For Erin

Quantitative Methods in Linguistics

Keith Johnson

Blackwell
Publishing

BLACKWELL PUBLISHING
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford OX4 2DQ, UK
550 Swanston Street, Carlton, Victoria 3053, Australia

First published 2008 by Blackwell Publishing Ltd

1 2008

Library of Congress Cataloging-in-Publication Data

Johnson, Keith, 1958–
Quantitative methods in linguistics / Keith Johnson.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4051-4424-7 (hardcover : alk. paper) — ISBN 978-1-4051-4425-4 (pbk. : alk. paper) 1. Linguistics—Statistical methods. I. Title.

P138.5.J64 2008
401.2'1—dc22

2007045515

ISBN-13: 978-1-4051-4424-7 (hardback)
ISBN-13: 978-1-4051-6181-7 (paperback)

A catalogue record for this title is available from the British Library.

Set in Palatino 10/12.5
by Graphicraft Limited Hong Kong

For further information on
Blackwell Publishing, visit our website at
www.blackwellpublishing.com

Contents

Acknowledgments

This book began at Ohio State University and Mary Beckman is largely responsible for the fact that I wrote it. She established a course in "Quantitative Methods in Linguistics" which I also got to teach a few times. Her influence on my approach to quantitative methods can be found throughout this book and in my own research studies, and of course I am very grateful to her for all of the many ways that she has encouraged me and taught me over the years.

I am also very grateful to a number of colleagues from a variety of institutions who have given me feedback on this volume, including: Susanne Gahl, Chris Manning, Christine Mooshammer, Geoff Nicholls, Gerald Penn, Bonny Sands, and a UC San Diego student reading group led by Klinton Bicknell. Students at Ohio State also helped sharpen the text and exercises – particularly Kathleen Currie-Hall, Matt Makashay, Grant McGuire, and Steve Winters. I appreciate their feedback on earlier handouts and drafts of chapters. Grant has also taught me some R graphing strategies. I am very grateful to UC Berkeley students Molly Babel, Russell Lee-Goldman, and Reiko Kataoka for their feedback on several of the exercises and chapters. Shira Katseff deserves special mention for reading the entire manuscript during fall 2006, offering copy-editing and substantive feedback. This was extremely valuable detailed attention – thanks! I am especially grateful to OSU students Amanda Boomershine, Hope Dawson, Robin Dodsworth, and David Durian who not only offered comments on chapters but also donated data sets from their own very interesting research projects. Additionally, I am very grateful to Joan Bresnan, Beth Hume, Barbara Luka, and Mark Pitt for sharing data sets for this book. The generosity and openness of all of these "data donors" is a high

standard of research integrity. Of course, they are not responsible for any mistakes that I may have made with their data. I wish that I could have followed the recommendation of Johanna Nichols and Balthasar Bickel to add a chapter on typology. They were great, donating a data set and a number of observations and suggestions, but in the end I ran out of time. I hope that there will be a second edition of this book so I can include typology – and perhaps by then some other areas of linguistic research as well.

Finally, I would like to thank Nancy Dick-Atkinson for sharing her cabin in Maine with us in the summer of 2006, and Michael for the whiffle-ball breaks. What a nice place to work!

Design of the Book

One thing that I learned in writing this book is that I had been wrongly assuming that we phoneticians were the main users of quantitative methods in linguistics. I discovered that some of the most sophisticated and interesting quantitative techniques for doing linguistics are being developed by sociolinguists, historical linguists, and syntacticians. So, I have tried with this book to present a relatively representative and usable introduction to current quantitative research across many different subdisciplines within linguistics.[1]

The first chapter "Fundamentals of quantitative analysis" is an overview of, well, fundamental concepts that come up in the remainder of the book. Much of this will be review for students who have taken a general statistics course. The discussion of probability distributions in this chapter is key. Least-square statistics – the mean and standard deviation, are also introduced.

The remainder of the chapters introduce a variety of statistical methods in two thematic organizations. First, the chapters (after the second general chapter on "Patterns and tests") are organized by linguistic subdiscipline – phonetics, psycholinguistics, sociolinguistics, historical linguistics, and syntax.

[1] I hasten to add that, even though there is very much to be gained by studying techniques in natual language processing (NLP), this book is not a language engineering book. For a very authoritative introduction to NLP I would recommend Manning and Schütze's *Foundations of Statistical Natural Language Processing* (1999).

This organization provides some familiar landmarks for students and a convenient backdrop for the other organization of the book which centers around an escalating degree of modeling complexity culminating in the analysis of syntactic data. To be sure, the chapters do explore some of the specialized methods that are used in particular disciplines – such as principal components analysis in phonetics and cladistics in historical linguistics – but I have also attempted to develop a coherent progression of model complexity in the book.

Thus, students who are especially interested in phonetics are well advised to study the syntax chapter because the methods introduced there are more sophisticated and potentially more useful in phonetic research than the methods discussed in the phonetics chapter! Similarly, the syntactician will find the phonetics chapter to be a useful precursor to the methods introduced finally in the syntax chapter.

The usual statistics textbook introduction suggests what parts of the book can be skipped without a significant loss of comprehension. However, rather than suggest that you ignore parts of what I have written here (naturally, I think that it was all worth writing, and I hope it will be worth your reading) I refer you to Table 0.1 that shows the continuity that I see among the chapters.

The book examines several different methods for testing research hypotheses. These focus on building statistical models and evaluating them against one or more sets of data. The models discussed in the book include the simple *t*-test which is introduced in Chapter 2 and elaborated in Chapter 3, analysis of variance (Chapter 4), logistic regression (Chapter 5), linear mixed effects models and logistic linear mixed effects models discussed in Chapter 7. The progression here is from simple to complex. Several methods for discovering patterns in data are also discussed in the book (in Chapters 2, 3, and 6) in progression from simpler to more complex. One theme of the book is that despite our different research questions and methodologies, the statistical methods that are employed in modeling linguistic data are quite coherent across subdisciplines and indeed are the same methods that are used in scientific inquiry more generally. I think that one measure of the success of this book will be if the student can move from this introduction – oriented explicitly around linguistic data – to more general statistics reference books. If you are able to make this transition I think I will have succeeded in helping you connect your work to the larger context of general scientific inquiry.

Table 0.1 The design of the book as a function of statistical approach (hypothesis testing vs. pattern discovery), type of data, and type of predictor variables.

Hypothesis testing		*Predictor variables*		
		Factorial (nominal)	**Continuous**	**Mixed random and fixed factors**
Type of data	Ratio (continuous)	*t*-test (Chs 2 & 3)	Linear regression (Chs 2 & 3)	Repeated measures ANOVA (Ch. 4)
		ANOVA (Ch. 4)		Linear mixed effects (Ch. 7)
	Nominal (counting)	χ^2 test (Ch. 5) Logistic regression (Ch. 5)	Logistic regression (Ch. 5)	Logistic linear mixed effects (Ch. 7)

Pattern discovery		*Type of pattern*	
		Categories	**Continuous**
Type of data	Many continuous dimensions	Principal components (Ch. 3)	Linear regression (Ch. 3) Principal components (Ch. 3)
	Distance matrix	Clustering (Ch. 6) MD Scaling (Ch. 6)	
	Shared traits	Cladistics (Ch. 6)	

A Note about Software

One thing that you should be concerned with in using a book that devotes space to learning how to use a particular software package is that some software programs change at a relatively rapid pace.

In this book, I chose to focus on a software package (called "R") that is developed under the GNU license agreement. This means that the software is maintained and developed by a user community and is distributed not for profit (students can get it on their home computers at no charge). It is serious software. Originally developed at AT&T Bell Labs, it is used extensively in medical research, engineering, and

science. This is significant because GNU software (like Unix, Java, C, Perl, etc.) is more stable than commercially available software – revisions of the software come out because the user community needs changes, not because the company needs cash. There are also a number of electronic discussion lists and manuals covering various specific techniques using R. You'll find these resources at the R project web page (http://www.r-project.org).

At various points in the text you will find short tangential sections called "R notes." I use the R notes to give you, in detail, the command language that was used to produce the graphs or calculate the statistics that are being discussed in the main text. These commands have been student tested using the data and scripts that are available at the book web page, and it should be possible to copy the commands verbatim into an open session of R and reproduce for yourself the results that you find in the text. The aim of course is to reduce the R learning curve a bit so you can apply the concepts of the book as quickly as possible to your own data analysis and visualization problems.

Contents of the Book Web Site

The data sets and scripts that are used as examples in this book are available for free download at the publisher's web site – www.blackwellpublishing.com. The full listing of the available electronic resources is reproduced here so you will know what you can get from the publisher.

Chapter 2 Patterns and Tests

Script: Figure 2.1.
Script: The central limit function from a uniform distribution (central.limit.unif).
Script: The central limit function from a skewed distribution (central.limit).
Script: The central limit function from a normal distribution.
Script: Figure 2.5.
Script: Figure 2.6 (shade.tails)
Data: Male and female F1 frequency data (F1_data.txt).
Script: Explore the chi-square distribution (chisq).

Chapter 3 Phonetics

Data: Cherokee voice onset times (cherokeeVOT.txt).
Data: The tongue shape data (chaindata.txt).
Script: Commands to calculate and plot the first principal component of tongue shape.
Script: Explore the F distribution (shade.tails.df).
Data: Made-up regression example (regression.txt).

Chapter 4 Psycholinguistics

Data: One observation of phonological priming per listener from Pitt and Shoaf's (2002).
Data: One observation per listener from two groups (overlap versus no overlap) from Pitt and Shoaf's study.
Data: Hypothetical data to illustrate repeated measures of analysis.
Data: The full Pitt and Shoaf data set.
Data: Reaction time data on perception of flap, /d/, and eth by Spanish-speaking and English-speaking listeners.
Data: Luka and Barsalou (2005) "by subjects" data.
Data: Luka and Barsalou (2005) "by items" data.
Data: Boomershine's dialect identification data for exercise 5.

Chapter 5 Sociolinguistics

Data: Robin Dodsworth's preliminary data on /l/ vocalization in Worthington, Ohio.
Data: Data from David Durian's rapid anonymous survey on /str/ in Columbus, Ohio.
Data: Hope Dawson's Sanskrit data.

Chapter 6 Historical Linguistics

Script: A script that draws Figure 6.1.
Data: Dyen, Kruskal, and Black's (1984) distance matrix for 84 Indo-European languages based on the percentage of cognate words between languages.
Data: A subset of the Dyen et al. (1984) data coded as input to the Phylip program "pars."

Data: IE-lists.txt: A version of the Dyen et al. word lists that is readable in the scripts below.
Script: make_dist: This Perl script tabulates all of the letters used in the Dyen et al. word lists.
Script: get_IE_distance: This Perl script implements the "spelling distance" metric that was used to calculate distances between words in the Dyen et al. list.
Script: make_matrix: Another Perl script. This one takes the output of get_IE_distance and writes it back out as a matrix that R can easily read.
Data: A distance matrix produced from the spellings of words in the Dyen et al. (1984) data set.
Data: Distance matrix for eight Bantu languages from the Tanzanian Language Survey.
Data: A phonetic distance matrix of Bantu languages from Ladefoged, Glick, and Criper (1971).
Data: The TLS Bantu data arranged as input for phylogenetic parsimony analysis using the Phylip program pars.

Chapter 7 Syntax

Data: Results from a magnitude estimation study.
Data: Verb argument data from CoNLL-2005.
Script: Cross-validation of linear mixed effects models.
Data: Bresnan et al.'s (2007) dative alternation data.

1 Fundamentals of Quantitative Analysis

In this chapter, I follow the outline of topics used in the first chapter of Kachigan, *Multivariate Statistical Analysis*, because I think that that is a very effective presentation of these core ideas.

Increasingly, linguists handle quantitative data in their research. Phoneticians, sociolinguists, psycholinguists, and computational linguists deal in numbers and have for decades. Now also, phonologists, syntacticians, and historical linguists are finding linguistic research to involve quantitative methods. For example, Keller (2003) measured sentence acceptibility using a psychophysical technique called magnitude estimation. Also, Boersma and Hayes (2001) employed probablistic reasoning in a constraint reranking algorithm for optimality theory.

Consequently, mastery of quantitative methods is increasingly becoming a vital component of linguistic training. Yet, when I am asked to teach a course on quantitative methods I am not happy with the available textbooks. I hope that this book will deal adequately with the fundamental concepts that underlie common quantitative methods, and more than that will help students make the transition from the basics to real research problems with explicit examples of various common analysis techniques.

Of course, the strategies and methods of quantitative analysis are of primary importance, but in these chapters practical aspects of handling quantitative linguistic data will also be an important focus. We will be concerned with how to use a particular statistical package (R) to discover patterns in quantitative data and to test linguistic hypotheses. This theme is very practical and assumes that it is appropriate and useful to look at quantitative measures of language structure and usage.

We will question this assumption. Salsburg (2001) talks about a "statistical revolution" in science in which the distributions of

measurements are the objects of study. We will, to some small extent, consider linguistics from this point of view. Has linguistics participated in the statistical revolution? What would a quantitative linguistics be like? Where is this approach taking the discipline?

Table 1.1 shows a set of phonetic measurements. These VOT (voice onset time) measurements show the duration of aspiration in voiceless stops in Cherokee. I made these measurements from recordings of one

Table 1.1 Voice onset time measurements of a single Cherokee speaker with a 30-year gap between recordings.

	1971		2001
k	67	k	84
k	127	k	82
k	79	k	72
k	150	k	193
k	53	k	129
k	65	k	77
k	75	k	72
k	109	k	81
t	109	k	45
t	126	k	74
t	129	k	102
t	119	k	77
t	104	k	187
t	153	t	79
t	124	t	86
t	107	t	59
t	181	t	74
t	166	t	63
		t	75
		t	70
		t	106
		t	54
		t	49
		t	56
		t	58
		t	97
Average	113.5		84.7
Standard Deviation	35.9		36.09

speaker, the Cherokee linguist Durbin Feeling, that were made in 1971 and 2001. The average VOT for voiceless stops /k/ and /t/ is shorter in the 2001 dataset. But is the difference "significant"? Or is the difference between VOT in 1971 and 2001 just an instance of random variation – a consequence of randomly selecting possible utterances in the two years that, though not identical, come from the same underlying distribution of possible VOT values for this speaker? I think that one of the main points to keep in mind about drawing conclusions from data is that it is all guessing. Really. But what we are trying to do with statistical summaries and hypothesis testing is to quantify just how reliable our guesses are.

1.1 What We Accomplish in Quantitative Analysis

Quantitative analysis takes some time and effort, so it is important to be clear about what you are trying to accomplish with it. Note that "everybody seems to be doing it" is not on the list. The four main goals of quantitative analysis are:

1 data reduction: summarize trends, capture the common aspects of a set of observations such as the average, standard deviation, and correlations among variables;
2 inference: generalize from a representative set of observations to a larger universe of possible observations using hypothesis tests such as the *t*-test or analysis of variance;
3 discovery of relationships: find descriptive or causal patterns in data which may be described in multiple regression models or in factor analysis;
4 exploration of processes that may have a basis in probability: theoretical modeling, say in information theory, or in practical contexts such as probabilistic sentence parsing.

1.2 How to Describe an Observation

An observation can be obtained in some elaborate way, like visiting a monastery in Egypt to look at an ancient manuscript that hasn't been read in a thousand years, or renting an MRI machine for an hour of brain imaging. Or an observation can be obtained on the cheap –

asking someone where the shoes are in the department store and noting whether the talker says the /r/'s in "fourth floor."

Some observations can't be quantified in any meaningful sense. For example if that ancient text has an instance of a particular form and your main question is "how old is the form?" then your result is that the form is at least as old as the manuscript. However, if you were to observe that the form was used 15 times in this manuscript, but only twice in a slightly older manuscript, then these frequency counts begin to take the shape of quantified linguistic observations that can be analyzed with the same quantitative methods used in science and engineering. I take that to be a good thing – linguistics as a member of the scientific community.

Each observation will have several descriptive properties – some will be qualitative and some will be quantitative – and descriptive properties (variables) come in one of four types:

Nominal: Named properties – they have no meaningful order on a scale of any type.

Examples: What language is being observed? What dialect? Which word? What is the gender of the person being observed? Which variant was used: *going* or *goin'*?

Ordinal: Orderable properties – they aren't observed on a measurable scale, but this kind of property is transitive so that if a is less than b and b is less than c then a is also less than c.

Examples: Zipf's rank frequency of words, rating scales (e.g. excellent, good, fair, poor)?

Interval: This is a property that is measured on a scale that does not have a true zero value. In an interval scale, the magnitude of differences of adjacent observations can be determined (unlike the adjacent items on an ordinal scale), but because the zero value on the scale is arbitrary the scale cannot be interpreted in any absolute sense.

Examples: temperature (Fahrenheit or Centigrade scales), rating scales?, magnitude estimation judgments.

Ratio: This is a property that we measure on a scale that does have an absolute zero value. This is called a ratio scale because ratios of these measurements are meaningful. For instance, a vowel that is 100 ms long is twice as long as a 50 ms vowel, and 200 ms is twice 100 ms. Contrast this with temperature – 80 degrees Fahrenheit is not twice as hot as 40 degrees.

Examples: Acoustic measures – frequency, duration, frequency counts, reaction time.

1.3 Frequency Distributions: A Fundamental Building Block of Quantitative Analysis

You must get this next bit, so pay attention. Suppose we want to know how grammatical a sentence is. We ask 36 people to score the sentence on a grammaticality scale so that a score of 1 means that it sounds pretty ungrammatical and 10 sounds perfectly OK. Suppose that the ratings in Table 1.2 result from this exercise.

Interesting, but what are we supposed to learn from this? Well, we're going to use this set of 36 numbers to construct a frequency

Table 1.2 Hypothetical data of grammaticality ratings for a group of 36 raters.

Person #	Rating	Person #	Rating
1	5	19	3
2	4	20	9
3	6	21	5
4	5	22	6
5	5	23	5
6	4	24	1
7	6	25	5
8	1	26	7
9	4	27	4
10	3	28	5
11	6	29	2
12	3	30	4
13	4	31	5
14	5	32	3
15	4	33	3
16	5	34	6
17	5	35	3
18	4	36	5

distribution and define some of the terms used in discussing frequency distributions.

R note. I guess I should confess that I made up the "ratings" in Table 1.2. I used a function in the R statistics package to draw 36 random integer observations from a normal distribution that had a mean value of 4.5 and a standard deviation of 2. Here's the command that I used to produce the made up data:

```
> round(rnorm(36,4.5,2))
```

If you issue this command in R you will almost certainly get a different set of ratings (that's the nature of random selection), but the distribution of your scores should match the one in the example.

Look again at Table 1.2. How many people gave the sentence a rating of "1"? How many rated it a "2"? When we answer these questions for all of the possible ratings we have the values that make up the frequency distribution of our sentence grammaticality ratings. These data and some useful recodings of them are shown in Table 1.3.

You'll notice in Table 1.3 that we counted two instances of rating "1", one instance of rating "2", six instances of rating "3", and so on. Since there were 36 raters, each giving one score to the sentence, we have a total of 36 observations, so we can express the frequency counts in relative terms – as a percentage of the total number of observations. Note that percentages (as the etymology of the word would suggest) are commonly expressed on a scale from 0 to 100, but you could express the same information as proportions ranging from 0 to 1.

The frequency distribution in Table 1.3 shows that most of the grammaticality scores are either "4" or "5," and that though the scores span a wide range (from 1 to 9) the scores are generally clustered in the middle of the range. This is as it should be because I selected the set of scores from a normal (bell-shaped) frequency distribution that centered on the average value of 4.5 – more about this later.

The set of numbers in Table 1.3 is more informative than the set in Table 1.2, but nothing beats a picture. Figure 1.1 shows the frequencies from Table 1.3. This figure highlights, for the visually inclined, the same points that we made regarding the numeric data in Table 1.3.

Table 1.3 Frequency distributions of the grammaticality rating data in Table 1.2.

Rating	Frequencies	Relative frequencies	Cumulative frequencies	Relative cumulative frequencies
1	2	5.6	2	5.6
2	1	2.8	3	8.3
3	6	16.7	9	25.0
4	8	22.2	17	47.2
5	12	33.3	29	80.6
6	5	13.9	34	94.4
7	1	2.8	35	97.2
8	0	0.0	35	97.2
9	1	2.8	36	100.0
Total	36	100.0	36	100.0

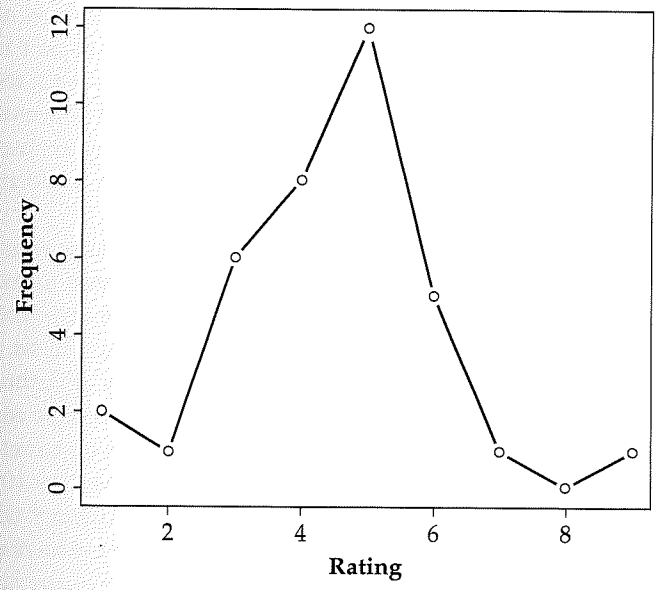

Figure 1.1 The frequency distribution of the grammaticality rating data that was presented in Table 1.2.

R note. I produced Figure 1.1 using the `plot()` command in R. I first typed in the frequency count data and the scores that correspond to these frequency counts, so that the vector `data` contains the counts and the vector `rating` has the rating values. Then I told `plot()` that we want a line plot with both (`type = "b"`) lines and points.

```
data = c(2,1,6,8,12,5,1,0,1)
rating = c(1,2,3,4,5,6,7,8,9)
plot(rating,data,type = "b", main = "Sentence rating frequency
     distribution", xlab = "Rating", ylab = "Frequency")
```

The property that we are seeking to study with the "grammaticality score" measure is probably a good deal more gradient than we permit by restricting our rater to a scale of integer numbers. It may be that not all sentences that he/she would rate as a "5" are exactly equivalent to each other in the internal feeling of grammaticality that they evoke. Who knows? But suppose that it is true that the internal grammaticality response that we measure with our rating scale is actually a continuous, gradient property. We could get at this aspect by providing a more and more continuous type of rating scale – we'll see more of this when we look at magnitude estimation later – but whatever scale we use, it will have some degree of granularity or quantization to it. This is true of all of the measurement scales that we could imagine using in any science.

So, with a very fine-grained scale (say a grammaticality rating on a scale with many decimal points) it doesn't make any sense to count the number of times that a particular measurement value appears in the data set because it is highly likely that no two ratings will be exactly the same. In this case, then, to describe the frequency distribution of our data we need to group the data into contiguous ranges of scores (bins) of similar values and then count the number of observations in each bin. For example, if we permitted ratings on the 1 to 10 grammaticality scale to have many decimal places, the frequency distribution would look like the histogram in Figure 1.2, where we have a count of 1 for each rating value in the data set.

Figure 1.3 shows how we can group these same data into ranges (here ratings between 0 and 1, 1 and 2, and so on) and then count the number of rating values in each range, just as we counted before, the

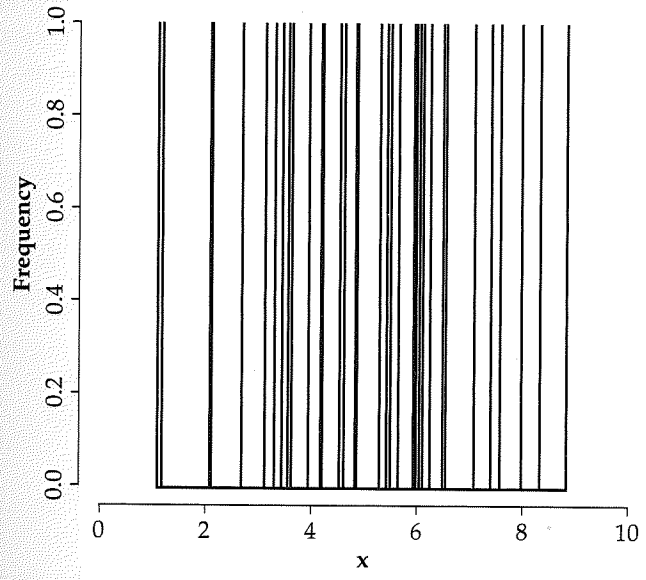

Figure 1.2 A histogram of the frequency distribution of grammaticality ratings when rating values come on a continuous scale.

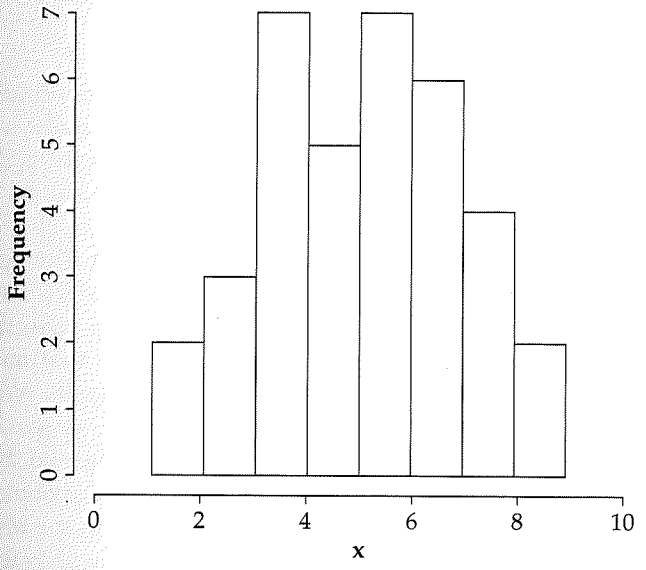

Figure 1.3 The same continuous rating data that was shown in Figure 1.2, but now the frequency distribution is plotted in bins.

number of ratings of a particular value. So, instead of counting the number of times the rating "6" was given, now we are counting the number of ratings that are greater than or equal to 5 and less than 6.

R note. The histograms in figures 1.2 and 1.3 are really easy to produce in R. First, I produced a small set of 36 "observations" from a normal distribution that has a mean rating of 4.5 and a standard deviation of ratings of 2.

```
x = rnorm(36, 4.5, 2)
```

Then to produce Figure 1.2, I used the `hist()` command and told it that I wanted lots and lots of vertical bars. This large number gave me a separate bar for each of the 36 observations in the "x" data set.

```
hist(x, breaks = 30000,xlim = c(0,10))
```

Then to produce Figure 1.3, I used the same command, this time permitting the command to choose a good bar width for my data. Nice that the simpler command gives you the more sensible output.

```
hist(x,xlim = c(0,10))
```

OK. This process of grouping measurements on a continuous scale is a useful, practical thing to do, but it helps us now make a serious point about theoretical frequency distributions. This point is the *foundation* of all of the hypothesis testing statistics that we will be looking at later. So, pay attention!

Let's suppose that we could draw an infinite data set. The larger our data set becomes the more detailed a representation of the frequency distribution we can get. For example, suppose I keep collecting sentence grammaticality data for the same sentence, so that instead of ratings from 36 people I had ratings from 10,000 people. Now even with a histogram that has 1,000 bars in it (Figure 1.4), we can see that ratings near 4.5 are more common than those at the edges of the rating scale. Now if we keep adding observations up to infinity (just play along with me here) and keep reducing the size of the bars in the histogram of the frequency distribution we come to a point at which the intervals between bars is vanishingly small – i.e. we end up with a continuous curve (see Figure 1.5). "Vanishingly small" should be a tip-off

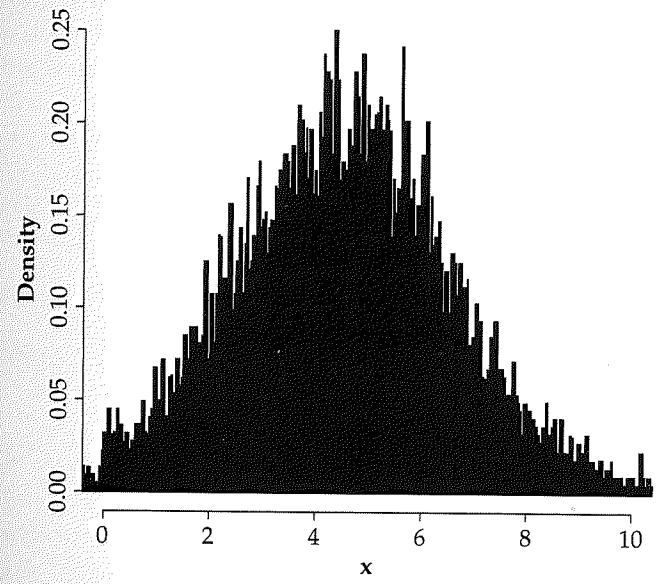

Figure 1.4 A frequency histogram with 1,000 bars plotting frequency in 10,000 observations.

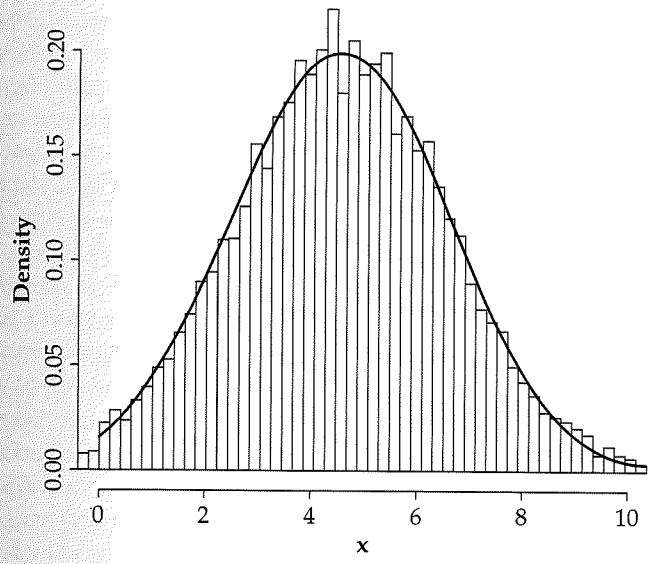

Figure 1.5 The probability density distribution of 10,000 observations and the theoretical probability density distribution of a normal distribution with a mean of 4.5 and a standard deviation of 2.

that we have entered the realm of calculus. Not to worry though, we're not going too far.

R note. The cool thing about Figure 1.5 is that we combine a histogram of the observed frequency distribution of a set of data with a theoretical normal distribution curve (see Chapter 2 regarding probability density). It is useful to be able to do this. Here are the commands I used:

```
x = rnorm(10000, 4.5, 2) # generate 10,000 data points
hist(x,breaks=100,freq=FALSE,xlim = c(0,10)) # plot them
in a histogram
  # now plot the normal curve
plot(function(x)dnorm(x, mean=4.5, sd=2), 0,10, add=TRUE)
```

Of course, the excellent fit between the "observed" and the theoretical distributions is helped by the fact that the data being plotted here were generated by random selection (`rnorm()`) of observations from the theoretical normal distribution (`dnorm()`).

The "normal distribution" is an especially useful theoretical function. It seems intuitively reasonable to assume that in most cases there is some underlying property that we are trying to measure – like grammaticality, or typical duration, or amount of processing time – and that there is some source of random error that keeps us from getting an exact measurement of the underlying property. If this is a good description of the source of variability in our measurements, then we can model this situation by assuming that the underlying property – the uncontaminated "true" value that we seek – is at the center of the frequency distribution that we observe in our measurements and that the spread of the distribution is caused by error, with bigger errors being less likely to occur than smaller errors.

These assumptions give us a bell-shaped frequency distribution which can be described by the normal curve, an extremely useful bell-shaped curve, which is an exponential function of the mean value (Greek letter μ "mew") and the variance (Greek letter σ "sigma").

$$f_x = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad \text{the normal distribution}$$

One useful aspect of this definition of a theoretical distribution of data (besides that it derives from just two numbers, the mean value and a measure of how variable the data are) is that sum of the area under the curve f_x is 1. So, instead of thinking in terms of a "frequency" distribution, the normal curve gives us a way to calculate the probability of any set of observations by finding the area under any portion of the curve. We'll come back to this.

1.4 Types of Distributions

Data come in a variety of shapes of frequency distributions (Figure 1.6).

For example, if every outcome is equally likely then the distribution is *uniform*. This happens for example with the six sides of a dice – each one is (supposed to be) equally likely, so if you count up the number of rolls that come up "1" it should be on average 1 out of every 6 rolls.

In the *normal* – bell-shaped – distribution, measurements tend to congregate around a typical value and values become less and less likely as they deviate further from this central value. As we saw in the section above, the normal curve is defined by two parameters – what the central tendency is (μ) and how quickly probability goes down as you move away from the center of the distribution (σ).

If measurements are taken on a scale (like the 1–9 grammaticality rating scale discussed above), as we approach one end of the scale the frequency distribution is bound to be *skewed* because there is a limit beyond which the data values cannot go. We most often run into skewed frequency distributions when dealing with percentage data and reaction time data (where negative reaction times are not meaningful).

The *J-shaped* distribution is a kind of skewed distribution with most observations coming from the very end of the measurement scale. For example, if you count speech errors per utterance you might find that most utterances have a speech error count of 0. So in a histogram, the number of utterances with a low error count will be very high and will decrease dramatically as the number of errors per utterance increases.

A *bimodal* distribution is like a combination of two normal distributions – there are two peaks. If you find that your data fall in a bimodal distribution you might consider whether the data actually represent two separate populations of measurements. For example, voice fundamental frequency (the acoustic property most closely related to the pitch of a person's voice) falls into a bimodal distribution when you

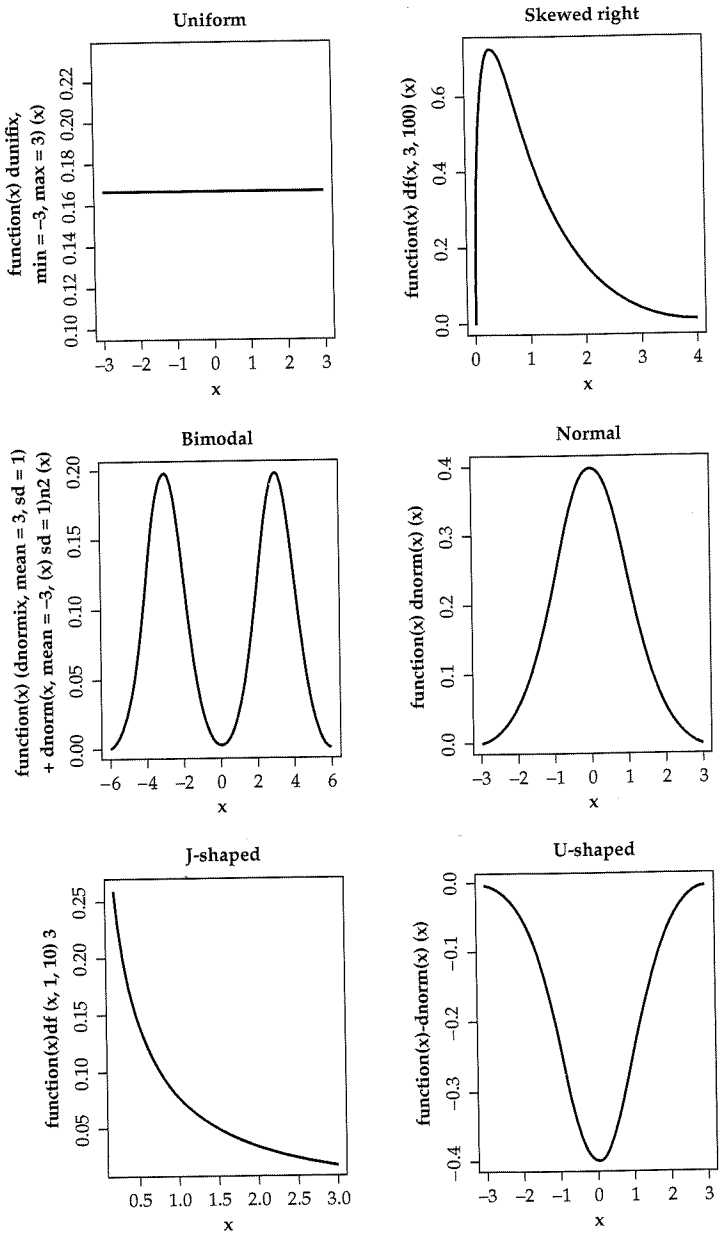

Figure 1.6 Types of probability distributions.

pool measurements from men and women because men tend to have lower pitch than women.

If you ask a number of people how strongly they supported the US invasion of Iraq you would get a very polarized distribution of results. In this *U-shaped* distribution most people would be either strongly in favor or strongly opposed with not too many in the middle.

R note. Figure 1.6 not only illustrates different types of probability distributions, it also shows how to combine several graphs into one figure in R. The command `par()` lets you set many different graphics parameters. I set the graph window to expect two rows that each have three graphs by entering this command:

```
> par(mfcol=c(2,3))
```

Then I entered the six plot commands in the following order:

```
> plot(function(x)dunif(x,min=-3,max=3),-3,3,
main="Uniform")
> plot(function(x)dnorm(x),-3,3, main="Normal")
> plot(function(x)df(x,3,100),0,4,main="Skewed right")
> plot(function(x)df(x,1,10)/3,0.2,3, main="J-shaped")
> plot(function(x)(dnorm(x, mean=3, sd=1)+dnorm(x,mean=-3,
sd=1))/2,-6,6, main="Bimodal")
> plot(function(x)-dnorm(x),-3,3,main="U-shaped")
```

And, voilá. The figure is done. When you know that you will want to repeat the same, or a very similar, sequence of commands for a new data set, you can save a list of commands like this as a custom command, and then just enter your own "plot my data my way" command.

1.5 Is Normal Data, Well, Normal?

The normal distribution is a useful way to describe data. It embodies some reasonable assumptions about how we end up with variability in our data sets and gives us some mathematical tools to use in two important goals of statistical analysis. In data reduction, we can describe the whole frequency distribution with just two numbers – the

mean and the standard deviation (formal definitions of these are just ahead.) Also, the normal distribution provides a basis for drawing inferences about the accuracy of our statistical estimates.

So, it is a good idea to know whether or not the frequency distribution of your data is shaped like the normal distribution. I suggested earlier that the data we deal with often falls in an approximately normal distribution, but as discussed in section 1.4, there are some common types of data (like percentages and rating values) that are not normally distributed.

We're going to do two things here. First, we'll explore a couple of ways to determine whether your data are normally distributed, and second we'll look at a couple of transformations that you can use to *make* data more normal (this may sound fishy, but transformations are legal!).

Consider again the Cherokee data that we used to start this chapter. We have two sets of data, thus, two distributions. So, when we plot the frequency distribution as a histogram and then compare that observed distribution with the best-fitting normal curve we can see that both the 2001 and the 1971 data sets are fairly similar to the normal curve. The 2001 set (Figure 1.7) has a pretty normal looking shape, but there are a couple of measurements at nearly 200 ms that hurt the fit. When we remove these two, the fit between the theoretical normal curve and the frequency distribtion of our data is quite good. The 1971 set

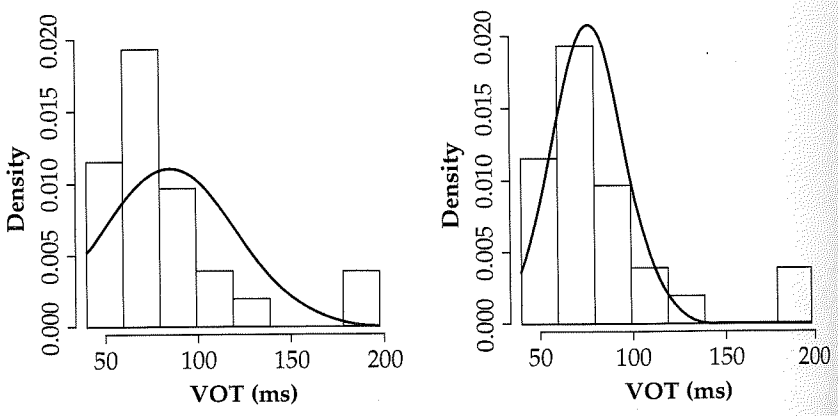

Figure 1.7 The probability density distribution of the Cherokee 2001 voice onset time data. The left panel shows the best-fitting normal curve for all of the data points. The right panel shows the best-fitting normal curve when the two largest VOT values are removed from the data set.

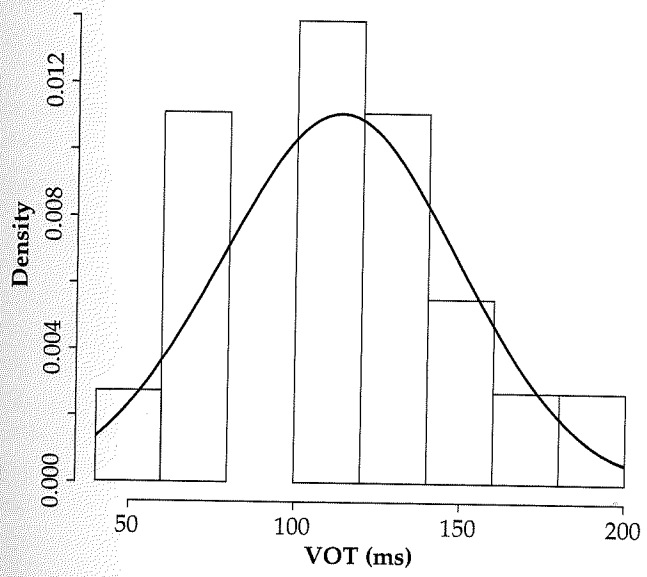

Figure 1.8 The probability density distribution of the Cherokee 1971 voice onset time data. The best-fitting normal curve is also shown.

(Figure 1.8) also looks roughly like a normally distributed data set, though notice that there were no observations between 80 and 100 ms in this (quite small) data set. Though if these data came from a normal curve we would have expected several observations in this range.

R note. In producing Figures 1.7 and 1.8, I used the `c()` function to type in the two vectors `vot01` for the 2001 data and `vot71` for the 1971 data. Just ahead I'll introduce methods for reading data from computer files into R – you don't usually have to type in your data. Then I used the `mean()` and `sd()` functions to calculate the means and standard deviations for these data sets. Finally, I used the `hist()` and `plot()` commands to draw the actual and theoretical frequency distributions in the figures.

```
> vot01 = c(84, 82, 72,193, 129, 77, 72, 81, 45, 74, 102, 77,
187, 79, 86, 59, 74, 63, 75, 70, 106, 54, 49, 56, 58, 97)
> vot71 =c(67, 127, 79, 150, 53, 65, 75, 109, 109, 126, 129,
119, 104, 153, 124, 107, 181, 166)
```

```
> mean(vot01)
  [1] 84.65385
> sd(vot01)
  [1] 36.08761
> hist(vot01,freq=FALSE)
> plot(function(x)dnorm(x, mean=84.654, sd=36.088), 40,
200, add=TRUE)
```

You might also be interested to see how to take the mean and standard deviation with outliers removed. I decided that the two VOT measurements in vot01 that are greater than 180 ms are outliers and so calcuated the mean and standard deviation for only those numbers in the vector that are less than 180 using the following statements.

```
> mean(vot01[vot01<180])
[1] 75.875
> sd(vot01[vot01<180])
[1] 19.218
```

Read vot01[vot01<180] as "the numbers in vot01 that are less than 180." When we have data sets that are composed of several linked vectors we can extract subsets of data using similar syntax.

Note though that I have just told you how to "remove outliers" as if it is perfectly fine to remove weird data. It is not! You should use *all* of the data you collect unless you have good independent reasons for not doing so. For example, data values can be removed if you know that there has been some measurement error that results in the weird value, or if you know that the person providing the data was different from the other participants in the study in some way that bears on the aims of the study (e.g. by virtue of having fallen asleep during a perception experiment, or by not being a native speaker of the language under study), or the token is different in some crucial way (e.g. by virtue of being spoken in error or with a disfluency). Because such variation on the part of the people we study is bound to happen, it is acceptable to trim the 5% most extreme data values from a large and noisy database where manual inspection of the entire database is not practical.

These frequency distribution graphs give an indication of whether our data is distributed on a normal curve, but we are essentially waving our hands at the graphs and saying "looks pretty normal to me." I guess you shouldn't underestimate how important it is to look at the data, but it would be good to be able to measure just how "normally distributed" these data are.

To do this we measure the degree of fit between the data and the normal curve with a quantile/quantile plot and a correlation between the actual quantile scores and the quantile scores that are predicted by the normal curve. The NIST *Handbook of Statistical Methods* (2004) has this to say about Q-Q plots.

> The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
>
> A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
>
> A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
>
> The advantages of the q-q plot are:
>
> 1 The sample sizes do not need to be equal.
> 2 Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.
>
> The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

Further regarding the "probability plot" the *Handbook* has this to say:

> The probability plot (Chambers et al. 1983) is a graphical technique for assessing whether or not a data set follows a given distribution such as the normal or Weibull.

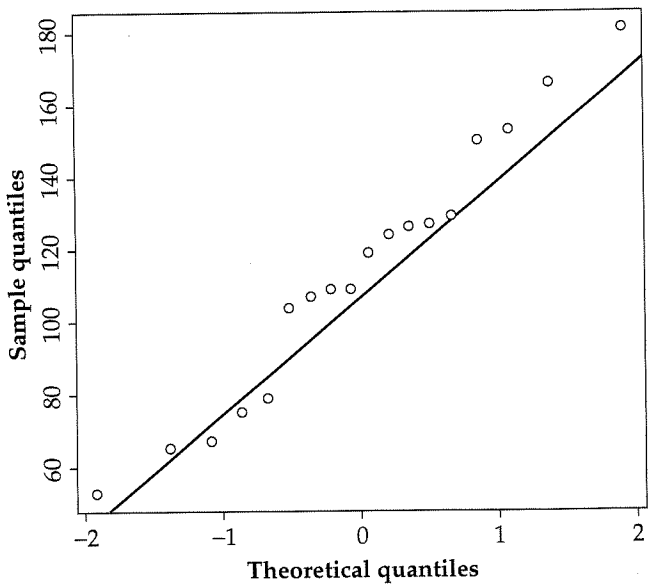

Figure 1.9 The quantiles-quantiles probability plot comparing the Cherokee 1971 data with the normal distribution.

The data are plotted against a theoretical distribution in such a way that the points should form approximately a straight line. Departures from this straight line indicate departures from the specified distribution.

As you can see in Figure 1.9 the Cherokee 1971 data are just as you would expect them to be if they came from a normal distribution. In fact, the data points are almost all on the line showing perfect identity between the expected "Theoretical quantiles" and the actual "Sample quantiles." This good fit between expected and actual quantiles is reflected in a correlation coefficient of 0.987 – almost a perfect 1 (you'll find more about correlation in the phonetics chapter, Chapter 3).

Contrast this excellent fit with the one between the normal distribution and the 2001 data (Figure 1.10). Here we see that most of the data points in the 2001 data set are just where we would expect them to be in a normal distribution. However the two (possibly three) largest VOT values are much larger than expected. Consequently, the correlation between expected and observed quantiles for this data set (r = 0.87) is lower than what we found for the 1971 data. It may be that this distribution would look more normal if we collected more data

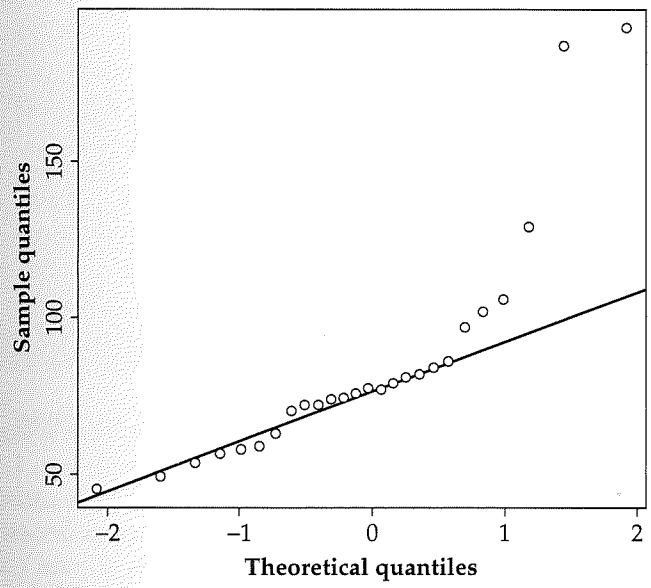

Figure 1.10 The quantiles-quantiles probability plot comparing the Cherokee 2001 data with the normal distribution.

points, or we might find that we have a bimodal distribution such that most data comes from a peak around 70 ms, but there are some VOTs (perhaps in a different speaking style?) that center around a much longer (190 ms) VOT value. We will eventually be testing the hypothesis that this speaker's VOT was shorter in 2001 than it was in 1971 and the outlying data values work against this hypothesis. But, even though these two very long VOT values are inconvenient, there is no valid reason to remove them from the data set (they are not errors of measurement, or speech dysfluencies), so we will keep them.

R note. Making a quantile-quantile plot in R is easy using the `qqnorm()` and `qqline()` functions. The function `qqnorm()` takes a vector of values (the data set) as input and draws a Q-Q plot of the data. I also captured the values used to plot the *x*-axis of the graph into the vector `vot71.qq` for later use in the correlation function `cor()`. `qqline()` adds the 45-degree reference line to the plot,

and cor() measures how well the points fit on the line (0 for no fit at all and 1 for a perfect fit).

```
vot71.qq = qqnorm(vot71)$x # make the quantile/quantile plot
vot01.qq = qqnorm(vot01)$x # and keep the x axis of the plot

qqline(vot71) # put the line on the plot
cor(vot71,vot71.qq) # compute the correlation
[1] 0.9868212
> cor(vot01,vot01.qq)
[1] 0.8700187
```

Now, let's look at a non-normal distribution. We have some rating data that are measured as proportions on a scale from 0 to 1, and in one particular condition several of the participants gave ratings that were very close to the bottom of the scale – near zero. So, when we plot these data in a quantile-quantile probability plot (Figure 1.11), you

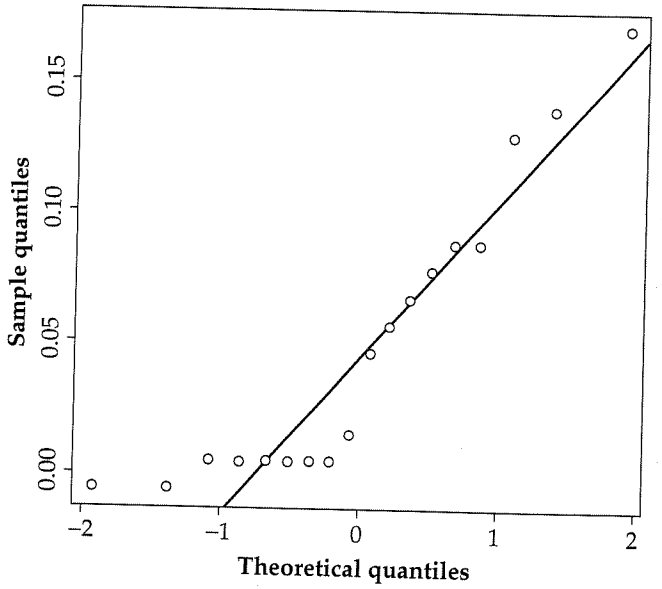

Figure 1.11 The Normal quantile-quantile plot for a set of data that is not normal because the score values (which are probabilities) cannot be less than zero.

can see that, as the sample quantile values approach zero, the data points fall on a horizontal line. Even with this non-normal distribution, though, the correlation between the expected normal distribution and the observed data points is pretty high (r = 0.92).

One standard method that is used to make a data set fall on a more normal distribution is to transform the data from the original measurement scale and put it on a scale that is stretched or compressed in helpful ways. For example, when the data are proportions it is usually recommended that they be transformed with the arcsine transform. This takes the original data x and converts it to the transformed data y using the following formula:

$$y = \frac{2}{\pi}\arcsin(\sqrt{x}) \quad \text{arcsine transformation}$$

This produces the transformation shown in Figure 1.12, in which values that are near 0 or 1 on the x-axis are spread out on the y-axis.

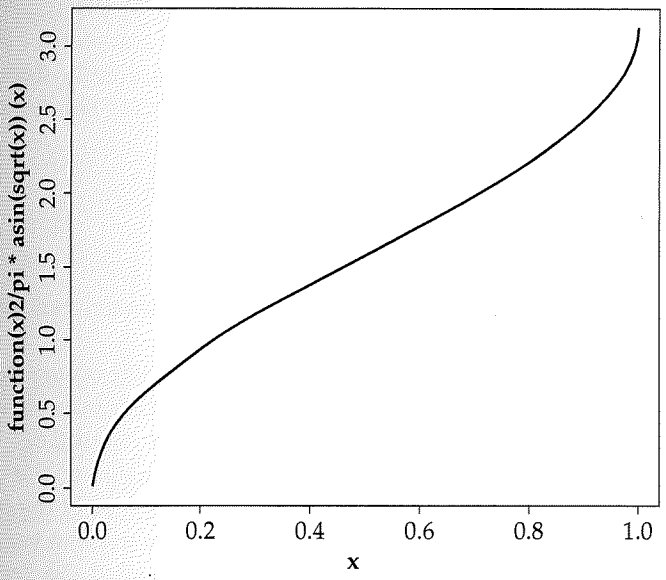

Figure 1.12 The arcsine transformation. Values of x that are near 0 or 1 are stretched out on the arcsine axis. Note that the transformed variable spans a range from 0 to π.

R note. The command for plotting the arcsine transformation in Figure 1.12 uses the plot function method.

```
> plot(function(x)2/pi*asin(sqrt(x)),0,1)
```

And as in this plot command, the command to transform the original data set (the vector "data") also uses the functions `asin()` and `sqrt()` to implement the arcsine and square root operations to create the new vector of values "x.arcsin." Read this as "x.arcsin is produced by taking the arcsine of the square root of data and multiplying it by 2 divided by π."

```
> x.arcsin = 2/pi*asin(sqrt(x))
```

The correlation between the expected values from a normal frequency distribution and the actual data values on the arcsine transformed measurement scale (r = 0.96) is higher than it was for the untransformed data. The better fit of the normal distribution to the observed data values is also apparent in the normal Q-Q plot of the transformed data (Figure 1.13). This indicates that the arcsine transform did what we needed it to do – it made our data more normally distributed so that we can use statistics that assume that the data fall in a normal distribution.

1.6 Measures of Central Tendency

Figure 1.14 shows three measures of the the central tendency, or mid-point, of a skewed distribution of data.

The *mode* of the distribution is the most frequently occurring value in the distribution – the tip of the frequency distribution. For the skewed distribution in Figure 1.14, the mode is at about 0.6.

Imagine ordering a data set from the smallest value to the largest. The *median* of the distribution is the value in the middle of the ordered list. There are as many data points greater than the median value then are less than the median. This is sometimes also called the "center of gravity".

The *mean* value, or the arithmetic average, is the least squares estimate of central tendency. First, how to calculate the mean – sum the data values and then divide by the number of values in the data set.

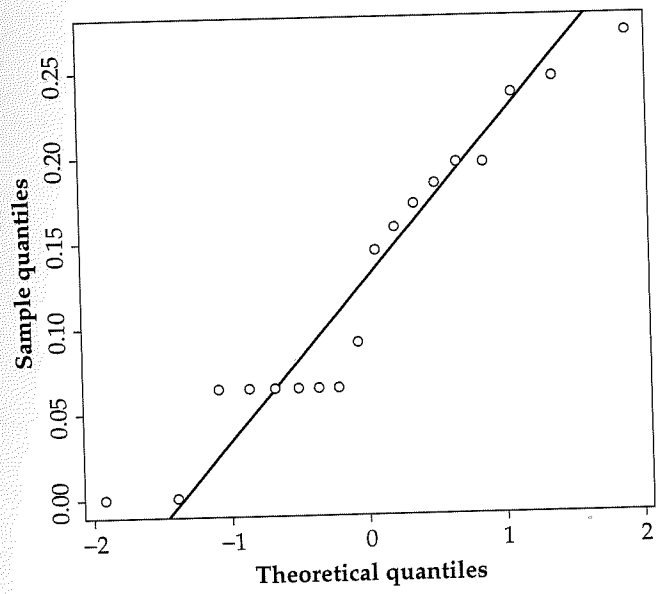

Figure 1.13 The normal quantile-quantile plot for the arcsine transform of the data shown in Figure 1.11.

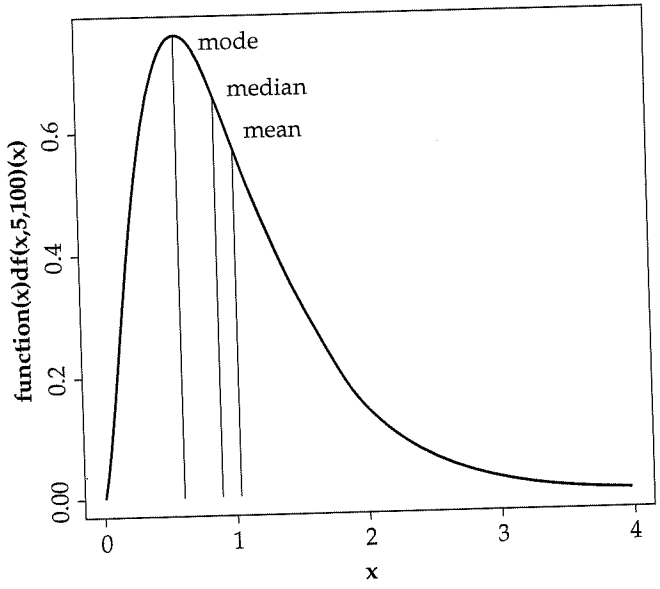

Figure 1.14 The mode, median, and mean of a skewed distribution.

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n} \quad \text{mean}$$

Second, what does it mean to be the least squares estimate of central tendency? This means that if we take the difference between the mean and each value in our data set, square these differences and add them up, we will have a smaller value than if we were to do the same thing with the median or any other estimate of the "mid-point" of the data set.

$$d^2 = \sum_{i=0}^{n} (x_i - \bar{x})^2 \quad \text{sum of the squared deviations (also called SS)}$$

So, in the data set illustrated in Figure 1.14, the value of d^2, the sum of the squared deviations from the mean, is 4,570, but if we calculate the sum of squared deviations from the median value we get a d^2 value of 4,794. This property, being the least squares estimate of central tendency, is a very useful one for the derivation of statistical tests of significance.

I should also note that I used a skewed distribution to show how the mode, median, and mean differ from each other because with a normal distribution these three measures of central tendency give the same value.

R note. The skewed distribution in Figure 1.14 comes from the "F" family of probability density distributions and is drawn in R using the `df()` density of "F" function.

```
plot(function(x)df(x,5,100),0,4,main="Measures of central
tendency")
```

The vertical lines were drawn with the `lines()` command. I used the `df()` function again to decide how tall to draw the lines and I used the `mean()` and `median()` commands with a data set drawn from this distribution to determine where on the x-axis to draw the mean and median lines.

```
lines(x = c(0.6,0.6), y = c(0,df(0.6,5,100)))
skew.data <- rf(10000,5,100)
lines(
    x = c(mean(skew.data),mean(skew.data)),
    y = c(0,df(mean(skew.data),5,100)))
lines(
    x = c(median(skew.data),median(skew.data)),
    y = c(0,df(median(skew.data),5,100)))
```


And finally, the text labels were added with the `text()` graphics command. I tried a couple of different x,y locations for each label before deciding on these.

```
text(1,0.75,labels="mode")
text(1.3,0.67,labels="median")
text(1.35,0.6,labels="mean")
```

Oh, and you might be interested in how I got the squared deviation d^2 values above. This illustrates how neatly you can do math in R. To square the difference between the mean and each data value in the vector I put the expression for the difference in `()` and then `^2` to square the differences. These then go inside the `sum()` function to add them up over the entire data vector. Your results will differ slightly because `skew.data` is a random sample from the F distribution and your random sample will be different from mine.

```
> sum((mean(skew.data)-skew.data)^2)
[1] 4570.231
> sum((median(skew.data)-skew.data)^2)
[1] 4794.141
```

We should probably also say something about the *weighted mean*. Suppose you asked someone to rate the grammaticality of a set of sentences, but you also let the person rate their ratings, to say that they feel very sure or not very sure at all about the rating given. These confidence values could be used as weights (w_i) in calculating the central tendency of the ratings, so that ratings given with high confidence influence the measure more than ratings given with a sense of confusion.

$$\bar{x} = \frac{\sum_{i=0}^{n} w_i x_i}{\sum_{i=0}^{n} w_i} \quad \text{weighted mean}$$

1.7 Measures of Dispersion

In addition to wanting to know the central point or most typical value in the data set we usually want to also know how closely clustered the data are around this central point – how dispersed are the data values away from the center of the distribution? The minimum possible amount of dispersion is the case in which every measurement has the same value. In this case there is no variation. I'm not sure what the maximum of variation would be.

A simple, but not very useful measure of dispersion is the *range* of the data values. This is the difference between the maximum and minimum values in the data set. The disadvantages of the range as a statistic are that (1) it is based on only two observations, so it may be sensitive to how lucky we were with the tails of the sampling distribution, and (2) range is undefined for most theoretical distributions like the normal distribution which extend to infinity.

I don't know of any measures of dispersion that use the median – the remaining measures discussed here refer to dispersion around the mean.

The average deviation, or the *mean absolute deviation*, measures the absolute difference between the mean and each observation. We take the absolute difference because if we took raw differences we would be adding positive and negative values for a sum of about zero no matter how dispersed the data are. This measure of deviation is not as well defined as is the standard deviation, partly because the mean is the least squares estimator of central tendency – so a measure of deviation that uses squared deviations is more comparable to the mean.

Variance is like the mean absolute deviation except that we square the deviations before averaging them. We have definitions for variance of a population and for a sample drawn from a larger population.

$$\sigma^2 = \sum(x_i - \mu)^2/N \qquad \text{population variance}$$

$$s^2 = \sum(x_i - \bar{x})^2/(n-1) \qquad \text{sample variance}$$

Notice that this formula uses the Sum of Squares (SS, also called d^2 above, the sum of squared deviations from the mean) and by dividing by N or $n - 1$, we get the Mean Squares (MS, also called s^2 here). We will see these names (SS, and MS) when we discuss the ANOVA later.

We take $(n - 1)$ as the denominator in the definition of s^2, sample variance, because $\bar{x}$ is not μ. The sample mean $\bar{x}$ is only an estimate of μ, derived from the x_i, so in trying to measure variance we have to keep in mind that our estimate of the central tendency $\bar{x}$ is probably wrong to a certain extent. We take this into account by giving up a "degree of freedom" in the sample formula. Degree of freedom is a measure of how much precision an estimate of variation has. Of course this is primarily related to the number of observations that serve as the basis for the estimate, but as a general rule the degrees of freedom decrease as we estimate more parameters with the same data set – here estimating both the mean and the variance with the set of observations x_i.

The variance is the average squared deviation – the units are squared – to get back to the original unit of measure we take the square root of the variance.

$$\sigma = \sqrt{\sigma^2} \qquad \text{population standard deviation}$$

$$s = \sqrt{s^2} \qquad \text{sample standard deviation}$$

This is the same as the value known as the RMS (root mean square), a measure of deviation used in acoustic phonetics (among other disciplines).

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}} \qquad \text{RMS = sample standard deviation}$$

1.8 Standard Deviation of the Normal Distribution

If you consider the formula for the normal distribution again, you will note that it can be defined for any mean value μ, and any standard deviation σ. However, I mentioned that this distribution is used to calculate probabilities, where the total area under the curve is equal to 1, so the area under any portion of the curve is equal to some proportion of 1. This is the case when the mean of the bell-shaped distribution is 0 and the standard deviation is 1. This is sometimes abbreviated as N(0,1) – a normal curve with mean 0 and standard deviation 1.

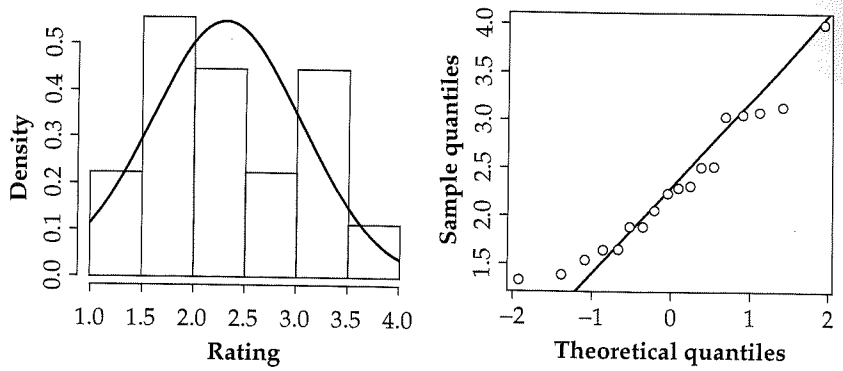

Figure 1.15 Histogram and Q-Q plot of some sample rating data.

$f_x = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ the normal distribution: N(0,1)

I would like to make two points about this.

First, because the area under the normal distribution curve is 1, we can state the probability (area under the curve) of finding a value larger than any value of x, smaller than any value of x, or between any two values of x.

Second, because we can often approximate our data with a normal distribution we can state such probabilities for our data given the mean and standard deviation.

Let's take an example of this from some rating data (Figure 1.15). Listeners were asked to rate how similar two sounds were on a scale from 1 to 5 and their average ratings for a particular condition in the experiment ("How different do [d] and [ɾ] sound?") will be analyzed here. Though the histogram doesn't look like a smooth normal curve (there are only 18 data points in the set), the Q-Q plot does reveal that the individual data points do follow the normal curve pretty well (r = 0.97). Now, how likely is it, given these data and the normal curve that they fall on, that an average rating of less than 1.5 would be given? The area to the left of 1.5 under the normal curve in the histogram plot is 0.134, so we can say that 13% of the distribution covers rating values less than 1.5, so that if we are drawing more average rating values from our population – ratings given by speakers of Latin

American Spanish – we could predict that 13% of them would have average ratings less than 1.5.

R note. Figure 1.15 comes from the following commands (assuming a vector of data). This is all pretty familiar by now.

```
par(mfrow = c(1,2))
hist(data, freq=F)
plot(function(x)dnorm(x,mean=mean(data),sd=sd(data)),1,4,
add=T)
qqnorm(data)
qqline(data)
```

I calculated the probability of a rating value less than 1.5 by calling the `pnorm()` function.

```
pnorm(1.5,mean=mean(data),sd=sd(data))
[1] 0.1340552
```

`Pnorm()` also gives the probability of a rating value greater than 3.5 but this time specifying that we want the probability of the upper tail of the distribution (values greater than 3.5).

```
pnorm(3.5,mean=mean(data),sd=sd(data),lower.tail=F)
[1] 0.05107909
```

How does this work? We can relate the frequency distribution of our data to the normal distribution because we know the mean and standard deviation of both. The key is to be able to express any value in a data set in terms of its distance in standard deviations from the mean.

For example, in these rating data the mean is 2.3 and the standard deviation is 0.7. Therefore, a rating of 3 is one standard deviation above the mean, and a rating of 1.6 is one standard deviation below the mean. This way of expressing data values, in standard deviation units, puts our data on the normal distribution – where the mean is 0 and the standard deviation is 1.

I'm talking about *standardizing* a data set – converting the data values into *z-scores*, where each data value is replaced by the distance between it and the sample mean where the distance is measured as the number of standard deviations between the data value and the mean. As a result of "standardizing" the data, z-scores always have a mean

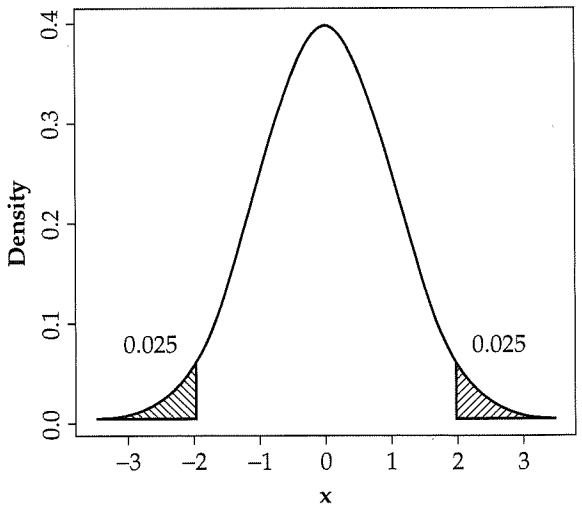

Figure 1.16 95% of the area under the normal distribution lies between $-1.96s$ and $1.96s$. 97.5% is above $-1.96s$ and 97.5% is less than $1.96s$.

of 0 and a standard deviation of 1, just like the normal distribution. Here's the formula for standardizing your data:

$$z_i = \frac{x_i - \bar{x}}{s} \quad z\text{-score standardization}$$

With standardized values we can easily make probability statements. For example, as illustrated in Figure 1.16, the area under the normal curve between −1.96 and 1.96 is 0.95. This means that 95% of the values we draw from a normal distribution will be between 1.96 standard deviations below the mean and 1.96 standard deviations above the mean.

EXERCISES

1 Open a dictionary of any language to a random page. Count the number of words that have 1, 2, 3, 4, etc. syllables. What kind of distribution do you get?
2 On the same page of the dictionary, count the number of instances of the different parts of speech – noun, verb, adjective, function word. What kind of variable is part of speech and why can't you draw a reasonable distribution of part of speech?
3 Calculate the average number of syllables per word on this page. You can do this as a weighted mean, using the count as the weight for each syllable length.
4 What is the standard deviation of the average length in syllables? How do you calculate this? Hint: The raw data have one observation per word, while the count data have several words summarized for each syllable length.
5 Are these data an accurate representation of word length in this language? How could you get a more accurate estimate?
6 Using your word length data from question 1 above, produce a quantile-quantile (Q-Q) plot of the data. Are these data approximately normally distributed? What is the correlation between the normal curve quantiles (the theoretical quantiles) and the observed data?
7 Make a histogram of the data. Do these data seem to follow a normal distribution? Hint: Plotting the syllable numbers on the x-axis and the word counts on the y-axis (like Figure 1.1) may be a good way to see the frequency distribution.
8 Assuming a normal distribution, what is the probability that a word will have more than three syllables? How does this relate to the observed percentage of words that have more than three syllables in your data?

2 Patterns and Tests

In this chapter, I will present two key strategies in the quantitative analysis of linguistic data. We will come back to these in several different contexts in later chapters, so this is supposed to provide a foundation for those later discussions of how to apply hypothesis testing and regression analysis to data.

2.1 Sampling

But first I would like to say something about sampling. In Chapter 1, I made the distinction between a population parameter (Greek letter symbols like μ, σ, σ^2) and sample statistics (Roman letter symbols like $\bar{x}$, s, s^2). These differ like this: If we take the average height of everyone in the room, then the mean value that we come up with is the population parameter μ, of the population "everyone in the room." But if we would like to think that this group of people is representative of a larger group like "everyone at this university" or "everyone in this town," then our measured mean value is a sample statistic $\bar{x}$ that may or may not be a good estimate of the larger population mean.

In the normal course of events as we study language, we rely on samples to represent larger populations. It isn't practical to directly measure a population parameter. Imagine trying to find the grammaticality of a sentence from everyone who speaks a language! So we take a small, and we hope, representative sample from the population of ultimate interest.

So, what makes a good sample? To be an adequate representation of a population, the sample should be (1) large enough, and (2) random. Small samples are too sensitive to the effects of the occasional "odd" value, and nonrandom samples are likely to have some bias (called sampling bias) in them.

To be random it must be the case that every member of the population under study has an equal chance of being included in the sample. Here are two ways in which our linguistic samples are usually nonrandom.

1. We limit participation in our research to only certain people. For example, a consultant must be bilingual in a language that the linguist knows, college students are convenient for our listening experiments, we design questionnaires and thereby require our participants to be literate.
2. We observe linguistic performance only in certain restricted contexts. For example, we make tape recordings while people are reading a list of words or sentences. We ask for sentence judgments of sentences in a particular order on a questionnaire.

Obviously, it is pretty easy to violate the maxims of good sampling, but what should you do if your sample isn't representative of the population that you would most like to study? One option is to try harder to find a way to get a more random, representative sample. For instance you might collect some data from monolingual speakers and compare this with your data drawn from bilingual speakers. Or you might try conducting a telephone survey, using the listing of people in the phone book as your "population." And to address the context issue, you might try asking people meaningful questions in a natural context, so that they don't know that you are observing their speech. Or you might simply reverse the order of your list of sentences on the questionnaire.

In sum, there is a tradeoff between the feasibility of research and the adequacy of the sample. We have to balance huge studies that address tiny questions against small studies that cover a wider range of interesting issues. A useful strategy for the discipline is probably to encourage a certain amount of "calibration" research that answers limited questions with better sampling.

2.2 Data

Some of this discussion may reveal that I have a particular attitude about what linguistic data are, and I think this attitude is not all that unusual but worth stating explicitly. The data in linguistics are any observations about language. So, I could observe people as they speak or as they listen to language, and call this a type of linguistic data. Additionally, a count of forms used in a text, whether it be modern newspaper corpora or ancient carvings, is data. I guess you could say that these are observations of people in the act of writing language and we could also observe people in the act of reading language as well. Finally, I think that when you ask a person directly about language, their answers are linguistic data. This includes native speaker judgments, perceptual judgments about sounds, and language consultants' answers to questions like "what is your word for finger?"

Let's consider an observation and some of its variables.

The observation is this: A native speaker of American English judges the grammaticality of the sentence "Josie didn't owe nobody nothing" to be a 3 on a 7-point scale.

There are a large number of variables associated with this observation. For example, there are some static properties of the person who provided the judgment – gender, age, dialect, socioeconomic status, size of vocabulary, linguistic training. Additionally, aspects of the situation in which the judgment occurs may influence the participant. One common factor is what prior judgments were given already in this session. Perhaps we can't try all possible orderings of the sentences that we want to test, but we should pay attention to the possibility that order matters. Additionally, the person's prior experience in judging sentences probably matters. I've heard syntacticians talk about how their judgments seem to evolve over time and sometimes reflect theoretical commitments.

The task given to the participant may also influence the type of answer we get. For example, we may find that a fine-grained judgment task provides greater separation of close cases, or we may find that variance goes up with a fine-grained judgment task because the participant tends to focus on the task instead of on the sentences being presented.

We may also try to influence the participant's performance by instructing them to pay particular attention to some aspect of the stimuli or approach the task in a particular way. I've done this to no effect (Johnson, Flemming, & Wright, 1993) and to startling effect (Johnson, Strand, & D'Imperio, 1999). The participants in Johnson, Flemming, & Wright gave the same answers regardless (so it seemed) of the instructions that we gave them. But the "instruction set manipulation" in Johnson, Strand, & D'Imperio changed listeners' expectations of the talker and thus changed their performance in a listening experiment. My main point here is that how we interact with participants may influence their performance in a data collection situation.

An additional, very important task variable is the list of materials. The context in which a judgment occurs influences it greatly. So if the test sentence appears in a list that has lots of "informal" sentences of the sort that language mavens would cringe at, it may get a higher rating than if it appeared in a list of "correct" sentences.

The observation "3 on a 7-point scale" might have been different if we had changed any one of these variables. This large collection of potentially important variables is typical when we study complex human behavior, especially learned behavior like language. There are too many possible experiments. So the question we have to address is: Which variables are you interested in studying and which would you like to ignore? You have to ignore variables that probably could affect the results, and one of the most important elements of research is learning how to ignore variables.

This is a question of research methods which lies beyond the scope of this chapter. However, I do want to emphasize that (1) our ability to generalize our findings depends on having a representative sample of data – good statistical analysis can't overcome sampling inadequacy – and (2) the observations that we are exploring in linguistics are complex with many potentially important variables. The balancing act that we attempt in research is to stay aware of the complexity, but not let it keep us from seeing the big picture.

2.3 Hypothesis Testing

Now, keeping in mind the complexities in collecting representative samples of linguistic data and the complexities of the data themselves, we come to the first of the two main points of this chapter – hypothesis testing.

We often want to ask questions about mean values. Is this average voice onset time (VOT) different from that one? Do these two constructions receive different average ratings? Does one variant occur more often than another? These all boil down to the question is $\bar{x}$ (the mean of the data x_i) different from $\bar{y}$ (the mean of the data y_i)?

The smarty-pants answer is that you just look at the numbers, and they either are different or they aren't. The sample mean simply is what it is. So $\bar{x}$ is either the same number as $\bar{y}$ or it isn't. So what we are *really* interested in is the population parameter estimated by $\bar{x}$ and $\bar{y}$ – call them μ_x and μ_y. Given that we know the sample mean values $\bar{x}$ and $\bar{y}$, can we say with some degree of confidence that μ_x is different from μ_y? Since the sample mean is just an estimate of the population parameter, if we could measure the error of the sample mean then we could put a confidence value on how well it estimates μ.

2.3.1 *The central limit theorem*

A key way to approach this is to consider the sampling distribution of $\bar{x}$. Suppose we take 100 samples from a particular population. What will the distribution of the means of our 100 samples look like?

Consider sampling from a uniform distribution of the values 1 . . . 6, i.e. roll a dice. If we take samples of two (roll the dice once, write down the number shown, roll it again, and write down that number), we have 6^2 possible results, as shown in Table 2.1. Notice that the average of the two rolls is the same for cells in the diagonals. For example, the only way to average 6 is to roll 6 in both trials, but there are two ways to average 5.5 – roll a 6 and then a 5 or a 5 and then a 6. As you can see from Table 2.1, there are six ways to get an average of 3.5 on two rolls of the dice. Just to drive the point home, excuse the excess of this, the average of the following six two dice trials is 3.5 – (6,1), (5,2), (4,3), (3,4), (2,5), and (1,6). So, if we roll two dice, the probability of having an average number of dots equal to 3.5 is 6 times out of 36 trials (6 of the 36 cells in table 2.1).

In general the frequency distribution of the mean for two rolls of a dice has a shape like the normal distribution – this is shown in Figure 2.1. This is the beginning of a proof of the central limit theorem, which states that as the number of observations in each sample increases, the distribution of the means drawn from these samples tends toward the normal distribution. We can see in this simple example that even

Table 2.1 The possible outcomes of rolling a dice twice – i.e. samples of size two from a uniform distribution of the integers 1 . . . 6. The number of the first roll is indicated by the row number and the number of the second roll is indicated by the column number.

	1	**2**	**3**	**4**	**5**	**6**
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6
$\bar{x}$	4	4.5	5	5.5	6	

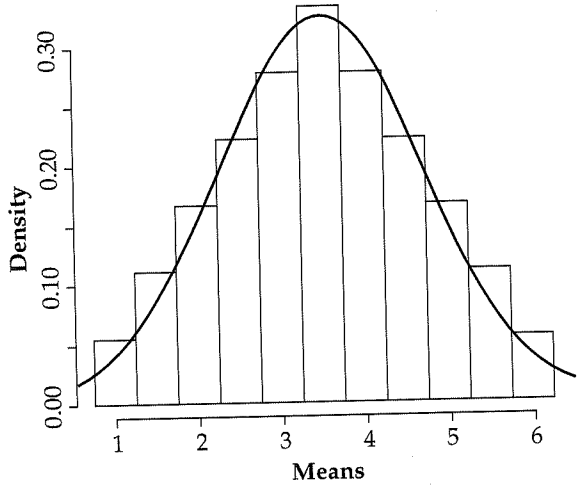

Figure 2.1 The frequency distribution of the mean for the samples illustrated in Table 2.1.

though the observations (dice throws) come from a uniform distribution in which each number on the dice has an equal probability of 1/6, the distribution of the means of just two observations looks remarkably like a normal distribution.

R note. Here's how I made Figure 2.1. First I entered a vector "means" that lists the mean value of each cell in Table 2.1. Then I entered a vector "b" to mark the edges of the bars that I want in the histogram. Then I plotted the histogram and best-fitting normal curve.

```
means = c(1, 1.5,1.5, 2,2,2, 2.5,2.5,2.5,2.5, 3,3,3,3,3,
3.5,3.5,3.5,3.5,3.5,3.5, 4,4,4,4,4, 4.5,4.5,4.5,4.5,
5,5,5, 5.5,5.5, 6)
b = c(0.75,1.25,1.75,2.25,2.75,3.25,3.75,4.25,4.75,
5.25,5.75,6.25)
hist(means,breaks=b,freq=F)
plot(function(x)dnorm(x,mean=mean(means), sd=sd(means)),
0.5,6.5, add=T)
```

Before we continue with this discussion of the central limit theorem we need step aside slightly to address one question that arises when you look at Figure 2.1. To the discriminating observer, the vertical axis doesn't seem right. The probability of averaging 3.5 dots on two rolls of the dice is 6/36 = 0.1666. So, why does the vertical axis in Figure 2.1 go up to 0.3? What does it mean to be labelled "density" and how is probability density different from probability?

Consider the probability of getting exactly some particular value on a continuous measurement scale. For example, if we measure the amount of time it takes someone to respond to a sound, we typically measure to some chosen degree of accuracy – typically the nearest millisecond. However, in theory we could have produced an arbitrarily precise measurement to the nanosecond and beyond. So, on a continuous measurement scale that permits arbitrarily precise values, the probability of finding exactly one particular value, say exactly 500 ms, is actually zero because we can always specify some greater degree of precision that will keep our observation from being exactly 500 ms – 500.00000001 ms. So on a continuous dimension, we can't give a probability for a specific value of the measurement variable. Instead we can only state the probability of a region under the cumulative distribution curve. For instance, we can't say what the probability of a measurement of 500 is, but we can say for example that about 16% of the cumulative distribution in Figure 2.2 falls to the left of 500 ms – that given a population like this one (mean = 600, standard deviation

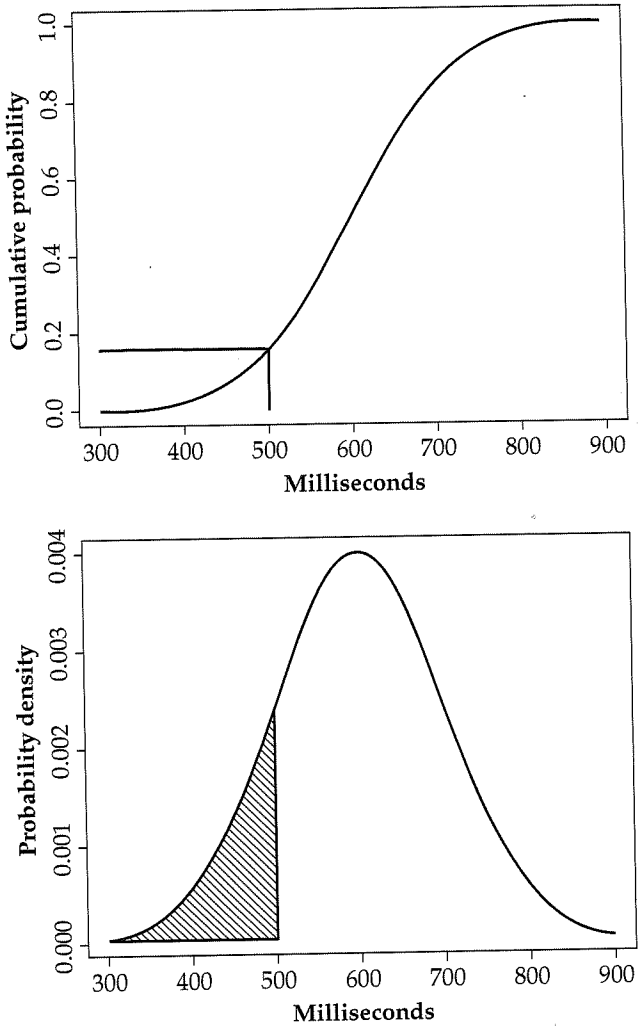

Figure 2.2 The cumulative probability function of the normal curve steadily increases from 0 to 1, while the probability density function (pdf) takes the familiar bell-shaped curve.

= 100) we expect 16% of the values in a representative sample to be lower than 500.

The probability density curve in the bottom panel of Figure 2.2 shows this same point, but in terms of the area under the curve instead of the

value of the function at a particular point. In the probability density function, the area under the curve from 0 to 500 ms is 16% of the total area under the curve, so the value of the cumulative density function at 500 ms is 0.16. This relationship is illustrated in figure 2.2.

The probability density that we see in Figures 2.1 and 2.2 indicates the amount of change in (the derivative of) the cumulative probability function. If $f(x)$ is the probability density function, and $F(x)$ is the cumulative probability function then the relationship is:

$$\frac{d}{dx}F(x) = f(x),$$ the density function from the cumulative probability function

The upshot is that we can't expect the density function to have the values on the y-axis that we would expect for a cumulative frequency curve, and what we get by going to the trouble of defining a probability density function in this way is a method for calculating probability for areas under the normal probability density function.

Let's return now to the main point of Figure 2.1. We have an equal probability in any *particular trial* of rolling any one of the numbers on the dice – a *uniform distribution* – but the frequency distribution of the *sample mean*, even of a small sample size of only two rolls, follows the *normal distribution*. This is only approximately true with an n of 2. As we take samples of larger and larger n the distribution of the means of those samples becomes more and more perfectly normal. In looking at this example of two rolls of the dice, I was struck by how normal the distribution of means is for such small samples. This property of average values – that they tend to fall in a normal distribution as n increases – is called the *central limit theorem*. The practical consequence of the central limit theorem is that we can use the normal distribution (or, as we will see, a close approximation) to make probability statements about the mean – like we did with z-scores – even though the population distribution is not normal.

Let's consider another example this time of a skewed distribution. To produce the left side of Figure 2.3, I started with a skewed population distribution as shown in the figure and took 1,000 random samples from the distribution with a sample size of 10 data points per sample. I calculated the mean of each of the 1,000 samples so that now I have a set of 1,000 means. These are plotted in a histogram and theoretical curve of the histogram that indicate that the frequency distribution of the mean is a normal distribution. Also, a Q-Q plot (not shown) of these

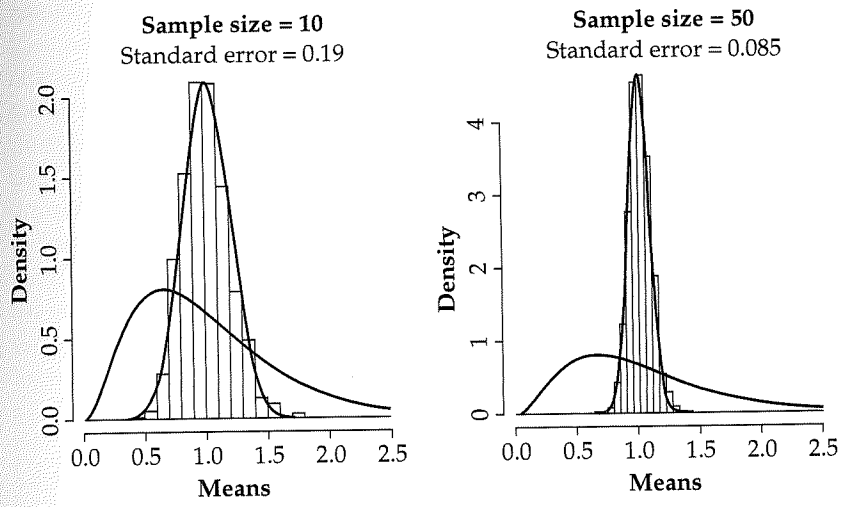

Figure 2.3 The sampling distribution of the mean taken from 1,000 samples that were drawn from a skewed population distribution when the sample size was only 10 observations, and when the sample size was 50 observations.

has a correlation of 0.997 between the normal distribution and the sampling distribution of the mean.

The panel on the right side of Figure 2.3 shows a similar situation except that each of the 1,000 samples had 50 observations in it instead of 10. Again the frequency distribution of the mean is normal (the Q-Q correlation was 0.998) and the fit to the normal curve is a little better than it was when we took samples of size 10 from this skewed distribution.

Notice in Figure 2.3 that I report the "standard error" of the 1,000 means in each panel. By standard error I mean simply the standard deviation of the sample of 1,000 means. As is apparent from the pictures, the standard deviation of the means is smaller when we take samples of 50 than when we take samples of 10 from the skewed distribution. This is a general feature. We are able to get a more accurate estimate of the population mean with a larger sample. You can see that this is so if we were to limit our samples to only one observation each. In such a case, the standard error of the mean would be the same as the standard deviation of the population being sampled. With larger samples the effects of observations from the tails of the distribution

are dampened by the more frequently occurring observations from the middle of the distribution.

In fact, this value, the standard deviation of a sample of means, tells us just that – how accurately we can measure the mean of our raw data population. If you take a bunch of different samples and find that their means are really different from each other, then you have to wonder how accurate any one particular mean is when it comes to estimating the true mean of the population. Clearly one factor that matters is the size of the sample. The means drawn from samples of size 10 were much more spread out than were the means drawn from samples of size 50. Next we'll look at another factor that determines how accurately we can measure the population mean from samples drawn from that population.

R note. To explore the central limit theorem I wrote a function in R called `central.limit()`. Then to produce the graphs shown in Figure 2.3 all I had to do was type:

```
source("central.limit")
par(mfrow=c(1,2)) # to have one row with two graphs
central.limit(10) # to make the first graph
central.limit(50) # to make the second graph
```

I also made it so we can look at a Q-Q plot of the distribution of the means by changing the function call slightly.

```
central.limit(10,qq=TRUE)
```

I put a little effort into this `central.limit()` function so that I could change the shape of the population distribution, change the sample size, and the number of samples drawn, and to add some information to the output. Here's the definition of `central.limit()` that I stored in a text file called "central.limit" and that is read into R with the `source()` command above.

```
#--------- central.limit ---------------
#
# The input parameters are:
#       n - size of each sample
#       m - number of samples of size n to select
#       qq - TRUE means show the Q-Q plots
#       df1, df2 - the df of the F() distribution from which
#       samples are drawn
#       xlow, xhigh - x-axis limits for plots
central.limit = function(n=15,m=1000,qq=FALSE,
df1=6,df2=200,xlow=0,xhigh=2.5) {
    means = vector() # I hereby declare that "means" is a
#   vector
    for (i in 1:m) { # get m samples from a skewed
        distribution
        data= rf(n,df1,df2) # the F() distribution is
#       nice and skewed
        means[i] = mean(data) # means is our array of means
    }
    if (qq) { # call with TRUE and it makes the q-q plots
        x=qqnorm(means)$x
        qqline(means)
        caption = paste("n =",n,", Correlation = ",
            signif(cor(means,x),3))
        mtext(caption)
    } else { # the default behavior
        title = paste("Sample size = ",n)
        hist(means, xlim = c(xlow,xhigh),main=title,
            freq=F)
        plot(function(x)dnorm(x,mean=mean(means),
            sd=sd(means)), xlow, xhigh, add=T)
        plot(function(x)df(x,df1,df2),xlow,xhigh,add=T)
        caption = paste("Standard error = ",
            signif(sd(means),3))
        mtext(caption)
    }
}
```

This looks pretty complicated, I know, but I really think it is worth knowing how to put together an R function because (1) being able to write functions is one of the strengths of R, and (2) having a function that does exactly what you want done is extremely valuable. I do a number of new things in this function, but I also do a number of things that we saw in Chapter 1. For

example, we saw earlier how to sample randomly from the F-distribution using the `rf()` command, and we saw how to plot a histogram together with a normal distribution curve. We also saw earlier how to draw Q-Q plots. So really, this function just puts together a number of things that we already know how to do.

There are only three new R capabilities that I used in writing the `central.limit()` function:

`function()`. The first is that `central.limit()` shows how to create a new command in R using the `function()` command. For example, this line:

```
>square = function(x) {return(x*x)}
```

creates, or defines, a function called `square()` that returns the square (x times x) of the input value. So now if you enter:

```
->square(1532)
```

R will return with the number "2347024". Note that in my definition of "central.limit()" I gave each input parameter a default value. This way the user doesn't have to enter values for these parameters, but can instead pick and choose which variable names to set manually. Because I set default values for the input variables in the central limit function, `central.limit(df1=3)` is a legal command that works so that the default values are used for each input parameter except `df1`. You can save function definitions in text files and use the `source()` command to read them into R and make them available for use.

`for (i...)`. The `central.limit()` function shows how to use a "for loop" to do something over and over. I wanted to draw lots (m) of random samples from the skewed F-distribution, store all of the means of these samples and then make a histogram of the means. So, I used a "for loop" to execute two commands over and over:

```
data= rf(n,df1,df2)
means[i] = mean(data)
```

The first one draws a new random sample of n data points from the F-distribution, and the second one calculates the mean of this

sample and stores it in the ith location in the "means" vector. By putting these two commands inside the following lines we indicate that we want the variable i to start with a value of 1 and we want to repeat the commands bracketed by { . . . } counting up from $i = 1$ until $i = m$. This gives us a vector of m mean values stored in means[1] . . . means[m].

```
for (i in 1:m) {
    # my repeated stuff here
}
```

`if ()`. Finally, `central.limit()` shows how to use an "if statement" to choose to do one set of commands or another. I wanted to have the option to look at a histogram and report the standard deviation, or to look at a Q-Q plot and report the correlation with the normal distribution. So, I had `central.limit()` choose one of these two options depending on whether the input parameter "qq" is TRUE or FALSE.

```
if (qq) {
    # do this if "qq" is TRUE
  else {
    # do this if "qq" is not TRUE
}
```

I made a new version of `central.limit()` that I decided to call `central.limit.norm()` because instead of sampling from a skewed population distribution, I sampled from a normal distribution. In all other regards `central.limit.norm()` is exactly like `central.limit()`. Figure 2.4 shows the population distribution (notice that now it is a normal distribution), a histogram of 5,000 means drawn from the population, and a normal curve fit to the histogram of means. The panels on the left had a population standard deviation (σ) of 1, while the panels on the right had a σ of 0.6. The top panels had a sample size (n) of 15, while for the bottom panels $n = 50$.

We saw in Figure 2.3 that our estimate of μ was more accurate (the standard deviation of $\bar{x}$ was smaller) when the sample size increased. What we see in Figure 2.4 is that the error of our estimates of μ are also smaller when the standard deviation of the population (σ) is smaller. In fact, figure 2.4 makes it clear that the standard error of

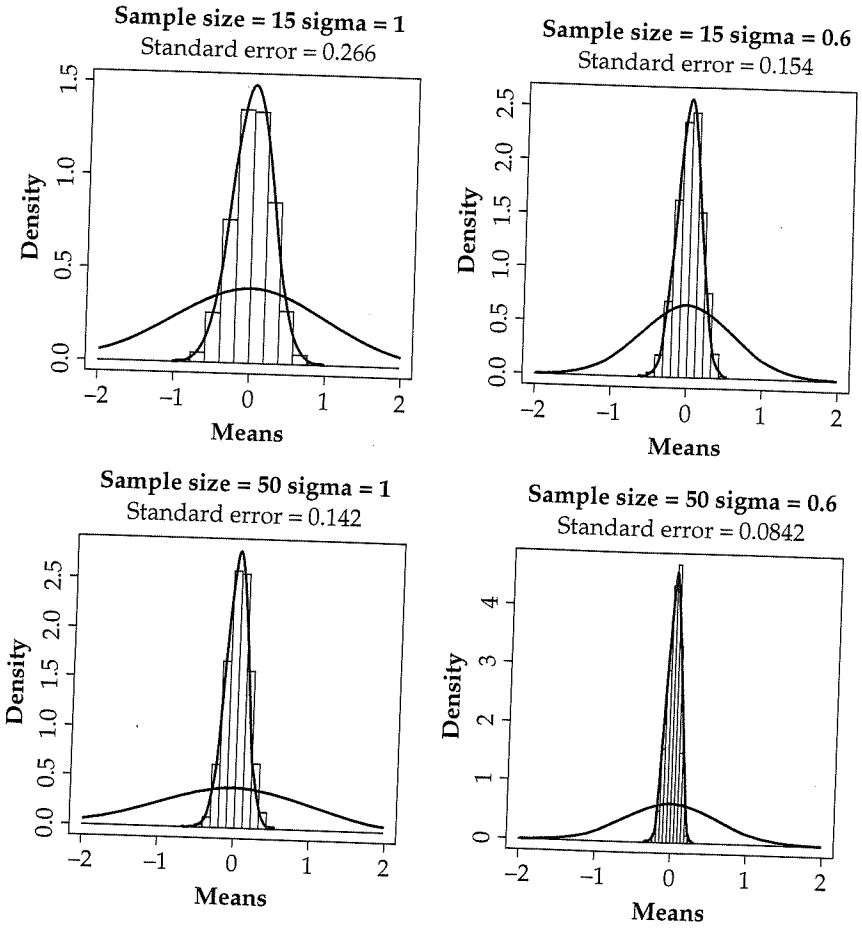

Figure 2.4 The sampling distribution of the mean of a normal distribution is a function of sample size and the population standard deviation (σ). Right panels: the population distribution has a σ of 1. Left panels: σ is 0.6. Top panels: Sample size (n) was 15. Bottom panels: n is 50.

the mean depends on both sample size and σ. So, we saw earlier in Figure 2.3 that the distribution of sample means $\bar{x}$ is more tightly focused around the true population mean μ when the sample n is larger. What we see in Figure 2.4 is that the distribution of sample means is also more tightly focused around the true mean when the population distribution is smaller.

Check this out. Let's abbreviate the standard error of the mean to SE – this is the standard deviation of $\bar{x}$ values that we calculate from successive samples from a population and it indicates how accurately we can estimate the population mean μ from a random sample of data drawn from that population. It turns out (as you might expect from the relationships apparent in Figure 2.4) that you can measure the standard error of the mean from a single sample – it isn't necessary to take thousands of samples and measure SE directly from the distribution of means. This is a good thing. Can you imagine having to perform every experiment 1,000 times so you can measure SE directly? The relationship between SE and σ (or our sample estimate of σ, s) is:

$$SE = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{standard error of the mean: population}$$

$$SE = s_{\bar{x}} = \frac{s_x}{\sqrt{n}} \quad \text{standard error of the mean: sample}$$

You can test this out on the values shown in Figure 2.4:

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{15}} = 0.258 \qquad \frac{\sigma}{\sqrt{n}} = \frac{0.6}{\sqrt{15}} = 0.155$$

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{50}} = 0.141 \qquad \frac{\sigma}{\sqrt{n}} = \frac{0.6}{\sqrt{50}} = 0.0849$$

The calculated values are almost exactly the same as the measured values of the standard deviation of the sets of 5,000 means.

2.3.2 *Score keeping*

Here's what we've got so far about how to test hypotheses regarding means.

1 You can make probability statements about variables in normal distributions.
2 You can estimate the parameters of empirical distributions as the least squares estimates of $\bar{x}$ and s.
3 Means themselves, of samples drawn from a population, fall in a normal distribution.
4 You can estimate the standard error (SE) of the normal distribution of $\bar{x}$ values from a single sample.

What this means for us is that we can make probability statements about means.

2.3.3 H_0: $\mu = 100$

Recall that when we wanted to make probability statements about observations using the normal distribution, we converted our observation scores into *z*-scores (the number of standard deviations different from the mean) using the *z*-score formula.

So, now to test a hypothesis about the population mean (μ) on the basis of our sample mean and the standard error of the mean we will use a very similar approach.

$$z = \frac{x_i - \bar{x}}{s} \quad z\text{-score}$$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad t\text{-value}$$

However, we usually (almost always) don't know the population standard deviation. Instead we estimate it with the sample standard deviation, and the uncertainty introduced by using s instead of σ means that we are off a bit and can't use the normal distribution to compare $\bar{x}$ to μ. Instead, to be a little more conservative, we use a distribution (or family of distributions), called the *t*-distribution that takes into account how certain we can be about our estimate of σ. Just as we saw that a larger sample size gives us a more stable estimate of the population mean, so we get a better estimate of the population standard deviation with larger sample sizes. So the larger the sample size, the closer the *t*-distribution is to normal. I show this in Figure 2.5 for the normal distribution and *t*-distributions for three different sample sizes. So we are using a slightly different distribution to talk about mean values, but the procedure is practically the same as if we were using the normal distribution. Nice that you don't have to learn something totally new.

To make a probability statement about a *z*-score you refer to the normal distribution, and to make a probability statement about a *t*-value you refer to the *t*-distribution. It may seem odd to talk about comparing the sample mean to the population mean because we we can easily calculate the sample mean but the population mean is not a value that we can know. However, if you think of this as a way to

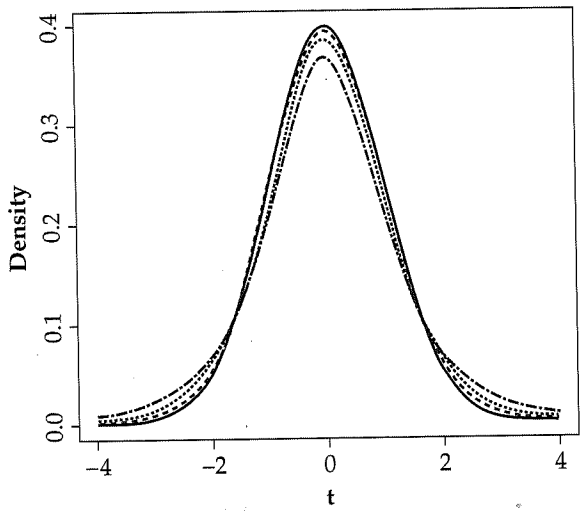

Figure 2.5 The normal distribution (solid line) and *t*-distributions for samples of size $n = 21$ (dash), $n = 7$ (dot) and $n = 3$ (dot, dash).

test a hypothesis, then we have something. For instance, with the Cherokee VOT data, where we observed that $\bar{x} = 84.7$ and $s = 36.1$ for the stops produced in 2001, we can now ask whether the population mean μ is different from 100. Let's just plug the numbers into the formula:

$$t = \frac{\bar{x}_i - \mu}{s_{\bar{x}}} = \frac{84.7 - 100}{36.1/\sqrt{26}} = \frac{-15.3}{7.08} = -2.168$$

So we can use the formula for t to find that the *t*-value in this test is −2.168. But what does that mean? We were testing the hypothesis that the average VOT value of 84.7 ms is not different from 100 ms. This can be written as H_0: $\mu = 100$. Meaning that the null hypothesis (the "no difference" hypothesis H_0) is that the population mean is 100. Recall that the statistic t is analogous to z – it measures how different the sample mean $\bar{x}$ is from the hypothesized population mean μ, as measured in units of the standard error of the mean. As we saw in Chapter 1, observations that are more than 1.96 standard deviations away from the mean in a normal distribution are pretty unlikely – only 5% of the area under the normal curve. So this *t*-value of −2.168 (a little more

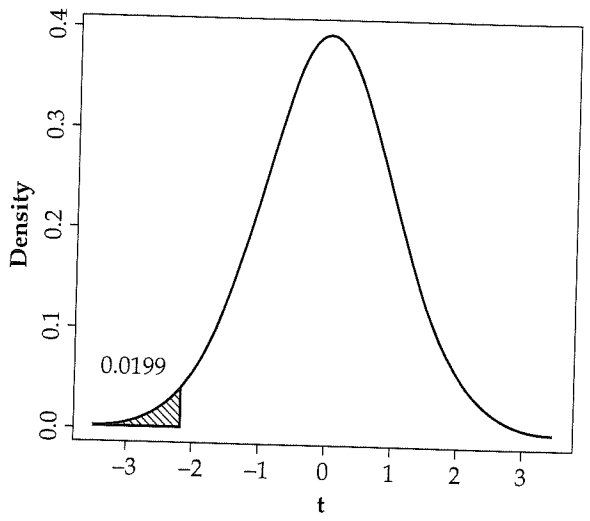

Figure 2.6 The probability density function of t with 25 degrees of freedom. The area of the shaded region at $t < -2.168$ indicates that only a little over 2% of the area under this curve has a t-value less than −2.168.

than 2 standard errors less than the hypothesized mean) might be a pretty unlikely one to find if the population mean is actually 100 ms.

How unlikely? Well, the probability density function of t with 25 degrees of freedom (since we had 26 observations in the VOT data set) shows that only 2% of all t-values in this distribution are less than −2.16 (Figure 2.6). Recall that we are evaluating the null hypothesis that $\mu = 100$. Therefore, this probability value says that if we assume that $\mu = 100$ it is pretty unlikely (2 times in 100) that we would draw a sample that has an $\bar{x}$ of 84.7. The more likely conclusion that we should draw is that the population mean is less than 100.

R note. I wrote a script to produce t probability density functions with shaded tails such as the one in Figure 2.6. The function is called "shade.tails" (check for it in the "Data sets and scripts" part of the book web page), and to produce Figure 2.6, I entered the t-value we calculated above, the degrees of freedom for the t-distribution, and indicated that we want the probability of a lower

t-value. The key functions in shade.tails are `pt()`, the probability of getting a smaller or larger t-value, and `dt()` the density functon of t. For example, the probability of finding a smaller t-value, given 25 degrees of freedom is calculated by the `pt()` function.

```
> shade.tails(2.16,tail="lower",df=25)
> pt(-2.168,25)
[1] 0.01994047
```

This hypothesis test – using the mean, standard deviation, and hypothesized population mean (μ) to calculate a t-value, and then look up the probability of the t-value – is a very common statistical test. Therefore, there is a procedure in R that does it for us.

```
> vot01 = c(84, 82, 72, 193, 129, 77, 72, 81, 45, 74, 102, 77,
187, 79, 86, 59, 74, 63, 75, 70, 106, 54, 49, 56, 58, 97)
> vot71 =c(67, 127, 79, 150, 53, 65, 75, 109, 109, 126, 129,
119, 104, 153, 124, 107, 181, 166)
> t.test(vot01,mu=100,alternative="less")

  One Sample t-test

data: vot01
t = -2.1683, df = 25, p-value = 0.01993
alternative hypothesis: true mean is less than 100
95 percent confidence interval:
  -Inf 96.74298
sample estimates:
mean of x
  84.65385
```

In this call to `t.test()`, I entered the name of the vector that contains my data, the hypothesized population mean for these data, and that I want to know how likely it is to have a lower t-value.

2.3.4 *Type I and type II error*

We seek to test the hypothesis that the true Cherokee VOT in 2001 (μ) is 100 ms by taking a sample from a larger population of possible

measurements. If the sample mean ($\bar{x}$) is different enough from 100 ms, then we reject this hypothesis; otherwise we accept it.

The question is, how different is different enough? We can quantify the difference between the sample mean and the hypothesized population mean in terms of a probability. As we saw above, if the population mean is 100 ms, then in only 2 times in 100 could we get a sample mean of 84.7 or less. Suppose that we decide then that this is a big enough difference – the probability of a sample of 84.7 mean coming from a population that has a mean of 100 ms is pretty darn low – so we reject the hypothesis that $\mu = 100$ (let's label it H_0), and instead accept the alternative hypothesis that $\mu < 100$ (call this H_1 and note that this is only one of several possible alternative hypotheses).

H_0: $\mu = 100$ Reject
H_1: $\mu < 100$ Accept

We have to admit, though, that 2 times out of 100 this decision would be wrong. It may be unlikely, but it is still possible that H_0 is correct – the population mean really could be 100 ms even though our sample mean is a good deal less than 100 ms. This error probability (0.02) is called the probability of making a type I error. A type I error is that we incorrectly reject the null hypothesis – we claim that the population mean is less than 100, when actually we were just unlucky and happened to draw one of the 2 out of 100 samples for which the sample mean was equal to or less than 84.7.

No matter what the sample mean is, you can't reject the null hypothesis with certainty because the normal distribution extends from negative infinity to positive infinity. So, even with a population mean of 100 ms we could have a really unlucky sample that has a mean of only 5 ms. This probably wouldn't happen, but it might. So we have to go with our best guess.

In practice, "going with your best guess" means choosing a type I error probability that you are willing to tolerate. Most often we are willing to accept a 1 in 20 chance that we just got an unlucky sample that leads us to make a type I error. This means that if the probability of the *t*-value that we calculate to test the hypothesis is less than 0.05, we are willing to reject H_0 ($\mu = 100$) and conclude that the sample mean comes from a population that has a mean that is less than 100 ($\mu < 100$). This criterion probability value ($p < 0.05$) is called the "alpha" (α) level of the test. The α level is the acceptable type I error rate for our hypothesis test.

Table 2.2 The decision to accept or reject the null hypothesis may be wrong in two ways. An incorrect rejection, a type I error, is when we claim that the means are different but in reality they aren't, and an incorrect acceptance, a type II error, is when we claim that the means are not different but in reality they are.

		Reality	
		H_0 is true	H_0 is false
Decision	accept H_0	correct	Type II error
	reject H_0	Type I error	correct

Where there is a type I error, there must be a type II error also (see Table 2.2). A type II error occurs when we incorrectly accept the null hypothesis. Suppose that we test the hypothesis that the average VOT for Cherokee (or at least this speaker) is 100 ms, but the actual true mean VOT is 95 ms. If our sample mean is 95 ms and the standard deviation is again about 35 ms we are surely going to conclude that the null hypothesis (H_0: $\mu = 100$) is probably true. At least our data is not inconsistent with the hypothesis because 24% of the time ($p = 0.24$) we can get a *t*-value that is equal to or less than −0.706.

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{95 - 100}{36.1/\sqrt{26}} = \frac{-5}{7.08} = -0.706 \quad \text{testing for a small difference}$$

Nonetheless, by accepting the null hypothesis we have made a type II error. Just as we can choose a criterion α level for the acceptable type I error rate, we can also require that our statistics avoid type II errors. The probability of making a type II error is called β, and the value we are usually interested in is $1 - \beta$, the *power* of our statistical test. As my example illustrates, to avoid type II errors you need to have statistical tests that are sensitive enough to catch small differences between the sample mean and the population mean – to detect that 95 really is different from 100. With only 26 observations ($n = 26$) and a standard deviation of 36.1, if we set the power of our test to 0.8 (that is, accept type II errors 20% of the time with $\beta = 0.2$) the difference between the hypothesized mean and the true population mean would

have to be 18 ms before we could detect the difference. To detect a smaller difference like 5 ms we would have to increase the power of the hypothesis test.

You'll notice that in the calculation of t there are two parameters other than the sample mean and the population mean that affect the t-value. These are the standard deviation of the sample (s) and the size of the sample (n). To increase the power of the t-test we need to either reduce the standard deviation or increase the number of observations. Sometimes you can reduce the standard deviation by controlling some uncontrolled sources of variance. For example, in this VOT data I pooled observations from both /t/ and /k/. These probably do have overall different average VOT, so by pooling them I have inflated the standard deviation. If we had a sample of all /k/ VOTs the standard deviation might be lower and thus the power of the t-test greater. Generally, though, the best way to increase the power of your test is to get more data. In this case, if we set the probability of a type I error at 0.05, the probability of a type II error at 0.2, and we want to be able to detect that 95 ms is different from the hypothesized 100 ms, then we need to have an n of 324 observations (see the R note for the magic).

R note. The R function `power.t.test()` provides a way to estimate how many observations you need to make in order to detect differences between means of any specified magnitude with α and β error probabilities controlled. In the call below, I specified that I want to detect a difference of 5 ms (`delta=5`), that I expect that the standard deviation of my observations will be 36.1 (`sd=36.1`), and that we are testing whether the sample mean is less than the hypothesized mean, not just different one way or the other (`alternative = "one.sided"`). With $\alpha = 0.05$ (`sig.level=0.05`) and $\beta = 0.2$ (`power=0.8`), this function reports that we need 324 (`n = 323.6439`) observations to detect the 5 ms difference.

```
power.t.test(power=0.8,sig.level=0.05,delta=5,sd=36.1,
type="one.sample",alternative="one.sided")
One-sample t test power calculation
  n = 323.6439
```



```
  delta = 5
  sd = 36.1
  sig.level = 0.05
  power = 0.8
  alternative = one.sided
```

Of course, collecting more data is time consuming, so it is wise to ask, as Ilse Lehiste once asked me, "sure it is significant, but is it important?" It may be that a 5 ms difference is too small to be of much practical or theoretical importance, so taking the trouble to collect enough data so that we can detect such a small difference is really just a waste of time.

2.4 Correlation

So far we have been concerned in this chapter with the statistical background assumptions that make it possible to test hypotheses about the population mean. This is the "tests" portion of this chapter on "Patterns and tests." You can be sure that we will be coming back to this topic in several practical applications in chapters to follow. However, because this chapter is aiming to establish some of the basic building blocks that we will return to over and over in the subsequent chapters, I would like to suspend the "tests" discussion at this point and turn to the "patterns" portion of the chapter. The aim here is to explain some of the key concepts that underlie studies of relationships among variables – in particular to review the conceptual and mathematical underpinnings of correlation and regression.

One way to explore the relationship between two variables is by looking at counts in a contingency table. For example, we have a data set of two measurements of the lowest vocal tract resonance frequency – the first formant (F1). We have F1 values for men and women for the vowels /i/, /e/, /a/, /o/, and /u/ in four different languages (see the data file "F1_data.txt"). Women tend to have shorter vocal tracts than men and thus to have higher resonance frequencies. This is the case in our data set, where the average F1 of the women is 534.6 Hz and the average F1 for men is 440.9. We can construct a contingency table by counting how many of the observations in this data set fall

Table 2.3 A 2 × 2 contingency table showing the number of F1 values above or below the average F1 for men and women in the small "F1_data.txt" data set.

		Female F1 below	Female F1 above
Male F1	above	0	6
	below	12	1

above or below the mean on each of the two variables being compared. For example, we have the five vowels in Sele (a language spoken in Ghana) measured on two variables – male F1 and female F1 – and we are interested in studying the relationship or correlation between male and female F1 frequency. The contingency table in Table 2.3 then shows the number of times that we have a vowel that has female F1 above the average female F1 value while for the same vowel in the same language we have male F1 above the male F1 average – this "both above" condition happens six times in this small data set. Most of the other vowels (twelve) have both male and female F1 falling below their respective means, and in only one vowel was there a discrepancy, where the female F1 was above the female mean while the male F1 was below the male F1.

So, Table 2.3 shows how the F1 values fall in a two by two (2 × 2) contingency table. This table shows that for a particular vowel in a particular language (say /i/ in Sele), if the male F1 falls below the average male F1, then the female F1 for that vowel will probably also fall below the average F1 for female speakers. In only one case does this relationship not hold.

I guess it is important to keep in mind that Table 2.3 didn't have to come out this way. For instance, if F1 was not acoustically related to vowel quality, then pairing observations of male and female talkers according to which vowel they were producing would not have resulted in matched patterns of F1 variation.

Contingency tables are a useful way to see the relationship, or lack of one, between two variables, and we will see in Chapter 5 that when

the counts are a good deal larger than these – particularly when we have more than 5 or 10 observations even in the smallest cell of the table – we can test the strength of the relationship using the χ^2 distribution. However, we threw away a lot of information by constructing this contingency table. From Table 2.3 all we know is that if the male F1 is above average so is the female F1, but we don't know whether they tend to be the same amount above average or if sometimes the amount above average for males is much more than it is for females. It would be much better to explore the relationship of these two variables without throwing out this information.

In Figure 2.7 you can see the four cells of Table 2.3. There are 6 data points in the upper right quadrant of the graph, 12 data points in the lower left, and 1 that just barely ended up in the lower right quadrant. These quadrants were marked in the graph by drawing a dashed line at the mean values for the male (441 Hz) and female (535 Hz) talkers. As you can see, the relationship between male and female F1 values goes beyond simply being in one quadrant of the graph or not. In fact,

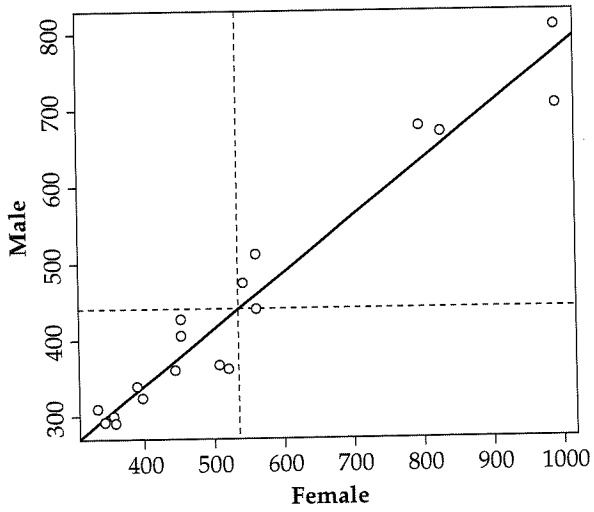

Figure 2.7 Nineteen pairs of male and female F1 values drawn from four different languages and 4 or 5 vowels in each language. The grid lines mark the average female (vertical line) and male (horizontal line) F1 values. The diagonal line is the best-fitting straight line (the linear regression) that relates female F1 to male F1.

it looks as though if we divided the lower left and the upper right quadrants into quadrants again we would still have the relationship, higher male F1 is associated with higher female F1. We need a measure of association that will give us a consistent indication of how closely related two variables are.

R note. For the example in Table 2.3 and Figure 2.7, I used data that are stored in a data file that I called "F1_data.txt". This is laid out like a spread sheet, but it is important to keep in mind that it is a text file. So, if you are preparing data for import into R you should be sure to save your data file as a ".txt" file (actually R doesn't care about the file name extension, but some programs do). The first three lines of my data file look like this:

```
female  male  vowel  language
391     339   i      W.Apache
561     512   e      W.Apache
......
```

The first row contains the names of the variables, and the following rows each contain one pair of observations. For example, the first row indicates that the vowel [i] in Western Apache has an F1 value of 391 Hz for women and 339 Hz for men. I used the `read.delim()` function to read my data from the file into an R `data.frame` object.

```
f1data = read.delim("F1_data.txt")
```

This object `f1data` is composed of four vectors, one for each column in the data file. So, if I would like to see the vector of female F1 measurements I can type the name of the data frame followed by a dollar sign and then the name of the vector within that data frame.

```
> f1data$female
 [1] 391 561 826 453 358 454 991 561 398 334 444 796 542 333
[15] 343 520 989 507 357
```

The command `summary()` is a useful one for verifying that your data file has been read correctly.


```
> summary(f1data)
     female          male       vowel  language
 Min.   :333.0  Min.   :291.0  a:4    CA English:5
 1st Qu.:374.5  1st Qu.:317.5  e:4    Ndumbea   :5
 Median :454.0  Median :367.0  i:4    Sele      :5
 Mean   :534.6  Mean   :440.9  o:4    W.Apache  :4
 3rd Qu.:561.0  3rd Qu.:493.0  u:3
 Max.   :991.0  Max.   :809.0
```

It is a bit of a pain to keep typing `f1data$female` to refer to a vector, so the `attach()` command is useful because once a data frame has been attached you don't have to mention the data frame name.

```
attach(f1data)
```

Now, having read in the data, here are the commands I used to produce Figure 2.7 (I'll discuss the diagonal line later):

```
plot(female,male)
lines(x=c(mean(female),mean(female)),y=c(200,900),lty=2)
lines(x=c(200,1100),y=c(mean(male),mean(male)),lty=2)
```

Finally, as you might expect, R has built-in functions to calcluate the covariance and correlation between two variables.

```
> cov(female,male)
[1] 33161.79
> cor(female,male)
[1] 0.9738566
```

2.4.1 *Covariance and correlation*

The key insight in developing a measure of association between two variables is to measure deviation from the mean $(x_i - \bar{x})$. As we saw in Figure 2.7, the association of male F1 and female F1 can be captured by noticing that when female F1 (let's name this variable x) was higher than the female mean, male F1 (y) was also higher than the male mean. That is, if $x_i - \bar{x}$ is positive then $y_i - \bar{y}$ is also positive. What is more, the association is strongest when the magnitudes of these deviations are matched – when x_i is quite a bit larger than the x mean and y_i is also

quite a bit larger than the y mean. We can get an overall sense of how strong the association of two variables is by multiplying the deviations of x and y and summing these products for all of the observations.

$$\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad \text{sum of product of deviations}$$

Notice that if x_i is much larger than the mean and y_i is also much larger than the mean then the product will be greater than if y_i is only a little larger than the mean. Notice also that if x_i is quite a bit less than the mean and y_i is also quite a bit less than the mean the product will again be a large positive value.

The product of the deviations will be larger as we have a larger and larger data set, so we need to normalize this value to the size of the data set by taking the average of the paired deviations. This average product of the deviations is called the covariance of X and Y.

$$\frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \quad \text{covariance of X and Y}$$

Of course, the size of a deviation from the mean can be standardized so that we can compare deviations from different data sets on the same measurement scale. We saw that deviation can be expressed in units of standard deviation with the z-score normalization. This is commonly done when we measure association too.

$$\frac{\sum_{i=0}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n} = \frac{\sum_{i=0}^{n}(z_x)(z_y)}{n} = r_{xy} \quad \text{correlation of X and Y}$$

The main result here is that the correlation coefficient r_{xy} is simply a scaled version of the sum of the product of the deviations using the idea that this value will be highest when x and y deviate from their means in comparable magnitude. Correlation is identical to covariance, except that correlation is scaled by the standard deviations. So covariance can have any value, and correlation ranges from 1 to –1 (perfect positive correlation is 1 and perfect negative correlation is –1).

2.4.2 *The regression line*

Notice in Figure 2.7 that I put a diagonal line through the data points that shows generally the relationship between female and male F1. This

line was not drawn by hand, but was calculated to be the best fitting straight line that could possibly be drawn. Well, let's qualify that by saying that it is the best-fitting least-squares estimate line. Here the squared deviations that we are trying to minimize as we find the best line are the differences between the predicted values of y_i, which we will write as y_i hat ($\hat{y}_i$), and the actual values. So the least-squares estimate will minimize $\sum(y_i - \hat{y}_i)^2$. The difference between the predicted value and the actual value for any observation of y_i is called the "residual."

Here's how to find the best-fitting line. If we have a perfect correlation between x and y then the deviations z_x and z_y are equal to each other for every observation. So we have:

$$\frac{y_i - \bar{y}}{s_y} = \frac{x_i - \bar{x}}{s_x} \quad \text{deviations are equivalent if } r_{xy} = 1$$

So, if we solve for y_i to get the formula for the predicted $\hat{y}_i$ if the correlation is perfect, then we have:

$$\hat{y}_i = \frac{s_y}{s_x}(x_i - \bar{x}) + \bar{y} \quad \text{predicting } y_i \text{ from } x_i \text{ when } r_{xy} = 1$$

Because we want our best prediction even when the correlation isn't perfect and the best prediction of z_y is r_{xy} times z_x, then our best prediction of y_i is:

$$\hat{y}_i = r_{xy}\frac{s_y}{s_x}(x_i - \bar{x}) + \bar{y} \quad \text{predicting } y_i \text{ from } x_i \text{ when } r_{xy} \neq 1$$

Now to put this into the form of an equation for a straight line ($\hat{y}_i = A + Bx_i$) we let the slope of the line $B = r_{xy}(s_x/s_y)$ and the intercept of the line $A = \bar{y} - B\bar{x}$.

R note. Let's return to the male versus female vowel F1 data and see if we can find the slope and intercept of the best-fitting line that relates these two variables. Note that we are justified in fitting a line to these data because when we look at the graph, the relationship looks linear.

Using the correlation function cor() and the standard deviation function sd(), we can calculate the slope of the regression line.

```
cor(male,female)
[1] 0.9738566
sd(male)
[1] 160.3828
sd(female)
[1] 212.3172
B = cor(male,female)*(sd(male)/sd(female))
A = mean(male) - B*mean(female)
A
[1] 47.59615
B
[1] 0.7356441
```

Now we have values for A and B, so that we can predict the male F1 value from the female F1:

male F1 = 0.736 * female F1 + 47.6

Consider, for example, what male F1 value we expect if the female F1 is 700 Hz. The line we have in Figure 2.7 leads us to expect a male F1 that is a little higher than 550 Hz. We get a more exact answer by applying the linear regression coefficients:

```
> B*700 + A
[1] 562.547
```

2.4.3 *Amount of variance accounted for*

So now we have a method to measure the association between two continuous variables giving us the Pearson's product moment correlation (r_{xy}), and a way to use that measure to determine the slope and intercept of the best-fitting line that relates x and y (assuming that a linear relationship is correct).

So what I'd like to present in this section is a way to use the correlation coefficient to measure the percent of variance in y that we can correctly predict as a linear function of x. Then we will see how to put all of this stuff together in R, which naturally has a function that does it all for you.

So, we have a statistic r_{xy} that ranges from –1 to 1 that indicates the degree of association between two variables. And we have a linear function $\hat{y}_i = A + Bx_i$ that uses r_{xy} to predict y from x. What we want now is a measure of how much of the variability of y is accurately predicted by this linear function. This will be the "amount of variance accounted for" by our linear model.

Is it a model or just a line? What's in a name? If it seems reasonable to think that x might cause y we can think of the linear function as a model of the causal relationship and call it a "regression model." If a causal relationship doesn't seem reasonable then we'll speak of the correlation of two variables.

As it turns out, you can simply square the correlation coefficient to get the amount of variance in y that the line $A + Bx$ accounts for. I hope that it will add some insight to look at how we come to this conclusion.

$r^2 = r_{xy}r_{xy}$ r-squared, the amount of variance accounted for

The variance of y is s_y^2. We are trying to measure how much of this variance can be accounted for by the line $\hat{y}_i = A + Bx_i$. The amount of variance that is predicted by this "linear regression function" is $s_{\hat{y}}^2$. Which means that the unpredicted variance is the variance of the deviation between the actual y_i values and the predicted values $\hat{y}_i$. Call this unpredicted, or residual, variance $s_{y-\hat{y}}^2$. Because we used the optimal rule (the least-squares criterion) to relate x and y, $s_{\hat{y}}^2$ and $s_{y-\hat{y}}^2$ are not correlated with each other, therefore

$$s_y^2 = s_{\hat{y}}^2 + s_{y-\hat{y}}^2.$$

In words that is: The total variance of y is composed of the part that can be predicted if we know x and the part that is independent of x.

If we consider this same relationship in terms of z-scores instead of in terms of the raw data ($s_{z_y}^2 = s_{z_{\hat{y}}}^2 + s_{z_{y-\hat{y}}}^2$) we can equivalently talk about it in terms of proportion of variance because the variance s_z^2 of the normal distribution is equal to one. Then instead of dividing the total amount of variance into a part that can be predicted by the line and a part that remains unpredicted we can divide it into a proportion can be predicted and a remainder.

In terms of z-scores the line equation $\hat{y} = A + Bx$ is $\hat{z}_y = rz_x$ and from the definition of variance, then,

$$s^2_{\hat{z}_y} = \frac{\sum(rz_x)^2}{n} = r^2\frac{\sum(z_x^2)}{n} = r^2 \quad \text{proportion of variance accounted for is } r^2$$

I guess, in deciphering this it helps to know that $\sum z_x^2 = n$ because the standard deviation of z is 1.

The key point here is that r^2 is equivilent to $s^2_{\hat{z}_y}$, the proportion of total variance of y that can be predicted by the line.

R note. We return once again to the male and female vowel F1 data. Earlier we calculated the regression coefficients from the standard deviations and the correlation of male and female F1 values. What I'd like to do here is to introduce the function `lm()`. This function calculates a linear model in which we try to predict one variable from another variable (or several).

In this example, I ask to see a summary of the linear model in which we try to predict male F1 from female F1. You can read "male~female" as "male as a function of female".

```
> summary(lm(male~female))

Call:
lm(formula = male ~ female)

Residuals:
    Min      1Q  Median      3Q     Max
-70.619 -18.170   3.767  26.053  51.707

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 47.59615   23.85501    1.995    0.0623  .
female       0.73564    0.04162   17.676  2.23e-12  ***
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 37.49 on 17 degrees of freedom
Multiple R-Squared: 0.9484,  Adjusted R-squared: 0.9454
F-statistic: 312.4 on 1 and 17 DF,  p-value: 2.230e-12
```

Notice that this summary statement gives us a report on several aspects of the linear regression fit. The first section of the report is on the residuals ($y_i - \hat{y}_i$) that shows their range, median, and

quartiles. The second section reports the coefficients. Notice that `lm()` calculates the same A and B coefficients that we calcuated in explicit formulas above. But now we also have a t-test for both the intercept and the slope.

Finally, `lm()` reports the r^2 value and again tests whether this amount of variance accounted for is greater than zero – using an F-test. The line $F1_{male} = 47.596 + 0.7356{*}F1_{female}$ accounts for almost 95% of the variance in the male F1 values.

The t-tests for A and B (the regression coefficients) above indicate that the slope (labeled `female`) is definitely different from zero but that the `intercept` may not be reliably different from zero. This means that we might simplify the predictive formula by rerunning `lm()` specifying that we don't want the equation to have an intercept value. When I did this using the command `lm(male ~ female-1)`, where adding "-1" to the formula means "leave out the intercept (A) parameter," the regression accounted for 99% of the male F1 variance by simply multiplying each paired female F1 value by 0.813. That is, the regression formula was $F1_{male} = 0.813{*}F1_{female}$. The two regression analyses (with and without the intercept coefficient) are shown in Figure 2.8.

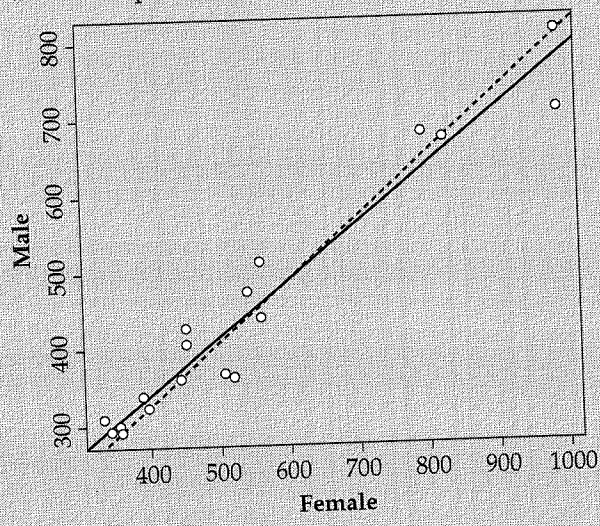

Figure 2.8 Regression lines at $y = 47.5 + 0.736x$ (solid line) and $y = 0.81x$ (dashed line).

EXERCISES

1 Given a skewed distribution (Figure 2.3) what distribution of the mean do you expect for samples of size $n = 1$?

2 I used the custom R function `central.limit()` to draw Q-Q plots for samples from this skewed distribution for samples of size $n = 1$ through $n = 8$. Here are the correlations between the normal distribution and the distribution of the means for samples of these sizes:

n	1	2	3	4	6	8
correlation	0.965	0.981	0.988	0.991	0.993	0.997

What do these values indicate about the distribution of the mean? Where was the biggest change in the correlation?

3 We test the hypothesis that the mean of our data set ($\bar{x} = 113.5$, $s = 35.9$, $n = 18$) is no different from 100, and find that the t is 1.59, and the probability of finding a higher t is 0.065. Show how to get this t-value and this probability from the t-distribution. What do you conclude from this t-test?

4 Calculate the covariance and correlation of the following data set by hand (well, use a calculator!). Plot the data and notice that the relationship is such that as Y gets bigger X gets smaller. How do your covariance and correlation values reflect this "down going" trend in the data?

X	Y
90	−7
82	−0.5
47	8
18	32
12	22
51	17
46	13
2	31
48	11
72	4
18	29
13	32

5 `Source()` the following function and explore the χ^2 distribution. It is said that the expression $(n-1)s^2/\sigma^2$ is distributed in a family of distributions (one slightly different distribution for each value of n) that is analogous to the t-distribution. Try this function out with different values of n, and different population standard deviations. Describe what this function does in a flow chart or a paragraph – whatever makes most sense to you. What is the effect of choosing samples of different size ($n = 4, \ldots, 50$)?


```
#--------- chisqu ---------------
#
#The input parameters are:
#      n - size of each sample
#      m - number of samples of size n to select
#      mu, sigma - the mean and sd the normal distribution from
       which samples are drawn

chisq = function(n=15,m=5000,mu=0,sigma=1) {

      sigsq=(sigma*sigma)
      xlow = 0
      xhigh = 2*n

      vars = vector() # I hereby declare that "vars" is a vector

      for (i in 1:m) { # get m samples
            data= rnorm(n,mu,sigma) # sample the normal dist
            vars[i] = var(data) # vars is our array of variances
      }
      title = paste("Sample size = ",n, "df = ",n-1)
      hist((n-1)*vars/sigsq,
         xlim=(xlow,xhigh),main=title,freq=F)
      plot(function(x)dchisq(x,df=(n-1)),xlow,xhigh,add=T)
}
```

6 Is the population mean (μ) of the 1971 Cherokee VOT data (Chapter 1, Table 1.1) 100 ms? How sure are you?

7 You want to be able to detect a reaction time difference as small as 20 ms between two conditions in a psycholinguistic experiment. You want your t-test to have a criterion type I error rate of 0.05 and you want the type II error rate to be 0.2. The standard deviation in such experiments is typically about 60 ms, so how many observations do you have to make in order to detect a 20 ms difference? Try this with the type of comparison being "paired" instead of "two. sample" and with the standard deviation of the differences being 20 ms. Which method do you prefer – two independent samples, or paired observations?