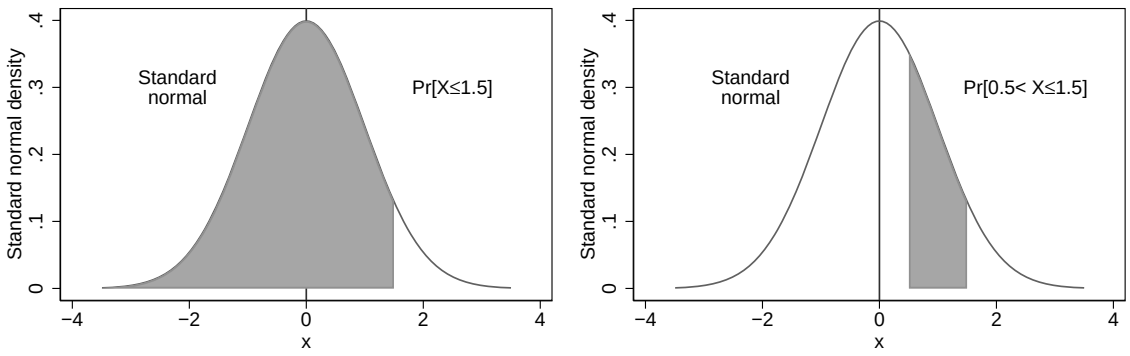

Figure B.1: Standard normal density: Graphs of $\Pr[X \leq 1.5]$ and $\Pr[0.5 \leq X \leq 1.5]$.

The shaded region in the left panel of Figure B.1 gives $\Pr[X \leq 1.5]$, since this is the area under the curve from $-\infty$ to 1.5. The total area under the curve is necessarily 1, and it appears visually that $\Pr[X \leq 1.5] \simeq 0.9$ since the shaded region is about 90% of the total area under the curve. From standard normal tables or a computer in fact $\Pr[X \leq 1.5] = 0.9332$.

The right panel of Figure B.1 gives $\Pr[0.5 < X \leq 1.5]$. This appears to be approximately equal to 0.2. From standard normal tables or a computer in fact $\Pr[0.5 < X \leq 1.5] = 0.2317$.

For those familiar with **integral calculus** the area under the curve is obtained by taking the integral. Thus, for example, $\Pr[X \leq 1.5] = \int_{-\infty}^{1.5} f(x)dx$. For the standard normal, where $f(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$, this integral has no exact solution. Instead one uses numerical approximations that are given in statistical tables or are calculated by the computer.

B.1.6 Other Continuous Distributions

The normal distribution is a generalization of the standard normal distribution that allows for a nonzero mean and a standard deviation other than one. In the more general case the **normal distribution,** with mean μ and standard deviation σ, has probability density function $f(x) = (1/\sqrt{2\pi\sigma^2})\exp\left(-(x-\mu)^2/2\sigma^2\right)$ and is denoted $N(\mu, \sigma^2)$.

A powerful property of the normal, shared by few other distributions, is that linear combinations of normally distributed random variables are also normally distributed. If $X \sim N(\mu, \sigma^2)$ then $Y = a + bX \sim N(a + b\mu, b^2\sigma^2)$. One consequence of this result is that if $X \sim (\mu, \sigma^2)$ then $Y = (X - \mu)/\sigma \sim N(0, 1)$. In words, for a normally distributed random variable, subtracting the mean and then dividing by the standard deviation leads to a random variable that is standard normal distributed.

Aside from the normal, the continuous probability distributions most often used in econometrics are the t, F and chi-squared distributions, as these are used to obtain critical values and p-values for hypothesis tests. The probability density functions for these distributions are complicated and are not presented here, and probabilities need to be obtained from tables or be calculated by the

computer. The t **distribution** is discussed in some detail in Chapter 4.2. The F **distribution** and the **chi-squared distribution** are introduced in Chapter 11.5. Appendix F gives tables.

B.2 Probability Theory for the Sample Mean

We provide in more detail results presented in Chapter 3 for the average $\bar{X}$ that is the sum of n independent random variables $X_1, ..., X_n$ divided by n.

B.2.1 Statistical Independence

The two random variables X and Y are **statistically independent** or, more simply, **independent** if the value taken by X is unrelated to the value taken by Y.

For example, let the two random variables X and Y represent the outcomes from two consecutive tosses of the coin. Then the two random variables are statistically independent if the result of the first toss, say a head, has no bearing on the result of the second toss. If the probability of heads on the first toss is p then, regardless of whether the first toss results in heads or tails, the probability of heads on the second toss is still p. (This is the case even if $p = 0.4$, say, so that the coin is not fair).

Random variables are not necessarily statistically independent. But under simple random sampling $X_1, ..., X_n$ are statistically independent.

B.2.2 Sums of Random Variables

The mean of a **weighted sum of random variables** equals the weighted sum of their means. That is,

$$\mathrm{E}[aX + bY] = a \times \mathrm{E}[X] + b \times \mathrm{E}[Y].$$

The variance of a sum of random variables is more complicated, as it depends on the statistical relationship between X and Y. Simplification occurs if the random variables are statistically independent. Then

$$\mathrm{Var}[aX + bY] = \mathrm{Var}[aX] + \ \mathrm{Var}[bY] = a^2 \times \mathrm{Var}[X] + b^2 \times \mathrm{Var}[Y].$$

This is a weighted sum of the individual variance, with weights that are the square of the original weights.

Applying this result, for statistically independent random variables the sum $X+Y$ has variance that is the sum of the variance of X and the variance of Y. Similarly, the difference $X - Y$ has variance that is the sum of the variance of X and the variance of Y.

B.2.3 Mean and Variance of the Sample Mean

The sample mean is the random variable $\bar{X}$ that equals the sum of X_1 to X_n divided by n. This can be written as a weighted sum of random variables

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n.$$

We first show that $\mathrm{E}[\bar{X}] = \mu$ given common mean $\mathrm{E}[X_i] = \mu$ (assumption A in Chapter 3.4). Then $\mathrm{E}[X_i/n] = \mu/n$ and hence

$$\mathrm{E}[\bar{X}] = \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu.$$

Next we want to show that $\mathrm{Var}[\bar{X}] = \sigma^2/n$ given assumptions A–C in Chapter 3.4. The X_i are statistically independent by assumption C, so the variance of a sum is the sum of the variances and

$$\begin{aligned}\mathrm{Var}[\bar{X}] &= \mathrm{Var}\left[\tfrac{1}{n}X_1 + \tfrac{1}{n}X_2 + ... + \tfrac{1}{n}X_n\right] \\ &= \mathrm{Var}\left[\tfrac{1}{n}X_1\right] + \mathrm{Var}\left[\tfrac{1}{n}X_2\right] + \cdots + \mathrm{Var}\left[\tfrac{1}{n}X_n\right] \\ &= \left(\tfrac{1}{n}\right)^2 \mathrm{Var}\left[X_1\right] + \cdots + \left(\tfrac{1}{n}\right)^2 \mathrm{Var}\left[X_n\right],\end{aligned}$$

where the third equality uses $\mathrm{Var}[bX] = b^2\mathrm{Var}[X]$. Given assumption B of a common variance σ^2, $\mathrm{Var}[\bar{X}] = \left(\frac{1}{n}\right)^2\sigma^2 + \cdots + \left(\frac{1}{n}\right)^2\sigma^2$, which equals n times $\left(\frac{1}{n}\right)^2\sigma^2$, which equals $\frac{1}{n}\sigma^2$.

It follows that the standard deviation of $\bar{X}$ is $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$.

B.2.4 Law of Large Numbers

Now consider the behavior of $\bar{X}$ as the sample size gets large. Then $\bar{X}$ has mean μ and variance that goes to zero, since the variance $\sigma^2/n \to 0$ as $n \to \infty$.

So the distribution of $\bar{X}$ is centered on μ with very little variation around μ. The formal statistical term used is that $\bar{X}$ **converges in probability** to μ if the probability that $|\bar{X} - \mu| > \varepsilon$ goes to zero as $n \to \infty$, no matter how small $\varepsilon > 0$ is chosen to be.

A **law of large numbers** states that, under some assumptions on the individual X_i, an average of random variables converges in probability to its expected value; here that $\bar{X}$ converges in probability μ. The simplest law of large numbers assumes that X_i are statistically independent and identically distributed and that the mean μ exists. This is the case for simple random sampling.

B.2.5 Central Limit Theorem

The standardized variable $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has mean zero and variance one.

A **central limit theorem** states that, under some assumptions on the individual X_i, Z is standard normally distributed as the sample size gets large. That is, as $n \to \infty$

$$Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Note that this result does not require that X_i is normally distributed. It follows that in large samples $\bar{X} \sim N(\mu, \sigma^2/n)$.

There are many central limit theorems that vary with the assumptions made about the individual random variables X_i. The simplest central limit theorem is the Lindberg-Levy central limit theorem that assumes that X_i are statistically independent and identically distributed with mean μ and variance σ^2. This is the case for simple random sampling.

More general central limit theorems do not require a common mean, a common variance and/or statistical independence. Then $Z = (\bar{X} - \mathrm{E}[\bar{X}])/\sqrt{\mathrm{Var}[\bar{X}]} \sim N(0, 1)$ where $\mathrm{E}[\bar{X}]$ may no longer simplify to μ and $\mathrm{Var}[\bar{X}]$ may no longer simplify to $\sigma/\sqrt{n}$.

B.3 Probability Theory for Two Related Random Variables

The bivariate regression model is a model of the conditional mean of Y given $X = x$. Here we define the conditional mean, conditional variance, covariance and correlation.

B.3.1 Joint Probabilities

Consider the statistical relationship between two random variables X and Y that are both discrete random variables. Their joint probability of occurrence is defined by the **joint probability mass function**

$$\Pr[X = x, Y = y], \quad x = x_1, x_2, ..., \quad y = y_1, y_2, ...$$

where upper case denotes the random variable and lower case denotes the values that the random variable might take.

Given knowledge of the joint probabilities of X and Y we can obtain the separate probabilities for X and for Y. For example, $\Pr[X = x]$ equals the sum over all possible values of y of the joint probability $\Pr[X = x, Y = y]$.

B.3.2 Example

Throughout this appendix we consider the following example:

$$\Pr[X = x, Y = y] = \begin{cases} 0.1 & \text{for } x = 1,\ y = 50 \\ 0.1 & \text{for } x = 1,\ y = 30 \\ 0.2 & \text{for } x = 0,\ y = 30 \\ 0.6 & \text{for } x = 0,\ y = 10. \end{cases}$$

Note that these probabilities sum to one.

For this example, $\Pr[X = 1] = \Pr[X = 1, Y = 50] + \Pr[X = 1, Y = 30] = 0.1 + 0.1 = 0.2$ and $\Pr[X = 0] = \Pr[X = 0, Y = 30] + \Pr[X = 0, Y = 10] = 0.2 + 0.6 = 0.8$. So $X = 1$ with probability 0.2 and $X = 0$ with probability 0.8. Thus

$$\Pr[X = x] = \begin{cases} 0.2 & \text{for } x = 1 \\ 0.8 & \text{for } x = 0. \end{cases}$$

By similar summation, now over the possible values of X, $\Pr[Y = 50] = 0.1$, $\Pr[Y = 30] = 0.1 + 0.2 = 0.3$, and $\Pr[Y = 10] = 0.6$. So

$$\Pr[Y = y] = \begin{cases} 0.1 & \text{for } y = 50 \\ 0.3 & \text{for } y = 30 \\ 0.6 & \text{for } y = 10. \end{cases}$$

B.3.3 Conditional Distribution

A very important result is **Bayes Theorem** that states that for any two events A and B, the probability that event A happens, given that event B happens, is the joint probability that A and B happen divided by the probability that B happens:

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]},$$

where $A \cap B$ means the intersection of events A and B.

For example, consider the probability of getting a three on the second toss of a six-sided die (event A), given that the combined sum of the first two tosses is five (event B). The only way that both A and B can occur is if the first toss was a 2 and the second was a 3. The two tosses can sum to five in four ways, with tosses $(1,4)$, $(2,3)$, $(3,2)$, and $(4,1)$. All these outcomes are equally likely, so the conditional probability of event A given B is one in four, or 0.25.

More generally, the **conditional probability** of Y given X

$$\Pr[Y = y|X = x] = \frac{\Pr[Y = y, X = x]}{\Pr[X = x]}.$$

Consider the earlier numerical example and condition on $X = 1$. Then $Y = 50$ or $Y = 30$ with $\Pr[X = 1, Y = 50] = 0.1$ and $\Pr[X = 1, Y = 30] = 0.1$. Also $\Pr[X = 1] = 0.2$. It follows that the conditional distribution of Y given $X = 1$ is $\Pr[Y = 30|X = 1] = 0.1/0.2 = 0.5$ and $\Pr[Y = 50|X = 1] = 0.1/0.2 = 0.5$. By similar reasoning, since $\Pr[X = 0] = 0.8$, the conditional distribution of Y given $X = 0$ is $\Pr[Y = 10|X = 0] = 0.6/0.8 = 0.75$ and $\Pr[Y = 30|X = 0] = 0.2/0.8 = 0.25$.

The conditional probabilities for Y given $X = 1$ are therefore

$$\Pr[Y = y|X = 1] = \begin{cases} 0.5 & \text{for } y = 50 \\ 0.5 & \text{for } x = 30, \end{cases}$$

and for Y given $X = 0$:

$$\Pr[Y = y|X = 0] = \begin{cases} 0.25 & \text{for } y = 30 \\ 0.75 & \text{for } y = 10. \end{cases}$$

If X and Y are **statistically independent** then the probability of Y taking a particular value is unaffected by the value taken by X. In that case the conditional probability $\Pr[Y = y|X = x]$ reduces to the unconditional probability $\Pr[Y = y]$.

In the example of this appendix X and Y are **not** statistically independent.

B.3.4 Conditional Mean

The **conditional expected value** of a function $g(Y)$ given $X = x$ is an extension of the usual unconditional expected value of $g(Y)$, except that the values $g(y)$ are weighted by the conditional probabilities of $Y|X = x$ rather than the unconditional probabilities of Y. Thus

$$\mathrm{E}[g(Y)|X = x] = g(y_1) \times \Pr[Y = y_1|X = x] + g(y_2) \times \Pr[Y = y_2|X = x] + \cdots$$

The **conditional mean** of Y given X is the mean of the conditional distribution. This is the weighted sum of the possible values of Y where the weighting is by the conditional probability of Y given x. So

$$\mathrm{E}[Y|X=x] = \sum\nolimits_y y \times \Pr[Y=y|X=x].$$

For the example of this appendix begin with the conditional mean of Y given $X=1$. When $X=1$, Y takes value 30 with $\Pr[Y=30|X=1]=0.5$, and Y takes value 50 with $\Pr[Y=50|X=1]=0.5$. It follows that $\mathrm{E}[Y|X=1] = 30\times 0.5 + 50 \times 0.5 = 40$. By similar calculation $\mathrm{E}[Y|X=0] = 10 \times \Pr[Y=10|X=0] + 30 \times \Pr[Y=30|X=0] = 10 \times 0.75 + 30 \times 0.25 = 15$. So

$$\mathrm{E}[Y|X=x] = \begin{cases} 40 & \text{for } x=1 \\ 15 & \text{for } x=0. \end{cases}$$

In this example the conditional mean of Y given X varies with the value of X.

For linear regression it is assumed that the **conditional mean $\mathrm{E}[Y|X=x]$ is a linear function of** x. Then the population relationship between Y and X is a line with intercept denoted β_1 and slope denoted β_2, so

$$\mathrm{E}[Y|X=x] = \beta_1 + \beta_2 x.$$

For example, suppose that $\mathrm{E}[Y|X=x] = 3 + 2x$. Then when X takes the values 1, 2, and 3, for example, the corresponding conditional means are $\mathrm{E}[Y|X=1]=5$, $\mathrm{E}[Y|X=2]=7$, and $\mathrm{E}[Y|X=3]=9$.

The conditional mean function is not necessarily linear. For example, suppose $\mathrm{E}[Y|X=1]=5$, $\mathrm{E}[Y|X=2]=7$, and $\mathrm{E}[Y|X=3]=12$. Then the conditional mean function is nonlinear in X since it increases by 2 from $X=1$ to $X=2$ but increases by 5 from $X=2$ to $X=3$. Chapter 15 presents some more general models for linear regression that relax the assumption of a conditional mean for Y that is linear in X.

B.3.5 Conditional Variance

The **conditional variance** of Y given X measures the variation in Y around the conditional mean $\mathrm{E}[Y|X=x]$, where the deviation is squared and is weighted by the conditional probabilities. Then

$$\mathrm{Var}[Y|X=x] = \mathrm{E}[(Y - \mathrm{E}[Y|X=x])^2|X=x],$$

is the probability-weighted average of all possible values of $(Y - \mathrm{E}[Y|X=x])^2$ when $X=x$.

For the example of this appendix, we have already calculated that $\Pr[Y=50|X=1]=0.5$, $\Pr[Y=30|X=1]=0.5$, and $\mathrm{E}[Y|X=1]=40$. It follows that $\mathrm{Var}[Y|X=1] = (50-40)^2 \times .5 + (30-40)^2 \times .5 = 100$. Similarly, $\mathrm{Var}[Y|X=0] = (30-15)^2 \times .25 + (10-15)^2 \times .75 = 75$. Note that the conditional variance here differs according to whether we condition on $X=0$ or on $X=1$.

Assumption 3 (homoskedastic errors) that $\mathrm{Var}[u|X=x]$ does not depend on x implies that $\mathrm{Var}[Y|X=x]$ does not depend on x. Alternative assumptions regarding $\mathrm{Var}[u|X=x]$ lead to different ways to estimate the precision of the least squares estimates; see Chapter 7.7.

B.3.6 Covariance

The **covariance** of Y and X measures the joint variation of X and Y around their respective means. Let μ_X denote $\mathrm{E}[X]$ and μ_Y denote $\mathrm{E}[Y]$. Then

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathrm{E}[(X-\mu_x)(Y-\mu_Y)] \\ &= \sum\nolimits_x \sum\nolimits_y (X-\mu_x)(Y-\mu_Y) \times \Pr[Y=y|X=x].\end{aligned}$$

It can be shown that $\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mu_x\mu_Y$.

For the example of this appendix, $\mu_X = 1 \times 0.2 + 0 \times 0.8 = 0.2$ and $\mu_Y = 50 \times 0.1 + 30 \times 0.3 + 10 \times 0.6 = 20$. Then $\mathrm{Cov}[X,Y] = (1-0.2) \times (50-20) \times 0.1 + (1-0.2) \times (30-20) \times 0.1 + (0-0.2) \times (30-20) \times 0.2 + (0-0.2) \times (10-20) \times 0.6 = 4$. So $\mathrm{Cov}[X,Y] = 4$.

The covariance between a random variable and a constant is zero. To see this, let a be a constant in which case $\mu_a = \mathrm{E}[a] = a$. Then $\mathrm{Cov}[X,a] = \mathrm{E}[(X-\mu_x)(a-\mu_a)] = \mathrm{E}[(X-\mu_x)(a-a)] = \mathrm{E}[0] = 0$.

The law of iterated expectation states that the overall mean (more formally the unconditional mean) of a random variable is the expected value of the conditional mean, where the expectation is with respect to the conditioning variable. Thus

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y|X=x]] = \sum\nolimits_x \mathrm{E}[Y|X=x] \times \Pr[X=x].$$

One consequence is that if the regression error term has mean zero conditional on the regressors then it has unconditional mean zero, since $\mathrm{E}[u|X=x] = 0$ implies that $\mathrm{E}[u] = \mathrm{E}[\mathrm{E}[u|X=x]] = \mathrm{E}[0] = 0$.

A second consequence is that if the regression error term has mean 0 conditional on regressors then it is uncorrelated with the regressors. To see this use $\mathrm{E}[Xu] = \mathrm{E}[\mathrm{E}[Xu|X=x]] = \mathrm{E}[X \times \mathrm{E}[u|X=x]] = \mathrm{E}[X \times 0] = 0$. So $\mathrm{Cov}[X,u] = \mathrm{E}[Xu] - \mu_x\mu_u = 0 - \mu_x \times 0 = 0$.

B.3.7 Correlation

The **correlation coefficient** of Y and X standardizes the covariance to lie between -1 and 1. We have

$$\mathrm{Cor}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X] \times \mathrm{Var}[Y]}}.$$

For the example of this appendix, $\mathrm{Var}[X] = (1-0.2)^2 \times 0.2 + (0-0.8)^2 \times 0.8 = 0.16$ and $\mathrm{Var}[Y] = (50-20)^2 \times 0.1 + (30-20)^2 \times 0.3 + (10-20)^2 \times 0.6 = 180$. So $\mathrm{Cor}[X,Y] = 4/\sqrt{0.16 \times 180} = 0.745$. In this example X and Y are quite highly positively correlated.

The covariance and correlation coefficients are the population analogs of the sample covariance and sample correlation coefficient that are defined in Chapter 5.4.

Appendix D

Solutions to Selected Exercises

Usually for most odd-numbered questions.

D.1 Solutions: Analysis of Economics Data

1. (a) Numerical time series. (b) Numerical cross-section data. (c) Numerical panel data. (d) Categorical cross-section. data. (e) Numerical repeated cross-section.

3. (a) Observational. (b) Experimental. (c) Observational.

D.2 Solutions: Univariate Data Summary

1. (a) 1+1+1+1+1=5. (b) (1+2+3+4+5)=15. (c) 110. (d) 2.28333. (e) 55.

3. (a) $\sum_{t=1}^{3} x_{1t} = x_{11} + x_{12} + x_{13} = 15$. $\sum_{t=1}^{3} x_{2t} = x_{21} + x_{22} + x_{23} = 18$.
   (b) $\sum_{i=1}^{2} x_{i1} = x_{11} + x_{21} = 5 + 8 = 13$. $\sum_{i=1}^{2} x_{i2} = x_{12} + x_{22} = 9$; $\sum_{i=1}^{2} x_{13} + x_{23} = 11$.

5. $\bar{x} = 2$; $s = \sqrt{8/3} = 1.633$; CV=$\frac{2}{1.663} = 1.225$; Skew= $0/(8/32)^{3/2} = 0$; Kurt= $(32/4)/(8/4)^2 = 2$.

7. Only the mean changes. It becomes 4.

9. $100 \pm 3 \times 14 = (58, 142)$ as 99.7% are within 3 standard deviations of the mean.

11. (a) Median. (b) Mean. (c) No. (d) 12.

13. (b) Data appear right-skewed: skewness $1.56 > 0$ (and mean= 253910 > median= 244000).
    (c) No. Histogram is clearly right-skewed. (d) No. Kernel density estimate is right-skewed.

15. (b) $\bar{x} = 13.033333$; $s = 1.991072$; interquartile range (12, 14).

17. (b) $\bar{x} = 2.7$; $s = 0.8932495$; interquartile range (2.1, 3.3).

21. (a) Somewhat right-skewed (1.01) and excess kurtosis (4.54>3). (b) Quite right-skewed. (c) Quite volatile - big peak in 2000 (technology stocks bubble) plus other peaks including around 1929 (before the Great Depression) and 2018. (d) Yes - over twice the mean (and median) in 2020.

23. (b) Yes (aside from very small computational error). (c) No. Not normal and big spike at low value reflecting the early periods. (d) Yes, the three series appear to move together.

25. (a) Yes. (d) Estimate in (b) is much more volatile than that in (c).

D.3 Solutions: The Sample Mean

1. (a) $\mu = 1000 \times 0.8 + 4000 \times 0.2 = 1{,}800$.
(b) $\sigma^2 = (1000 - 1800)^2 \times 0.8 + (5000 - 1800)^2 \times 0.2 = 2{,}560{,}000$. (c) $\sigma = 1{,}600$.

3. (a) $X + 3$ has mean $5 + 3 = 8$, variance 4 (and standard deviation $\sqrt{4} = 2$).
(b) $\text{E}[2X] = 2 \times 5 = 10$, $\text{Var}[2X] = 2^2 \times 4 = 16$. (c) $\text{E}[2X + 3] = 2 \times 5 + 3 = 13$, $\text{Var}[2X + 3] = 2^2 \times 4 = 16$. (d) $\text{E}[(X - 5)/2] = 0/2 = 0$, $\text{Var}[(X - 5)/2] = 4/2^2 = 1$.

5. (a) 200. (b) $\sigma^2 = 400/100 = 4$; $\sigma = \sqrt{4} = 2$. (c) Most likely given sample size of 100 is reasonably large.

7. (a) Yes. (b) 100. (c) Yes. (d) Data appear random. (e) Observations appear unrelated.

9. $\text{E}[\bar{X}]$ unchanged, $\text{Var}[\bar{X}]$ one-quarter as large, standard deviation of $\bar{X}$ one-half as large.

11. (a) $\mu = 1 \times \frac{1}{6} + 0 \times \frac{5}{6} = \frac{1}{6} \simeq 0.167$. (b) $\sigma^2 = (1 - \frac{1}{6})^2 \times \frac{1}{6} + (0 - \frac{1}{6})^2 \times \frac{5}{6} = \frac{5}{36} \simeq 0.139$. (d) Sample mean and variance should be quite close to 0.167 and 0.139. (e) No.

12. (a) Yes since $\bar{x} = 0.169975 \simeq \frac{1}{6}$ and $s = 0.037151 \simeq \sqrt{\frac{5}{36}/100} = 0.037268$. (b) Histogram is not exactly normal. With a larger sample size we would get closer to exactly normal.

15. (a) $\text{E}[\bar{X}] = 5{,}000$, $\text{St.Dev.}[\bar{X}] = 20000/\sqrt{10000} = 200$. (b) Low probability of 0.025 of making a loss this large since \$5,400 is two standard deviations higher than the expected average loss of \$5,000 and $\bar{X}$ is normally distributed given the large sample. $\Pr(X > 5400) = 0.0228$.

17. (a) $\bar{x} = 100.0564$ and $se(\bar{x}) = 7.97252$ are close to $\mu = 100$ and $\sigma/\sqrt{n} = 16/\sqrt{4} = 8$.
(b) The mean of $se(\bar{x}) = 14.75808$ is close to $\sigma = 16$. (c) $(X - \mu)/(\sigma/\sqrt{n}) = (X - 100)/8$ is normal even in small samples as X is normal. (d) z has mean $0.007 \simeq 0$, standard deviation $0.997 \simeq 1$, skewness $0.028 \simeq 0$, kurtosis $3.177 \simeq 3$. (e) Yes, it appears standard normal.

19. (a) Representative. (b) Most likely unrepresentative. (c) Most likely unrepresentative.

21. (a) 7436390. (b) 467.099. (c) Unweighted mean is 446.66. (d) 32578.1.
(e) $\sqrt{32578.1} = 180.49$ vs. 181.31. (f) In Stata `sum net_worth[weight=number].`

23. (b) Yes, essentially 0 and 1. (c) No. They are very left skewed. (d) Yes.

25. (a) Yes. (d) The part (b) measure is much more variable than the part (c) measure.

D.4 Solutions: Statistical Inference for the Mean

1. (a) 0.028998. (b) 0.057996. (c) 1.717144. (d) 2.073873.

3. (a) 0.025. (b) 0.05. (c) 2.0 (or 1.96 more precise). (d) 2.0 (or 1.96).

5. (b) The $z's$ are draws of $Z = (\overline{X} - \mu)/\sigma$ where X_i's are normal so expect z has mean 0, variance 1 and standard normal distribution.
(c) Similar as the 1,000 z's have mean 0.00705 and standard deviation 0.996.
(d) The $t's$ are draws of $Z = (\overline{X} - \mu)/se(\overline{X})$ where X_i's are normal so Z is t–distributed with $n-1$ degrees of freedom and here $n-1=3$. (e) The 1,000 t's have mean 0.00415 and variance 2.197. (f) The density for t is squashed (lower peak and fatter tails) compared to density for z.

7. $\bar{x} \pm t_{30;.10} \times se(\bar{x}) = 40 \pm 1.3 \times 10 = (27, 53)$.

9. $\bar{x} \pm t_{24;.025} \times (20/\sqrt{25}) = 10 \pm 2.064 \times 4 = (1.744, 18.256)$.

11. (b) $(239688, 268133)$.
(c) $\bar{x} \pm t_{28;.025} \times se(\bar{x}) = 253910.3 \pm 2.0484071 \times 37390.71/\sqrt{29} = (239688, 268133)$.

13. (a) Wider. (b) Narrower. (c) Similar width but larger values. (d) Wider.

15. (b) $H_0 : \mu = 270000$ against $H_a : \mu \neq 270000$. $p = 0.0280 < 0.05$ so reject H_0 at level 0.05.
(c) $c = t_{28,.025} = 2.0484$. $|t| = 2.31730 > 2.0484$ so reject H_0 at level 0.05.
(d) $t = (\bar{x} - \mu^*)/se(\bar{x}) = (253910.3 - 270000)/(37390.71/\sqrt{29}) = -2.31730$.
$p = \Pr[|T_{28}| < |-2.31730|] = 0.0280 < 0.05$.

17. (a) $2.0/\sqrt{25} = 0.4$. (b) $1.2 \pm t_{24,.024} \times 0.4 = (0.3744, 2.0256)$.
(c) $H_0 : \mu = 0$ vs. $H_a : \mu \neq 0$. $t = 1.2/0.4 = 3$. Reject H_0 as $|t| > t_{24;.025} = 2.064$ or as $p = 0.006 < 0.05$.
(d) H_0:$\mu < 0$ vs. H_a:$\mu \geq 0$. $t = 3$. Reject H_0 as $|t| > t_{24;.05} = 1.711$ or as $p = 0.003 < 0.05$. Support the claim.
(e) Observations have common mean and variance and are independent (and for exact t distribution are normally distributed).
(f) Observations are unlikely to be independent and are possibly (downward) biased.

19. (a) Less likely - $p > 0.05$ was given so $p > 0.01$. (b) More likely given more precise estimation.

21. (b) 951 of 1,000 or 95.1% of 95% confidence intervals included 100, close to 95% as expected.
(d) 49 of 1,000 or 4.9% of tests rejected, close to 5% as expected.

23. Answers will vary with package. Stata gave the following.
(b) 953 of 1,000 or 95.3% of 95% confidence intervals included 200.
(d) 47 of 1,000 or 4.7% of tests rejected $H_0 : \mu = 200$.

25. (a) $19.73/\sqrt{109} = 1.890$. (b) (29.06, 36.56). (c) $H_0 : \mu = 35$ against $H_a : \mu \neq 35$. $t = -1.159$. Do not reject H_0 as $|t| > t_{108;.025} = 1.982$ or as $p = 0.249 > 0.05$. (d) $H_0 : \mu > 35$ against $H_a : \mu < 35$. $t = -1.159$. Do not reject H_0 as $t > -t_{108;.05} = -1.659$ or as $p = 0.124 > 0.05$.

27. (a) Do not reject as $p = 0.0868 > 0.05$. (b) Reject as $p = 0.0434 < 0.05$.
(c) $62.468 \pm t_{84,.05} \times 10.0796 = (45.70, 79.23)$.
(d) $t = (62.468 - 35)/10.079 = 2.75$ has $p = 0.007 < 0.05$. So reject H_0.

29. (b) (2.245, 2.515). (c) $H_0 : \mu = 2.5$ against $H_a : \mu \neq 2.5$, $t = (32.81 - 35)/1.89 = -1.758$, $p = 0.081 > 0.05$ so do not reject H_0 at level 0.05.
(d) $H_0 : \mu \geq 2.5$ against $H_a : \mu < 2.5$, $t = -1.758$, $p = 0.040 < 0.05$ so reject H_0 at level 0.05.

31. (a) $25 \pm 2 \times 4 = (17, 33)$. (b) $2 \times 4 = 8$. (c) $t = (25 - 15)/4 = 2.5$. Reject H_0 at level 0.05.

33. (a) $\bar{x} = 550/1025 = 0.5366$ and $se(\bar{x}) = \sqrt{.5366 \times (1 - .5366)/1025)} = 0.0156$.
(b) $0.5366 \pm 1.960 \times 0.0156 = (0.506, 0.567)$. (c) $2 \times 0.0156 = 0.0312$.
(d) $H_0 : \mu \leq 0.5$ vs. $H_a : \mu > 0.5$. $t = (0.5366 - 0.5)/0.0156 = 2.35$. Reject H_0 at level 0.05 as $t = 2.35 > 1.645$ or $p = 0.0094 < 0.05$.

D.5 Solutions: Bivariate Data Summary

1. (a) Yes. (b) $\bar{x} = 4$, $\bar{y} = 14$; $\sum_{i=1}^{3}(x_i - \bar{x})(y_i - \bar{y}) = (2-4)(4-14) + (4-4)(10-14) + (6-4)(28-14) = 48$, so $\text{Cov}(x, y) = 48/2 = 24$. Unclear whether this means a strong relationship.
(c) $\sum_{i=1}^{3}(x_i - \bar{x})^2 = (2-4)^2 + (4-4)^2 + (6-4)^2 = 8$. $\sum_{i=1}^{3}(y_i - \bar{y})^2 = (4-14)^2 + (10-14)^2 + (28-14)^2 = 312$; $\text{Cor}(x, y) = 48/\sqrt{8 \times 312} = 0.9608$. Yes, very strong as close to 1.

3. (a) It is not immediately obvious. It requires some calculation. e.g. *dvisits*=0 for 82% ($100 \times 3686/4491$) with *hospadmi*=0 compared to only 68% with *hospadmi*=1.
(b) Now clearer. e.g. if no relationship we expect *dvisits*=0 and *hospadmi*=0 for 3583.3 which is less than actual 3686. (c) Correlation is 0.2484 so clearly a positive relationship.

5. (a) $\bar{x} = 4$, $\bar{y} = 14$; $\sum_{i=1}^{3}(x_i - \bar{x})(y_i - \bar{y}) = (2-4)(4-14) + (4-4)(10-14) + (6-4)(28-14) = 48$.
$\sum_{i=1}^{3}(x_i - \bar{x})^2 = (2-4)^2 + (4-4)^2 + (6-4)^2 = 8$. So $b_2 = 48/8 = 6$.
(b) $b_1 = \bar{y} - b_2\bar{x} = 14 - 4 \times 6 = -10$.

7. (a)-(b) indicate a strong relationship. (c) Correlation = 0.9315. (d) $price = 82559 + 292 \times size$.
(e) One more square foot is associated with a \$292 increase in house price.

9. (a) $10/(\sqrt{100} \times \sqrt{25}) = 0.2$. (b) $10/100 = 0.1$. (c) 0.2 standard deviations.

11. (c) 0.9949. (b) $realgdppc = 17024 + 169.7 \times quarter$. (e) Real GDP per capita rises on average by \$169.71 each quarter.

				Second decimal value of z						
z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Example: The probability that standard normal is less than 0.65 equals 0.7422.

Example: The probability that standard normal is greater than 0.65 equals 1-0.7422 = 0.2378.

Figure E.1: Standard normal distribution: Probabilities in the left tail

		Significance Level			
Degrees	20% (2-sided)	10% (2-sided)	5% (2-sided)	2% (2-sided)	1% (2-sided)
of Freedom	10%(1-sided)	5%(1-sided)	2.5%(1-sided)	1%(1-sided)	0.5%(1-sided)
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75
60	1.30	1.67	2.00	2.39	2.66
90	1.29	1.66	1.99	2.37	2.63
120	1.29	1.66	1.98	2.36	2.62
Infinity	1.28	1.65	1.96	2.33	2.58

Example: For one-sided test at 10% the area to the right of the critical value is 0.10.
Example: For two-sided test at 10% the area to the right of the critical value is 0.05 and the area to the left of minus one times the critical value is also 0.05.

Figure E.2: Student t distribution: Key critical values