

Introduction to Statistical Thought

Michael Lavine

August 25, 2009

Copyright © 2005 by Michael Lavine

CONTENTS

LIST OF FIGURES	v
LIST OF TABLES	x
PREFACE	xi
1 PROBABILITY	1
1.1 BASIC PROBABILITY	1
1.2 PROBABILITY DENSITIES	6
1.3 PARAMETRIC FAMILIES OF DISTRIBUTIONS	13
1.3.1 THE BINOMIAL DISTRIBUTION	14
1.3.2 THE POISSON DISTRIBUTION	17
1.3.3 THE EXPONENTIAL DISTRIBUTION	20
1.3.4 THE NORMAL DISTRIBUTION	22
1.4 CENTERS, SPREADS, MEANS, AND MOMENTS	29
1.5 JOINT, MARGINAL AND CONDITIONAL PROBABILITY	40
1.6 ASSOCIATION, DEPENDENCE, INDEPENDENCE	50
1.7 SIMULATION	56
1.7.1 CALCULATING PROBABILITIES	56
1.7.2 EVALUATING STATISTICAL PROCEDURES	60
1.8 R	71
1.9 SOME RESULTS FOR LARGE SAMPLES	76
1.10 EXERCISES	80
2 MODES OF INFERENCE	92
2.1 DATA	92
2.2 DATA DESCRIPTION	93
2.2.1 SUMMARY STATISTICS	93

2.2.2	DISPLAYING DISTRIBUTIONS	99
2.2.3	EXPLORING RELATIONSHIPS	112
2.3	LIKELIHOOD	131
2.3.1	THE LIKELIHOOD FUNCTION	131
2.3.2	LIKELIHOODS FROM THE CENTRAL LIMIT THEOREM	137
2.3.3	LIKELIHOODS FOR SEVERAL PARAMETERS	142
2.4	ESTIMATION	151
2.4.1	THE MAXIMUM LIKELIHOOD ESTIMATE	151
2.4.2	ACCURACY OF ESTIMATION	153
2.4.3	THE SAMPLING DISTRIBUTION OF AN ESTIMATOR	156
2.5	BAYESIAN INFERENCE	161
2.6	PREDICTION	171
2.7	HYPOTHESIS TESTING	175
2.8	EXERCISES	189
3	REGRESSION	199
3.1	INTRODUCTION	199
3.2	NORMAL LINEAR MODELS	207
3.2.1	INTRODUCTION	207
3.2.2	INFERENCE FOR LINEAR MODELS	218
3.3	GENERALIZED LINEAR MODELS	231
3.3.1	LOGISTIC REGRESSION	231
3.3.2	POISSON REGRESSION	241
3.4	PREDICTIONS FROM REGRESSION	245
3.5	EXERCISES	249
4	MORE PROBABILITY	260
4.1	MORE PROBABILITY DENSITY	260
4.2	RANDOM VECTORS	261
4.2.1	DENSITIES OF RANDOM VECTORS	261
4.2.2	MOMENTS OF RANDOM VECTORS	263
4.2.3	FUNCTIONS OF RANDOM VECTORS	263
4.3	REPRESENTING DISTRIBUTIONS	267
4.4	EXERCISES	272
5	SPECIAL DISTRIBUTIONS	275
5.1	BINOMIAL AND NEGATIVE BINOMIAL	275
5.2	MULTINOMIAL	285
5.3	POISSON	287

5.4	UNIFORM	300
5.5	GAMMA, EXPONENTIAL, CHI SQUARE	301
5.6	BETA	308
5.7	NORMAL	311
5.7.1	THE UNIVARIATE NORMAL DISTRIBUTION	311
5.7.2	THE MULTIVARIATE NORMAL DISTRIBUTION	316
5.7.3	MARGINAL, CONDITIONAL, AND RELATED DISTRIBUTIONS	323
5.8	THE t DISTRIBUTION	326
5.9	EXERCISES	331
6	BAYESIAN STATISTICS	341
6.1	MULTIDIMENSIONAL BAYESIAN ANALYSIS	341
6.2	METROPOLIS, METROPOLIS-HASTINGS, AND GIBBS	353
6.3	EXERCISES	372
7	MORE MODELS	376
7.1	FIXED EFFECTS, RANDOM EFFECTS, HIERARCHICAL MODELS	376
7.2	TIME SERIES AND MARKOV CHAINS	390
7.3	SURVIVAL ANALYSIS	403
7.4	EXERCISES	410
8	MATHEMATICAL STATISTICS	413
8.1	PROPERTIES OF STATISTICS	413
8.1.1	SUFFICIENCY	413
8.1.2	CONSISTENCY, BIAS, AND MEAN-SQUARED ERROR	416
8.2	INFORMATION	421
8.3	EXPONENTIAL FAMILIES	424
8.4	ASYMPTOTICS	426
8.4.1	MODES OF CONVERGENCE	432
8.4.2	THE δ -METHOD	436
8.4.3	THE ASYMPTOTIC BEHAVIOR OF ESTIMATORS	439
8.5	EXERCISES	444
BIBLIOGRAPHY		449

LIST OF FIGURES

1.1	PDF FOR TIME ON HOLD AT HELP LINE	7
1.2	p_Y FOR THE OUTCOME OF A SPINNER	9
1.3	(A): OCEAN TEMPERATURES; (B): IMPORTANT DISCOVERIES	11
1.4	CHANGE OF VARIABLES	13
1.5	BINOMIAL PROBABILITIES	16
1.6	$P[X = 3 \lambda]$ AS A FUNCTION OF λ	19
1.7	EXPONENTIAL DENSITIES	21
1.8	NORMAL DENSITIES	24
1.9	OCEAN TEMPERATURES AT 45°N , 30°W , 1000M DEPTH	25
1.10	NORMAL SAMPLES AND NORMAL DENSITIES	27
1.11	HYDROGRAPHIC STATIONS OFF THE COAST OF EUROPE AND AFRICA	30
1.12	WATER TEMPERATURES	31
1.13	WATER TEMPERATURES WITH STANDARD DEVIATIONS	36
1.14	TWO PDF'S WITH ± 1 AND ± 2 SD's.	38
1.15	PERMISSIBLE VALUES OF N AND X	44
1.16	FEATURES OF THE JOINT DISTRIBUTION OF (X, Y)	48
1.17	LENGTHS AND WIDTHS OF SEPALS AND PETALS OF 150 IRIS PLANTS	51
1.18	CORRELATIONS	54
1.19	1000 SIMULATIONS OF $\hat{\theta}$ FOR $\text{n.sim} = 50, 200, 1000$	59
1.20	1000 SIMULATIONS OF $\hat{\theta}$ UNDER THREE PROCEDURES	63
1.21	MONTHLY CONCENTRATIONS OF CO_2 AT MAUNA LOA	65
1.22	1000 SIMULATIONS OF A FACE EXPERIMENT	68
1.23	HISTOGRAMS OF CRAPS SIMULATIONS	81
2.1	QUANTILES	96
2.2	HISTOGRAMS OF TOOTH GROWTH	100
2.3	HISTOGRAMS OF TOOTH GROWTH	101

2.4	HISTOGRAMS OF TOOTH GROWTH	102
2.5	CALORIE CONTENTS OF BEEF HOT DOGS	106
2.6	STRIP CHART OF TOOTH GROWTH	108
2.7	QUIZ SCORES FROM STATISTICS 103	111
2.8	QQ PLOTS OF WATER TEMPERATURES ($^{\circ}$ C) AT 1000M DEPTH	113
2.9	MOSAIC PLOT OF UCBADMISSIONS	117
2.10	MOSAIC PLOT OF UCBADMISSIONS	118
2.11	OLD FAITHFUL DATA.	121
2.12	WAITING TIME VERSUS DURATION IN THE OLD FAITHFUL DATASET	122
2.13	TIME SERIES OF DURATION AND WAITING TIME AT OLD FAITHFUL	123
2.14	TIME SERIES OF DURATION AND WAITING TIME AT OLD FAITHFUL	124
2.15	TEMPERATURE VERSUS LATITUDE FOR DIFFERENT VALUES OF LONGITUDE	126
2.16	TEMPERATURE VERSUS LONGITUDE FOR DIFFERENT VALUES OF LATITUDE	127
2.17	SPIKE TRAIN FROM A NEURON DURING A TASTE EXPERIMENT. THE DOTS SHOW THE TIMES AT WHICH THE NEURON FIRED. THE SOLID LINES SHOW TIMES AT WHICH THE RAT RECEIVED A DROP OF A .3 M SOLUTION OF NaCl.	129
2.18	LIKELIHOOD FUNCTION FOR THE PROPORTION OF RED CARS	132
2.19	$\ell(\theta)$ AFTER $\sum y_i = 40$ IN 60 QUADRATS.	135
2.20	LIKELIHOOD FOR SLATER SCHOOL	136
2.21	MARGINAL AND EXACT LIKELIHOODS FOR SLATER SCHOOL	139
2.22	MARGINAL LIKELIHOOD FOR MEAN CEO SALARY	141
2.23	FACE EXPERIMENT: DATA AND LIKELIHOOD	144
2.24	LIKELIHOOD FUNCTION FOR QUIZ SCORES	146
2.25	LOG OF THE LIKELIHOOD FUNCTION FOR (λ, θ_f) IN EXAMPLE 2.13	150
2.26	LIKELIHOOD FUNCTION FOR THE PROBABILITY OF WINNING CRAPS	155
2.27	SAMPLING DISTRIBUTION OF THE SAMPLE MEAN AND MEDIAN	157
2.28	HISTOGRAMS OF THE SAMPLE MEAN FOR SAMPLES FROM $\text{Bin}(n, .1)$	159
2.29	PRIOR, LIKELIHOOD AND POSTERIOR IN THE SEEDLINGS EXAMPLE	166
2.30	PRIOR, LIKELIHOOD AND POSTERIOR DENSITIES FOR λ WITH $n = 1, 4, 16$	168
2.31	PRIOR, LIKELIHOOD AND POSTERIOR DENSITIES FOR λ WITH $n = 60$	169
2.32	PRIOR, LIKELIHOOD AND POSTERIOR DENSITY FOR SLATER SCHOOL	170
2.33	PLUG-IN PREDICTIVE DISTRIBUTION FOR SEEDLINGS	172
2.34	PREDICTIVE DISTRIBUTIONS FOR SEEDLINGS AFTER $n = 0, 1, 60$	176
2.35	PDF OF THE $\text{Bin}(100, .5)$ DISTRIBUTION	180
2.36	PDFS OF THE $\text{Bin}(100, .5)$ (DOTS) AND $\text{N}(50, 5)$ (LINE) DISTRIBUTIONS	181
2.37	APPROXIMATE DENSITY OF SUMMARY STATISTIC t	183
2.38	NUMBER OF TIMES BABOON FATHER HELPS OWN CHILD	187
2.39	HISTOGRAM OF SIMULATED VALUES OF W.TOT	188

3.1	FOUR REGRESSION EXAMPLES	200
3.2	1970 DRAFT LOTTERY. DRAFT NUMBER VS. DAY OF YEAR	203
3.3	DRAFT NUMBER VS. DAY OF YEAR WITH SMOOTHERS	204
3.4	TOTAL NUMBER OF NEW SEEDLINGS 1993 – 1997, BY QUADRAT.	206
3.5	CALORIE CONTENT OF HOT DOGS	208
3.6	DENSITY ESTIMATES OF CALORIE CONTENTS OF HOT DOGS	210
3.7	THE PLANTGROWTH DATA	212
3.8	ICE CREAM CONSUMPTION VERSUS MEAN TEMPERATURE	219
3.9	LIKELIHOOD FUNCTIONS FOR $(\mu, \delta_M, \delta_P)$ IN THE HOT DOG EXAMPLE.	225
3.10	PAIRS PLOT OF THE MTCARS DATA	227
3.11	MTCARS — VARIOUS PLOTS	230
3.12	LIKELIHOOD FUNCTIONS FOR $\beta_1, \gamma_1, \delta_1$ AND δ_2 IN THE MTCARS EXAMPLE.	232
3.13	PINE CONES AND O-RINGS	235
3.14	PINE CONES AND O-RINGS WITH REGRESSION CURVES	236
3.15	LIKELIHOOD FUNCTION FOR THE PINE CONE DATA	239
3.16	ACTUAL VS. FITTED AND RESIDUALS VS. FITTED FOR THE SEEDLING DATA	244
3.17	DIAGNOSTIC PLOTS FOR THE SEEDLING DATA	246
3.18	ACTUAL MPG AND FITTED VALUES FROM THREE MODELS	248
3.19	HAPPINESS QUOTIENT OF BANKERS AND POETS	253
4.1	THE (X_1, X_2) PLANE AND THE (Y_1, Y_2) PLANE	266
4.2	PMF'S, PDF'S, AND CDF'S	269
5.1	THE BINOMIAL PMF	281
5.2	THE NEGATIVE BINOMIAL PMF	284
5.3	POISSON PMF FOR $\lambda = 1, 4, 16, 64$	290
5.4	RUTHERFORD AND GEIGER'S FIGURE 1	295
5.5	NUMBERS OF FIRINGS OF A NEURON IN 150 msec AFTER FIVE DIFFERENT TASTANTS. TASTANTS: 1=MSG .1M; 2=MSG .3M; 3=NaCl .1M; 4=NaCl .3M; 5=WATER. PANELS: A: A STRIPCHART. EACH CIRCLE REPRESENTS ONE DELIVERY OF A TASTANT. B: A MOSAIC PLOT. C: EACH LINE REPRESENTS ONE TASTANT. D: LIKELIHOOD FUNCTIONS. EACH LINE REPRESENTS ONE TASTANT.	297
5.6	THE LINE SHOWS POISSON PROBABILITIES FOR $\lambda = 0.2$; THE CIRCLES SHOW THE FRACTION OF TIMES THE NEURON RESPONDED WITH 0, 1, ..., 5 SPIKES FOR EACH OF THE FIVE TASTANTS.	299
5.7	GAMMA DENSITIES	302
5.8	EXPONENTIAL DENSITIES	306
5.9	BETA DENSITIES	310
5.10	WATER TEMPERATURES ($^{\circ}$ C) AT 1000M DEPTH	312

5.11	BIVARIATE NORMAL DENSITY	319
5.12	BIVARIATE NORMAL DENSITY	321
5.13	t DENSITIES FOR FOUR DEGREES OF FREEDOM AND THE $N(0, 1)$ DENSITY	330
6.1	POSTERIOR DENSITIES OF β_0 AND β_1 IN THE ICE CREAM EXAMPLE USING THE PRIOR FROM EQUATION 6.4.	345
6.2	NUMBERS OF PINE CONES IN 1998 AS A FUNCTION OF DBH	349
6.3	NUMBERS OF PINE CONES IN 1999 AS A FUNCTION OF DBH	350
6.4	NUMBERS OF PINE CONES IN 2000 AS A FUNCTION OF DBH	351
6.5	10,000 MCMC SAMPLES OF THE $Be(5, 2)$ DENSITY. TOP PANEL: HISTOGRAM OF SAMPLES FROM THE METROPOLIS-HASTINGS ALGORITHM AND THE $Be(5, 2)$ DENSITY. MIDDLE PANEL: θ_i PLOTTED AGAINST i . BOTTOM PANEL: $p(\theta_i)$ PLOTTED AGAINST i	356
6.6	10,000 MCMC SAMPLES OF THE $Be(5, 2)$ DENSITY. LEFT COLUMN: $(\theta^* \theta) = U(\theta - 100, \theta + 100)$; RIGHT COLUMN: $(\theta^* \theta) = U(\theta - .00001, \theta + .00001)$. TOP: HISTOGRAM OF SAMPLES FROM THE METROPOLIS-HASTINGS ALGORITHM AND THE $Be(5, 2)$ DENSITY. MIDDLE: θ_i PLOTTED AGAINST i . BOTTOM: $p(\theta_i)$ PLOTTED AGAINST i	358
6.7	TRACE PLOTS OF MCMC OUTPUT FROM THE PINE CONE CODE ON PAGE 360.	362
6.8	TRACE PLOTS OF MCMC OUTPUT FROM THE PINE CONE CODE WITH A SMALLER PROPOSAL RADIUS.	363
6.9	TRACE PLOTS OF MCMC OUTPUT FROM THE PINE CONE CODE WITH A SMALLER PROPOSAL RADIUS AND 100,000 ITERATIONS. THE PLOTS SHOW EVERY 10' TH ITERATION.	364
6.10	TRACE PLOTS OF MCMC OUTPUT FROM THE PINE CONE CODE WITH PROPOSAL FUNCTION G.ONE AND 100,000 ITERATIONS. THE PLOTS SHOW EVERY 10' TH ITERATION.	366
6.11	PAIRS PLOTS OF MCMC OUTPUT FROM THE PINE CONES EXAMPLE.	367
6.12	TRACE PLOTS OF MCMC OUTPUT FROM THE PINE CONE CODE WITH PROPOSAL FUNCTION G.GROUP AND 100,000 ITERATIONS. THE PLOTS SHOW EVERY 10' TH ITERATION.	370
6.13	PAIRS PLOTS OF MCMC OUTPUT FROM THE PINE CONES EXAMPLE WITH PROPOSAL G.GROUP	371
6.14	POSTERIOR DENSITY OF β_2 AND γ_2 FROM EXAMPLE 6.3.	372
7.1	PLOTS OF THE ORTHODONT DATA: DISTANCE AS A FUNCTION OF AGE, GROUPED BY SUBJECT, SEPARATED BY SEX.	378
7.2	PLOTS OF THE ORTHODONT DATA: DISTANCE AS A FUNCTION OF AGE, SEPARATED BY SUBJECT.	381

7.3	PERCENT BODY FAT OF MAJOR (BLUE) AND MINOR (PURPLE) <i>Pheidole morrisi</i> ANTS AT THREE SITES IN TWO SEASONS.	385
7.4	RESIDUALS FROM MODEL 7.4. EACH POINT REPRESENTS ONE COLONY. THERE IS AN UPWARD TREND, INDICATING THE POSSIBLE PRESENCE OF COLONY EFFECTS.	387
7.5	SOME TIME SERIES	392
7.6	Y_{t+1} VS. Y_t FOR THE BEAVER AND PRESIDENTS DATA SETS	393
7.7	Y_{t+k} VS. Y_t FOR THE BEAVER DATA SET AND LAGS 0–5	394
7.8	COPLOT OF $Y_{t+1} \sim Y_{t-1} Y_t$ FOR THE BEAVER DATA SET	396
7.9	FIT OF CO ₂ DATA	399
7.10	DAX CLOSING PRICES	401
7.11	DAX RETURNS	402
7.12	SURVIVAL CURVE FOR BLADDER CANCER. SOLID LINE FOR PLACEBO; DASHED LINE FOR THIOTEPAN.	406
7.13	CUMULATIVE HAZARD AND LOG(HAZARD) CURVES FOR BLADDER CANCER. SOLID LINE FOR THIOTEPAN; DASHED LINE FOR PLACEBO.	409
8.1	MEAN SQUARED ERROR FOR ESTIMATING BINOMIAL θ . SAMPLE SIZE = 5, 20, 100, 1000. $\alpha = \beta = 0$: SOLID LINE. $\alpha = \beta = 0.5$: DASHED LINE. $\alpha = \beta = 1$: DOTTED LINE. $\alpha = \beta = 4$: DASH-DOTTED LINE.	420
8.2	THE BE(.39, .01) DENSITY	430
8.3	DENSITIES OF \bar{Y}_{in}	431
8.4	DENSITIES OF Z_{in}	433
8.5	THE δ -METHOD	437
8.6	TOP PANEL: ASYMPTOTIC STANDARD DEVIATIONS OF δ_n AND δ'_n FOR $\Pr[X \leq a]$. THE SOLID LINE SHOWS THE ACTUAL RELATIONSHIP. THE DOTTED LINE IS THE LINE OF EQUALITY. BOTTOM PANEL: THE RATIO OF ASYMPTOTIC STANDARD DEVIATIONS.	441

LIST OF TABLES

1.1	PARTY AFFILIATION AND REFERENDUM SUPPORT	41
1.2	STEROID USE AND TEST RESULTS	43
2.1	NEW AND OLD SEEDLINGS IN QUADRAT 6 IN 1992 AND 1993	148
3.1	CORRESPONDENCE BETWEEN MODELS 3.3 AND 3.4	212
3.2	β 'S FOR FIGURE 3.14	234
5.1	RUTHERFORD AND GEIGER'S DATA	293
6.1	THE NUMBERS OF PINE CONES ON TREES IN THE FACE EXPERIMENT, 1998–2000.	347
7.1	FAT AS A PERCENTAGE OF BODY WEIGHT IN ANT COLONIES. THREE SITES, TWO SEASONS, TWO CASTES.	389

PREFACE

This book is intended as an upper level undergraduate or introductory graduate textbook in statistical thinking with a likelihood emphasis for students with a good knowledge of calculus and the ability to think abstractly. By “statistical thinking” is meant a focus on ideas that statisticians care about as opposed to technical details of how to put those ideas into practice. By “likelihood emphasis” is meant that the likelihood function and likelihood principle are unifying ideas throughout the text. Another unusual aspect is the use of statistical software as a pedagogical tool. That is, instead of viewing the computer merely as a convenient and accurate calculating device, we use computer calculation and simulation as another way of explaining and helping readers understand the underlying concepts.

Our software of choice is R (R DEVELOPMENT CORE TEAM [2006]). R and accompanying manuals are available for free download from [HTTP://WWW.R-PROJECT.ORG](http://www.r-project.org). You may wish to download **An Introduction to R** to keep as a reference. It is highly recommended that you try all the examples in R. They will help you understand concepts, give you a little programming experience, and give you facility with a very flexible statistical software package. And don’t just try the examples as written. Vary them a little; play around with them; experiment. You won’t hurt anything and you’ll learn a lot.

CHAPTER 1

PROBABILITY

1.1 Basic Probability

Let \mathcal{X} be a set and \mathcal{F} a collection of subsets of \mathcal{X} . A *probability measure*, or just a *probability*, on $(\mathcal{X}, \mathcal{F})$ is a function $\mu : \mathcal{F} \rightarrow [0, 1]$. In other words, to every set in \mathcal{F} , μ assigns a probability between 0 and 1. We call μ a *set function* because its domain is a collection of sets. But not just any set function will do. To be a probability μ must satisfy

1. $\mu(\emptyset) = 0$ (\emptyset is the empty set.),
2. $\mu(\mathcal{X}) = 1$, and
3. if A_1 and A_2 are disjoint then $\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2)$.

One can show that property 3 holds for any finite collection of disjoint sets, not just two; see Exercise 1. It is common practice, which we adopt in this text, to assume more — that property 3 also holds for any countable collection of disjoint sets.

When \mathcal{X} is a finite or countably infinite set (usually integers) then μ is said to be a *discrete* probability. When \mathcal{X} is an interval, either finite or infinite, then μ is said to be a *continuous* probability. In the discrete case, \mathcal{F} usually contains all possible subsets of \mathcal{X} . But in the continuous case, technical complications prohibit \mathcal{F} from containing all possible subsets of \mathcal{X} . See CASELLA AND BERGER [2002] or SCHERVISH [1995] for details. In this text we deemphasize the role of \mathcal{F} and speak of probability measures on \mathcal{X} without mentioning \mathcal{F} .

In practical examples \mathcal{X} is the set of outcomes of an “experiment” and μ is determined by experience, logic or judgement. For example, consider rolling a six-sided die. The set of outcomes is $\{1, 2, 3, 4, 5, 6\}$ so we would assign $\mathcal{X} \equiv \{1, 2, 3, 4, 5, 6\}$. If we believe the

die to be fair then we would also assign $\mu(\{1\}) = \mu(\{2\}) = \dots = \mu(\{6\}) = 1/6$. The laws of probability then imply various other values such as

$$\begin{aligned}\mu(\{1, 2\}) &= 1/3 \\ \mu(\{2, 4, 6\}) &= 1/2 \\ &\text{etc.}\end{aligned}$$

Often we omit the braces and write $\mu(2)$, $\mu(5)$, etc. Setting $\mu(i) = 1/6$ is not automatic simply because a die has six faces. We set $\mu(i) = 1/6$ because we believe the die to be fair.

We usually use the word “probability” or the symbol P in place of μ . For example, we would use the following phrases interchangeably:

- The probability that the die lands 1
- $P(1)$
- $P[\text{the die lands } 1]$
- $\mu(\{1\})$

We also use the word *distribution* in place of *probability measure*.

The next example illustrates how probabilities of complicated events can be calculated from probabilities of simple events.

Example 1.1 (The Game of Craps)

Craps is a gambling game played with two dice. Here are the rules, as explained on the website www.ONLINE-CRAPS-GAMBLING.COM/CRAPS-RULES.HTML.

For the dice thrower (shooter) the object of the game is to throw a 7 or an 11 on the first roll (a win) and avoid throwing a 2, 3 or 12 (a loss). If none of these numbers (2, 3, 7, 11 or 12) is thrown on the first throw (the Come-out roll) then a Point is established (the point is the number rolled) against which the shooter plays. The shooter continues to throw until one of two numbers is thrown, the Point number or a Seven. If the shooter rolls the Point before rolling a Seven he/she wins, however if the shooter throws a Seven before rolling the Point he/she loses.

Ultimately we would like to calculate $P(\text{shooter wins})$. But for now, let's just calculate

$$P(\text{shooter wins on Come-out roll}) = P(7 \text{ or } 11) = P(7) + P(11).$$

Using the language of page 1, what is X in this case? Let d_1 denote the number showing on the first die and d_2 denote the number showing on the second die. d_1 and d_2 are integers from 1 to 6. So X is the set of ordered pairs (d_1, d_2) or

$$\begin{aligned} & (6, 6) \quad (6, 5) \quad (6, 4) \quad (6, 3) \quad (6, 2) \quad (6, 1) \\ & (5, 6) \quad (5, 5) \quad (5, 4) \quad (5, 3) \quad (5, 2) \quad (5, 1) \\ & (4, 6) \quad (4, 5) \quad (4, 4) \quad (4, 3) \quad (4, 2) \quad (4, 1) \\ & (3, 6) \quad (3, 5) \quad (3, 4) \quad (3, 3) \quad (3, 2) \quad (3, 1) \\ & (2, 6) \quad (2, 5) \quad (2, 4) \quad (2, 3) \quad (2, 2) \quad (2, 1) \\ & (1, 6) \quad (1, 5) \quad (1, 4) \quad (1, 3) \quad (1, 2) \quad (1, 1) \end{aligned}$$

If the dice are fair, then the pairs are all equally likely. Since there are 36 of them, we assign $P(d_1, d_2) = 1/36$ for any combination (d_1, d_2) . Finally, we can calculate

$$\begin{aligned} P(7 \text{ or } 11) &= P(6, 5) + P(5, 6) + P(6, 1) + P(5, 2) \\ &\quad + P(4, 3) + P(3, 4) + P(2, 5) + P(1, 6) = 8/36 = 2/9. \end{aligned}$$

The previous calculation uses desideratum 3 for probability measures. The different pairs $(6, 5), (5, 6), \dots, (1, 6)$ are disjoint, so the probability of their union is the sum of their probabilities.

Example 1.1 illustrates a common situation. We know the probabilities of some simple events like the rolls of individual dice, and want to calculate the probabilities of more complicated events like the success of a Come-out roll. Sometimes those probabilities can be calculated mathematically as in the example. Other times it is more convenient to calculate them by computer simulation. We frequently use R to calculate probabilities. To illustrate, Example 1.2 uses R to calculate by simulation the same probability we found directly in Example 1.1.

Example 1.2 (Craps, continued)

To simulate the game of craps, we will have to simulate rolling dice. That's like randomly sampling an integer from 1 to 6. The `sample()` command in R can do that. For example, the following snippet of code generates one roll from a fair, six-sided die and shows R's response:

```
> sample(1:6, 1)
[1] 1
>
```

When you start R on your computer, you see `>`, R's prompt. Then you can type a command such as `sample(1:6, 1)` which means "take a sample of size 1 from the

numbers 1 through 6". (It could have been abbreviated `sample(6, 1)`.) R responds with [1] 1. The [1] says how many calculations R has done; you can ignore it. The 1 is R's answer to the `sample` command; it selected the number "1". Then it gave another >, showing that it's ready for another command. Try this several times; you shouldn't get "1" every time.

Here's a longer snippet that does something more useful.

```
> x <- sample ( 6, 10, replace=T ) # take a sample of
                                # size 10 and call it x
> x # print the ten values
[1] 6 4 2 3 4 4 3 6 6 2

> sum ( x == 3 ) # how many are equal to 3?
[1] 2
>
```

Note

- # is the comment character. On each line, R ignores all text after #.
- We have to tell R to take its sample *with replacement*. Otherwise, when R selects "6" the first time, "6" is no longer available to be sampled a second time. In `replace=T`, the T stands for True.
- <- does *assignment*. I.e., the result of `sample (6, 10, replace=T)` is assigned to a variable called x. The assignment symbol is two characters: < followed by -.
- A variable such as x can hold many values simultaneously. When it does, it's called a *vector*. You can refer to individual elements of a vector. For example, `x[1]` is the first element of x. `x[1]` turned out to be 6; `x[2]` turned out to be 4; and so on.
- == does *comparison*. In the snippet above, `(x==3)` checks, for each element of x, whether that element is equal to 3. If you just type `x == 3` you will see a string of T's and F's (True and False), one for each element of x. Try it.
- The `sum` command treats T as 1 and F as 0.
- R is almost always tolerant of spaces. You can often leave them out or add extras where you like.

On average, we expect 1/6 of the draws to equal 1, another 1/6 to equal 2, and so on. The following snippet is a quick demonstration. We simulate 6000 rolls of a die and expect about 1000 1's, 1000 2's, etc. We count how many we actually get. This snippet also introduces the `for` loop, which you should try to understand now because it will be *extremely* useful in the future.

```
> x <- sample(6, 6000, replace=T)

> for ( i in 1:6 ) print ( sum ( x==i ) )
[1] 995
[1] 1047
[1] 986
[1] 1033
[1] 975
[1] 964
>
```

Each number from 1 through 6 was chosen about 1000 times, plus or minus a little bit due to chance variation.

Now let's get back to craps. We want to simulate a large number of games, say 1000. For each game, we record either 1 or 0, according to whether the shooter wins on the Come-out roll, or not. We should print out the number of wins at the end. So we start with a code snippet like this:

```
# make a vector of length 1000, filled with 0's
wins <- rep ( 0, 1000 )
for ( i in 1:1000 ) {
  simulate a Come-out roll
  if shooter wins on Come-out, wins[i] <- 1
}

sum ( wins ) # print the number of wins
```

Now we have to figure out how to simulate the Come-out roll and decide whether the shooter wins. Clearly, we begin by simulating the roll of two dice. So our snippet expands to

```
# make a vector of length 1000, filled with 0's
  wins <- rep ( 0, 1000 )
for ( i in 1:1000 ) {
  d <- sample ( 1:6, 2, replace=T )
  if ( sum(d) == 7 || sum(d) == 11 ) wins[i] <- 1
}
sum ( wins ) # print the number of wins
```

The “`||`” stands for “or”. So that line of code sets `wins[i] <- 1` if the sum of the rolls is either 7 or 11. When I ran this simulation R printed out 219. The calculation in Example 1.1 says we should expect around $(2/9) \times 1000 \approx 222$ wins. Our calculation and simulation agree about as well as can be expected from a simulation. Try it yourself a few times. You shouldn’t always get 219. But you should get around 222 plus or minus a little bit due to the randomness of the simulation.

Try out these R commands in the version of R installed on your computer. Make sure you understand them. If you don’t, print out the results. Try variations. Try any tricks you can think of to help you learn R.

1.2 Probability Densities

So far we have dealt with *discrete* probabilities, or the probabilities of at most a countably infinite number of outcomes. For discrete probabilities, X is usually a set of integers, either finite or infinite. Section 1.2 deals with the case where X is an interval, either of finite or infinite length. Some examples are

Medical trials the time until a patient experiences a relapse

Sports the length of a javelin throw

Ecology the lifetime of a tree

Manufacturing the diameter of a ball bearing

Computing the amount of time a Help Line customer spends on hold

Physics the time until a uranium atom decays

Oceanography the temperature of ocean water at a specified latitude, longitude and depth

Probabilities for such outcomes are called *continuous*. For example, let Y be the time a Help Line caller spends on hold. The random variable Y is often modelled with a density similar to that in Figure 1.1.

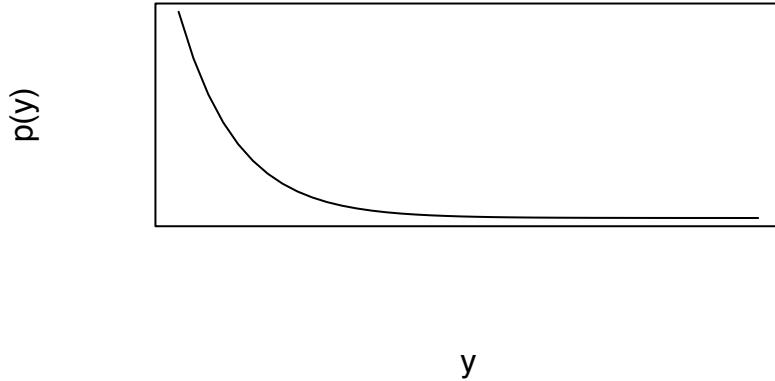


Figure 1.1: pdf for time on hold at Help Line

The curve in the figure is a *probability density function* or *pdf*. The pdf is large near $y = 0$ and monotonically decreasing, expressing the idea that smaller values of y are more likely than larger values. (Reasonable people may disagree about whether this pdf accurately represents callers' experience.) We typically use the symbols p , π or f for pdf's. We would write $p(50)$, $\pi(50)$ or $f(50)$ to denote the height of the curve at $y = 50$. For a pdf, probability is the same as area under the curve. For example, the probability that a caller waits less than 60 minutes is

$$P[Y < 60] = \int_0^{60} p(t) dt.$$

Every pdf must satisfy two properties.

1. $p(y) \geq 0$ for all y .
2. $\int_{-\infty}^{\infty} p(y) dy = 1$.

The first property holds because, if $p(y) < 0$ on the interval (a, b) then $P[Y \in (a, b)] = \int_a^b p(y) dy < 0$; and we can't have probabilities less than 0. The second property holds because $P[Y \in (-\infty, \infty)] = \int_{-\infty}^{\infty} p(y) dy = 1$.

One peculiar fact about any continuous random variable Y is that $P[Y = a] = 0$, for every $a \in \mathbb{R}$. That's because

$$P[Y = a] = \lim_{\epsilon \rightarrow 0} P[Y \in [a, a + \epsilon]] = \lim_{\epsilon \rightarrow 0} \int_a^{a+\epsilon} p_Y(y) dy = 0.$$

Consequently, for any numbers $a < b$,

$$P[Y \in (a, b)] = P[Y \in [a, b]] = P[Y \in (a, b]] = P[Y \in [a, b]].$$

The use of “density” in statistics is entirely analogous to its use in physics. In both fields

$$\text{density} = \frac{\text{mass}}{\text{volume}} \quad (1.1)$$

In statistics, we interpret density as *probability density*, mass as *probability mass* and volume as *length of interval*. In both fields, if the density varies from place to place (In physics it would vary within an object; in statistics it would vary along the real line.) then the density at a particular location is the limit of Equation 1.1 as volume $\rightarrow 0$.

Probability density functions are derivatives of probabilities. For any fixed number a

$$\frac{d}{db} P[X \in (a, b]] = \frac{d}{db} \int_a^b f_X(x) dx = f_X(b). \quad (1.2)$$

Similarly, $d/da P[X \in (a, b]] = -f_X(a)$.

Sometimes we can specify pdf's for continuous random variables based on the logic of the situation, just as we could specify discrete probabilities based on the logic of dice rolls. For example, let Y be the outcome of a spinner that is marked from 0 to 1. Then Y will be somewhere in the unit interval, and all parts of the interval are equally likely. So the pdf p_Y must look like Figure 1.2.

Figure 1.2 was produced by the following snippet.

```
plot ( c(0,1), c(1,1), xlab="y", ylab="p(y)",
      ylim=c(0,1.1), type="l" )
```

Note:

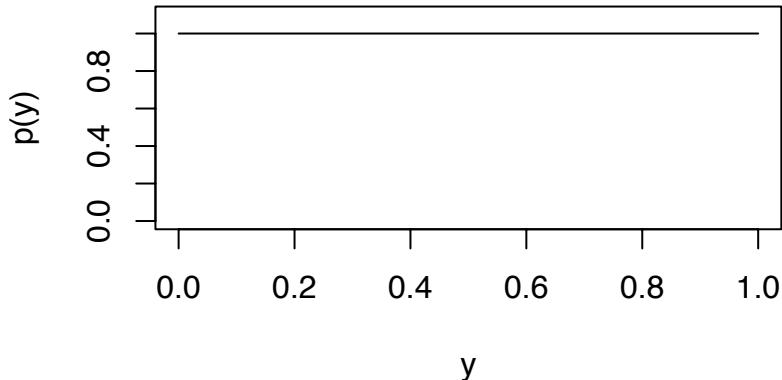


Figure 1.2: p_Y for the outcome of a spinner

- `c(0, 1)` collects 0 and 1 and puts them into the vector `(0,1)`. Likewise, `c(1, 1)` creates the vector `(1,1)`.
- `plot(x, y, ...)` produces a plot. The `plot(c(0, 1), c(1, 1), ...)` command above plots the points `(x[1], y[1]) = (0,1)` and `(x[2], y[2]) = (1,1)`.
- `type="l"` says to plot a line instead of individual points.
- `xlab` and `ylab` say how the axes are labelled.
- `ylim=c(0, 1.1)` sets the limits of the y-axis on the plot. If `ylim` is not specified then R sets the limits automatically. Limits on the x-axis can be specified with `xlim`.

At other times we use probability densities and distributions as models for data, and estimate the densities and distributions directly from the data. Figure 1.3 shows how that works. The upper panel of the figure is a histogram of 112 measurements of ocean temperature at a depth of 1000 meters in the North Atlantic near 45° North latitude and 20° degrees West longitude. Example 1.5 will say more about the data. Superimposed on the histogram is a pdf f . We think of f as underlying the data. The idea is that measuring a temperature at that location is like randomly drawing a value from f . The 112 measure-

ments, which are spread out over about a century of time, are like 112 independent draws from f . Having the 112 measurements allows us to make a good estimate of f . If oceanographers return to that location to make additional measurements, it would be like making additional draws from f . Because we can estimate f reasonably well, we can predict with some degree of assurance what the future draws will be like.

The bottom panel of Figure 1.3 is a histogram of the `discoveries` data set that comes with R and which is, as R explains, “The numbers of ‘great’ inventions and scientific discoveries in each year from 1860 to 1959.” It is overlaid with a line showing the $\text{Poi}(3.1)$ distribution. (Named distributions will be introduced in Section 1.3.) It seems that the number of great discoveries each year follows the $\text{Poi}(3.1)$ distribution, at least approximately. If we think the future will be like the past then we should expect future years to follow a similar pattern. Again, we think of a distribution underlying the data. The number of discoveries in a single year is like a draw from the underlying distribution. The figure shows 100 years, which allow us to estimate the underlying distribution reasonably well.

Figure 1.3 was produced by the following snippet.

```
par ( mfrow=c(2,1) )

good <- abs ( med.1000$lon + 20 ) < 1 &
           abs ( med.1000$lat - 45 ) < 1
hist ( med.1000$temp[good], xlab="temperature", ylab="",
       main="", prob=T, xlim=c(5,11) )
m <- mean ( med.1000$temp[good] )
s <- sqrt ( var ( med.1000$temp[good] ) )
x <- seq ( 5, 11, length=40 )
lines ( density(med.1000$temp[good]) )

hist ( discoveries, xlab="discoveries", ylab="", main="",
       prob=T, breaks=seq(-.5,12.5,by=1) )
lines ( 0:12, dpois(0:12, 3.1), type="b" )
```

Note:

- `par` sets R’s *graphical parameters*. `mfrow=c(2,1)` tells R to make an array of multiple figures in a 2 by 1 layout.
- `med.1000` is a data set of North Atlantic ocean temperatures at a depth of 1000 meters. `med.1000$lon` and `med.1000$lat` are the longitude and latitude of the measurements. `med.1000$temp` are the actual temperatures.

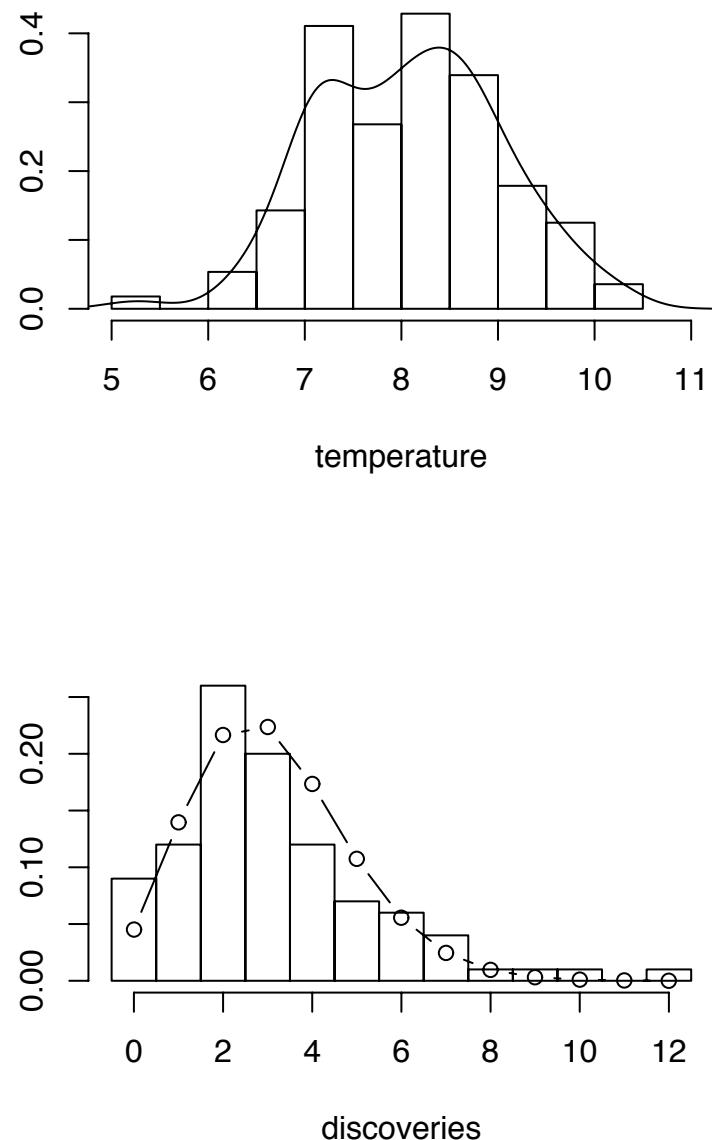


Figure 1.3: (a): Ocean temperatures at 1000m depth near 45°N latitude, -20° longitude;
(b) Numbers of important discoveries each year 1860–1959

- `abs` stands for absolute value.
- `good <- ...` calls those points *good* whose longitude is between -19 and -21 and whose latitude is between 44 and 46.
- `hist()` makes a histogram. `prob=T` turns the y-axis into a probability scale (area under the histogram is 1) instead of counts.
- `mean()` calculates the mean. `var()` calculates the variance. Section 1.4 defines the mean and variance of distributions. Section 2.2.1 defines the mean and variance of data sets.
- `lines()` adds lines to an existing plot.
- `density()` estimates a density from a data set.

It is often necessary to transform one variable into another as, for example, $Z = g(X)$ for some specified function g . We might know p_X (The subscript indicates which random variable we're talking about.) and want to calculate p_Z . Here we consider only monotonic functions g , so there is an inverse $X = h(Z)$.

Theorem 1.1. *Let X be a random variable with pdf p_X . Let g be a differentiable, monotonic, invertible function and define $Z = g(X)$. Then the pdf of Z is*

$$p_Z(t) = p_X(g^{-1}(t)) \left| \frac{d g^{-1}(t)}{d t} \right|$$

Proof. If g is an increasing function then

$$\begin{aligned} p_Z(b) &= \frac{d}{db} P[Z \in (a, b)] = \frac{d}{db} P[X \in (g^{-1}(a), g^{-1}(b))] \\ &= \frac{d}{db} \int_{g^{-1}(a)}^{g^{-1}(b)} p_X(x) dx = \left. \frac{d g^{-1}(t)}{d t} \right|_a^b \times p_X(g^{-1}(b)) \end{aligned}$$

The proof when g is decreasing is left as an exercise. \square

To illustrate, suppose that X is a random variable with pdf $p_X(x) = 2x$ on the unit interval. Let $Z = 1/X$. What is $p_Z(z)$? The inverse transformation is $X = 1/Z$. Its derivative is $dx/dz = -z^{-2}$. Therefore,

$$p_Z(z) = p_X(g^{-1}(z)) \left| \frac{d g^{-1}(z)}{d z} \right| = \frac{2}{z} \left| -\frac{1}{z^2} \right| = \frac{2}{z^3}$$

And the possible values of Z are from 1 to ∞ . So $p_Z(z) = 2/z^3$ on the interval $(1, \infty)$. As a partial check, we can verify that the integral is 1.

$$\int_1^\infty \frac{2}{z^3} dz = -\frac{1}{z^2} \Big|_1^\infty = 1.$$

Theorem 1.1 can be explained by Figure 1.4. The figure shows an x , a z , and the function $z = g(x)$. A little interval is shown around x ; call it I_x . It gets mapped by g into a little interval around z ; call it I_z . The density is

$$p_Z(z) \approx \frac{P[Z \in I_z]}{\text{length}(I_z)} = \frac{P[X \in I_x]}{\text{length}(I_x)} \frac{\text{length}(I_x)}{\text{length}(I_z)} \approx p_X(x) |h'(z)| \quad (1.3)$$

The approximations in Equation 1.3 are exact as the lengths of I_x and I_z decrease to 0.

If g is not one-to-one, then it is often possible to find subsets of \mathbb{R} on which g is one-to-one, and work separately on each subset.

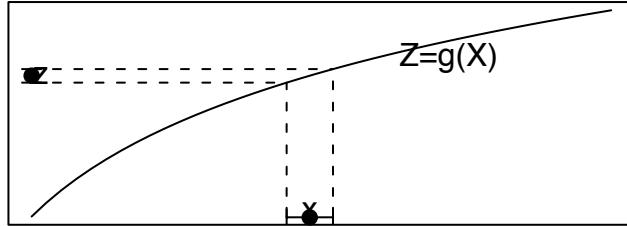


Figure 1.4: Change of variables

1.3 Parametric Families of Distributions

Probabilities often depend on one or more unknown numerical constants. Suppose, for example, that we have a biased coin. Let θ be the chance that it lands H. Then $P(H) = \theta$; but we might not know θ ; it is an unknown numerical constant. In this case we have a family of probability measures, one for each value of θ , and we don't know which one is

right. When we need to be explicit that probabilities depend on θ , we use the notation, for example, $P(H|\theta)$ or $P(H|\theta = 1/3)$. The vertical bar is read “given” or “given that”. So $P(H|\theta = 1/3)$ is read “the probability of Heads given that θ equals $1/3$ ” and $P(H|\theta)$ is read “the probability of Heads given θ .” This notation means

$$\begin{aligned} P(H|\theta = 1/3) &= 1/3, \\ P(T|\theta = 1/3) &= 2/3, \\ P(T|\theta = 1/5) &= 4/5 \end{aligned}$$

and so on. Instead of “given” we also use the word “conditional”. So we would say “the probability of Heads conditional on θ ”, etc.

The unknown constant θ is called a *parameter*. The set of possible values for θ is denoted Θ (upper case θ). For each θ there is a probability measure μ_θ . The set of all possible probability measures (for the problem at hand),

$$\{\mu_\theta : \theta \in \Theta\},$$

is called a *parametric family* of probability measures. The rest of this chapter introduces four of the most useful parametric families of probability measures.

1.3.1 The Binomial Distribution

Statisticians often have to consider observations of the following type.

- A repeatable event results in either a success or a failure.
- Many repetitions are observed.
- Successes and failures are counted.
- The number of successes helps us learn about the probability of success.

Such observations are called *binomial*. Some examples are

Medical trials A new treatment is given to many patients. Each is either cured or not.

Toxicity tests Many laboratory animals are exposed to a potential carcinogen. Each either develops cancer or not.

Ecology Many seeds are planted. Each either germinates or not.

Quality control Many supposedly identical items are subjected to a test. Each either passes or not.

Because binomial experiments are so prevalent there is specialized language to describe them. Each repetition is called a *trial*; the number of trials is usually denoted N ; the unknown probability of success is usually denoted either p or θ ; the number of successes is usually denoted X . We write $X \sim \text{Bin}(N, p)$. The symbol “ \sim ” is read *is distributed as*; we would say “ X is distributed as Binomial N p ” or “ X has the Binomial N, p distribution”. Some important assumptions about binomial experiments are that N is fixed in advance, θ is the same for every trial, and the outcome of any trial does not influence the outcome of any other trial. When $N = 1$ we say X has a Bernoulli(θ) distribution and write $X \sim \text{Bern}(\theta)$; the individual trials in a binomial experiment are called Bernoulli trials.

When a binomial experiment is performed, X will turn out to be one of the integers from 0 to N . We need to know the associated probabilities; i.e. $P[X = k | \theta]$ for each value of k from 0 to N . These probabilities are given by Equation 1.4 whose derivation is given in Section 5.1.

$$P[X = k | \theta] = \binom{N}{k} \theta^k (1 - \theta)^{N-k} \quad (1.4)$$

The term $\binom{N}{k}$ is called a *binomial coefficient* and is read “ N choose k ”. $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ and is equal to the number of subsets of size k that can be formed from a group of N distinct items. In case $k = 0$ or $k = N$, 0! is defined to be 1. Figure 1.5 shows binomial probabilities for $N \in \{3, 30, 300\}$ and $p \in \{.1, .5, .9\}$.

Example 1.3 (Craps, continued)

This example continues the game of craps. See Examples 1.1 and 1.2.

What is the probability that at least one of the next four players wins on his Come-out roll?

This is a Binomial experiment because

1. We are looking at repeated trials. Each Come-out roll is a trial. It results in either success, or not.
2. The outcome of one trial does not affect the other trials.
3. We are counting the number of successes.

Let X be the number of successes. There are four trials, so $N = 4$. We calculated the probability of success in Example 1.1; it's $p = 2/9$. So $X \sim \text{Bin}(4, 2/9)$. The probability of success in at least one Come-out roll is

$$\begin{aligned} P[\text{success in at least one Come-out roll}] &= P[X \geq 1] \\ &= \sum_{i=1}^4 P[X = i] = \sum_{i=1}^4 \binom{4}{i} (2/9)^i (7/9)^{4-i} \approx 0.634 \end{aligned} \quad (1.5)$$

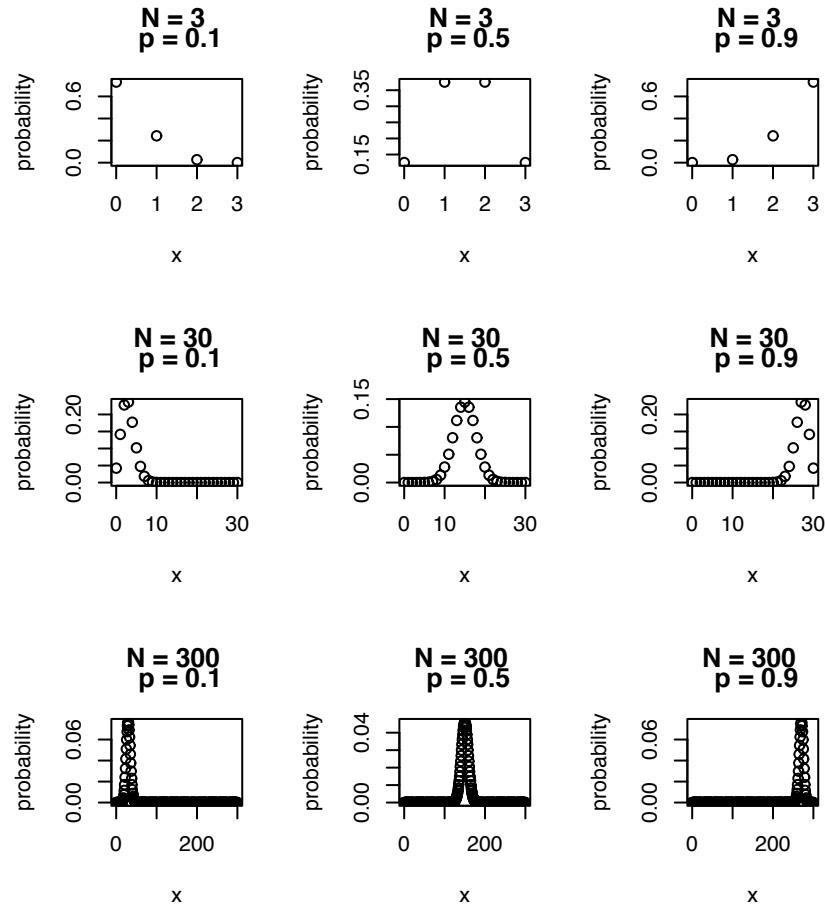


Figure 1.5: Binomial probabilities

A convenient way to re-express Equation 1.5 is

$$P[X \geq 1] = 1 - P[X = 0],$$

which can be quickly calculated in R. The `dbinom()` command computes Binomial probabilities. To compute Equation 1.5 we would write

```
1 - dbinom(0, 4, 2/9)
```

The `0` says what value of X we want. The `4` and the `2/9` are the number of trials and the probability of success. **Try it. Learn it.**

1.3.2 The Poisson Distribution

Another common type of observation occurs in the following situation.

- There is a domain of study, usually a block of space or time.
- Events arise seemingly at random in the domain.
- There is an underlying rate at which events arise.

Such observations are called *Poisson* after the 19th century French mathematician Siméon-Denis Poisson. The number of events in the domain of study helps us learn about the rate. Some examples are

Ecology Tree seedlings emerge from the forest floor.

Computer programming Bugs occur in computer code.

Quality control Defects occur along a strand of yarn.

Genetics Mutations occur in a genome.

Traffic flow Cars arrive at an intersection.

Customer service Customers arrive at a service counter.

Neurobiology Neurons fire.

The rate at which events occur is often called λ ; the number of events that occur in the domain of study is often called X ; we write $X \sim \text{Poi}(\lambda)$. Important assumptions about Poisson observations are that two events cannot occur at exactly the same location in space or time, that the occurrence of an event at location ℓ_1 does not influence whether an event occurs at any other location ℓ_2 , and the rate at which events arise does not vary over the domain of study.

When a Poisson experiment is observed, X will turn out to be a nonnegative integer. The associated probabilities are given by Equation 1.6.

$$P[X = k | \lambda] = \frac{\lambda^k e^{-\lambda}}{k!}. \quad (1.6)$$

One of the main themes of statistics is the quantitative way in which data help us learn about the phenomenon we are studying. Example 1.4 shows how this works when we want to learn about the rate λ of a Poisson distribution.

Example 1.4 (Seedlings in a Forest)

Tree populations move by dispersing their seeds. Seeds become seedlings, seedlings become saplings, and saplings become adults which eventually produce more seeds. Over time, whole populations may migrate in response to climate change. One instance occurred at the end of the Ice Age when species that had been sequestered in the south were free to move north. Another instance may be occurring today in response to global warming. One critical feature of the migration is its speed. Some of the factors determining the speed are the typical distances of long range seed dispersal, the proportion of seeds that germinate and emerge from the forest floor to become seedlings, and the proportion of seedlings that survive each year.

To learn about emergence and survival, ecologists return annually to forest quadrats (square meter sites) to count seedlings that have emerged since the previous year. One such study was reported in LAVINE ET AL. [2002]. A fundamental quantity of interest is the rate λ at which seedlings emerge. Suppose that, in one quadrat, three new seedlings are observed. What does that say about λ ?

Different values of λ yield different values of $P[X = 3 | \lambda]$. To compare different values of λ we see how well each one explains the data $X = 3$; i.e., we compare

$P[X = 3 | \lambda]$ for different values of λ . For example,

$$P[X = 3 | \lambda = 1] = \frac{1^3 e^{-1}}{3!} \approx 0.06$$

$$P[X = 3 | \lambda = 2] = \frac{2^3 e^{-2}}{3!} \approx 0.18$$

$$P[X = 3 | \lambda = 3] = \frac{3^3 e^{-3}}{3!} \approx 0.22$$

$$P[X = 3 | \lambda = 4] = \frac{4^3 e^{-4}}{3!} \approx 0.14$$

In other words, the value $\lambda = 3$ explains the data almost four times as well as the value $\lambda = 1$ and just a little bit better than the values $\lambda = 2$ and $\lambda = 4$. Figure 1.6 shows $P[X = 3 | \lambda]$ plotted as a function of λ . The figure suggests that $P[X = 3 | \lambda]$ is maximized by $\lambda = 3$. The suggestion can be verified by differentiating Equation 1.6 with respect to lambda, equating to 0, and solving. The figure also shows that any value of λ from about 0.5 to about 9 explains the data not too much worse than $\lambda = 3$.

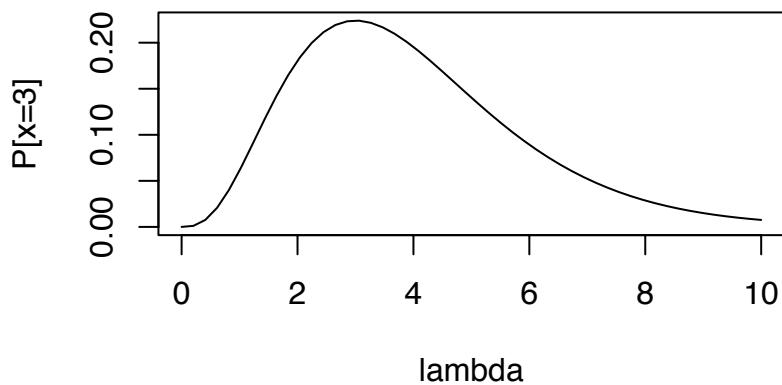


Figure 1.6: $P[X = 3 | \lambda]$ as a function of λ

Figure 1.6 was produced by the following snippet.

```
lam <- seq ( 0, 10, length=50 )
y <- dpois ( 3, lam )
plot ( lam, y, xlab="lambda", ylab="P[x=3]", type="l" )
```

Note:

- `seq` stands for “sequence”. `seq(0,10,length=50)` produces a sequence of 50 numbers evenly spaced from 0 to 10.
- `dpois` calculates probabilities for Poisson distributions the way `dbinom` does for Binomial distributions.
- `plot` produces a plot. In the `plot(...)` command above, `lam` goes on the x-axis, `y` goes on the y-axis, `xlab` and `ylab` say how the axes are labelled, and `type="l"` says to plot a line instead of individual points.

Making and interpreting plots is a big part of statistics. Figure 1.6 is a good example. Just by looking at the figure we were able to tell which values of λ are plausible and which are not. Most of the figures in this book were produced in R.

1.3.3 The Exponential Distribution

It is often necessary to model a continuous random variable X whose density decreases away from 0. Some examples are

Customer service time on hold at a help line

Neurobiology time until the next neuron fires

Seismology time until the next earthquake

Medicine remaining years of life for a cancer patient

Ecology dispersal distance of a seed

In these examples it is expected that most calls, times or distances will be short and a few will be long. So the density should be large near $x = 0$ and decreasing as x increases.

A useful pdf for such situations is the *Exponential density*

$$p(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \quad \text{for } x > 0. \quad (1.7)$$

We say X has an exponential distribution with parameter λ and write $X \sim \text{Exp}(\lambda)$. Figure 1.7 shows exponential densities for several different values of λ .

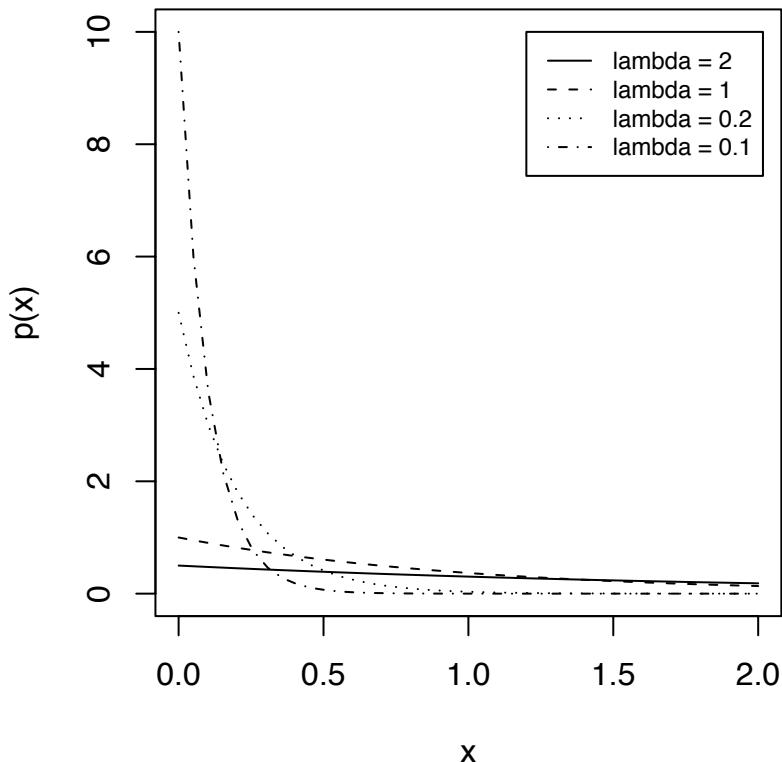


Figure 1.7: Exponential densities

Figure 1.7 was produced by the following snippet.

```
x <- seq(0, 2, length=40) # 40 values from 0 to 2
lam <- c(2, 1, .2, .1)    # 4 different values of lambda
y <- matrix(NA, 40, 4)    # y values for plotting
```

```

for ( i in 1:4 )
  y[,i] <- dexp ( x, 1/lam[i] ) # exponential pdf
  matplot ( x, y, type="l", xlab="x", ylab="p(x)", col=1 )
  legend ( 1.2, 10, paste ( "lambda =", lam ),
            lty=1:4, cex=.75 )

```

- We want to plot the exponential density for several different values of λ so we choose 40 values of x between 0 and 2 at which to do the plotting.
- Next we choose 4 values of λ .
- Then we need to calculate and save $p(x)$ for each combination of x and λ . We'll save them in a matrix called y . `matrix(NA, 40, 4)` creates the matrix. The size of the matrix is 40 by 4. It is filled with `NA`, or *Not Available*.
- `dexp` calculates the exponential pdf. The argument `x` tells R the x values at which to calculate the pdf. `x` can be a vector. The argument `1/lam[i]` tells R the value of the parameter. R uses a different notation than this book. Where this book says `Exp(2)`, R says `Exp(.5)`. That's the reason for the `1/lam[i]`.
- `matplot` plots one matrix versus another. The first matrix is `x` and the second is `y`. `matplot` plots each column of `y` against each column of `x`. In our case `x` is vector, so `matplot` plots each column of `y`, in turn, against `x`. `type="l"` says to plot lines instead of points. `col=1` says to use the first color in R's library of colors.
- `legend (. . .)` puts a legend on the plot. The `1.2` and `10` are the x and y coordinates of the upper left corner of the legend box. `lty=1:4` says to use line types 1 through 4. `cex=.75` sets the *character expansion factor* to .75. In other words, it sets the font size.
- `paste(. . .)` creates the words that go into the legend. It pastes together "lambda =" with the four values of `lam`.

1.3.4 The Normal Distribution

It is often necessary to model a continuous random variable Y whose density is mound-shaped. Some examples are

Biological Anthropology Heights of people

Oceanography Ocean temperatures at a particular location

Quality Control Diameters of ball bearings

Education SAT scores

In each case the random variable is expected to have a central value around which most of the observations cluster. Fewer and fewer observations are farther and farther away from the center. So the pdf should be unimodal — large in the center and decreasing in both directions away from the center. A useful pdf for such situations is the *Normal density*

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}. \quad (1.8)$$

We say Y has a *Normal distribution with mean μ and standard deviation σ* and write $Y \sim N(\mu, \sigma)$. Figure 1.8 shows Normal densities for several different values of (μ, σ) . As illustrated by the figure, μ controls the center of the density; each pdf is centered over its own value of μ . On the other hand, σ controls the spread. pdf's with larger values of σ are more spread out; pdf's with smaller σ are tighter.

Figure 1.8 was produced by the following snippet.

```
x <- seq( -6, 6, len=100 )
y <- cbind( dnorm( x, -2, 1 ),
            dnorm( x, 0, 2 ),
            dnorm( x, 0, .5 ),
            dnorm( x, 2, .3 ),
            dnorm( x, -.5, 3 )
)
matplot( x, y, type="l", col=1 )
legend( -6, 1.3, paste( "mu =", c(-2,0,0,2,-.5),
                       "; sigma =", c(1,2,.5,.3,3) ),
        lty=1:5, col=1, cex=.75 )
```

- `dnorm(...)` computes the Normal pdf. The first argument is the set of x values; the second argument is the mean; the third argument is the standard deviation.

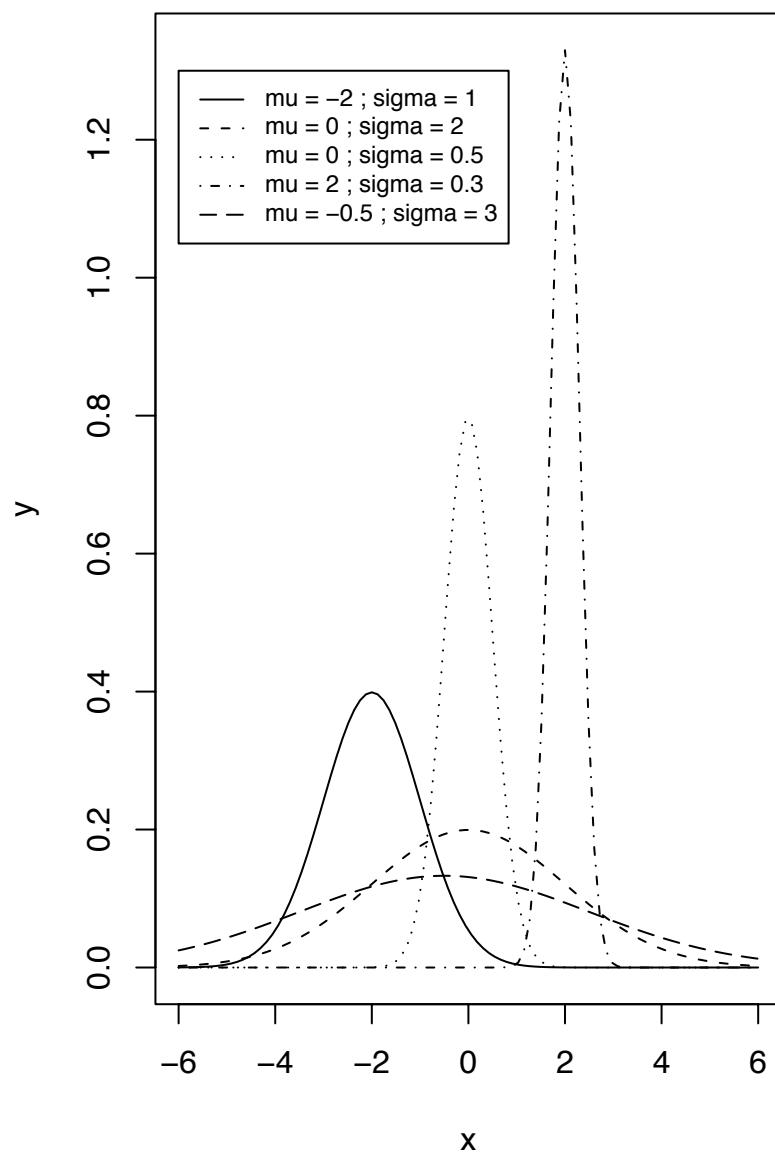


Figure 1.8: Normal densities

As a further illustration, Figure 1.9 shows a histogram of 105 ocean temperatures ($^{\circ}\text{C}$) recorded in the Atlantic Ocean from about 1938 to 1997 at a depth of 1000 meters, near 45 degrees North latitude and 30 degrees West longitude. The $\text{N}(5.87, .72)$ density is superimposed on the histogram. The Normal density represents the data moderately well. We will study ocean temperatures in much more detail in a series of examples beginning with Example 1.5.

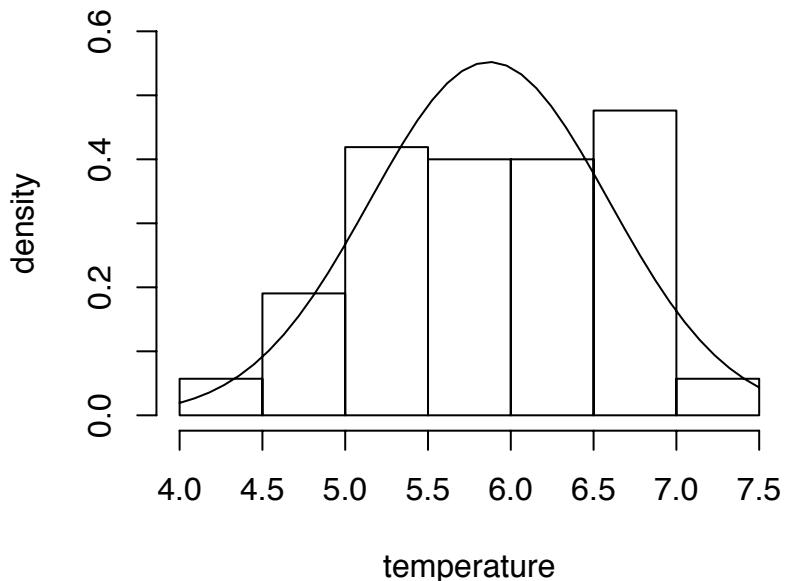


Figure 1.9: Ocean temperatures at 45°N , 30°W , 1000m depth. The $\text{N}(5.87, .72)$ density.

Figure 1.9 was produced by

```
hist ( y, prob=T, xlab="temperature", ylab="density",
      ylim=c(0,.6), main="" )
t <- seq ( 4, 7.5, length=40 )
```

```
lines ( t, dnorm ( t, mean(y), sd(y) ) )
```

- The 105 temperatures are in a vector `y`.
- `hist` produces a histogram. The argument `prob=T` causes the vertical scale to be probability density instead of counts.
- The line `t <- ...` sets 40 values of `t` in the interval [4, 7.5] at which to evaluate the Normal density for plotting purposes.
- `lines` displays the Normal density.

As usual, you should try to understand the R commands.

The function `rnorm(n, mu, sig)` generates a random sample from a Normal distribution. `n` is the sample size; `mu` is the mean; and `sig` is the standard deviation. To demonstrate, we'll generate a sample of size 100 from the $N(5.87, .72)$ density, the density in Figure 1.9, and compare the sample histogram to the theoretical density. Figure 1.10(a) shows the comparison. It shows about how good a fit can be expected between a histogram and the Normal density, for a sample of size around 100 in the most ideal case when the sample was actually generated from the Normal distribution. It is interesting to consider whether the fit in Figure 1.9 is much worse.

Figure 1.10(a) was produced by

```
samp <- rnorm ( 100, 5.87, .72 )
y.vals <- seq ( 4, 7.5, length=40 )
hist ( samp, prob=T, main="(a)",
       xlim=c(4,7.5), xlab="degrees C",
       ylim=c(0,.6), ylab="density" )
lines ( y.vals, dnorm(y.vals,5.87,.72) )
```

When working with Normal distributions it is *extremely* useful to think in terms of units of standard deviation, or simply *standard units*. One standard unit equals one standard deviation. In Figure 1.10(a) the number 6.6 is about 1 standard unit above the mean, while the number 4.5 is about 2 standard units below the mean. To see why that's a useful way to think, Figure 1.10(b) takes the sample from Figure 1.10(a), multiplies by

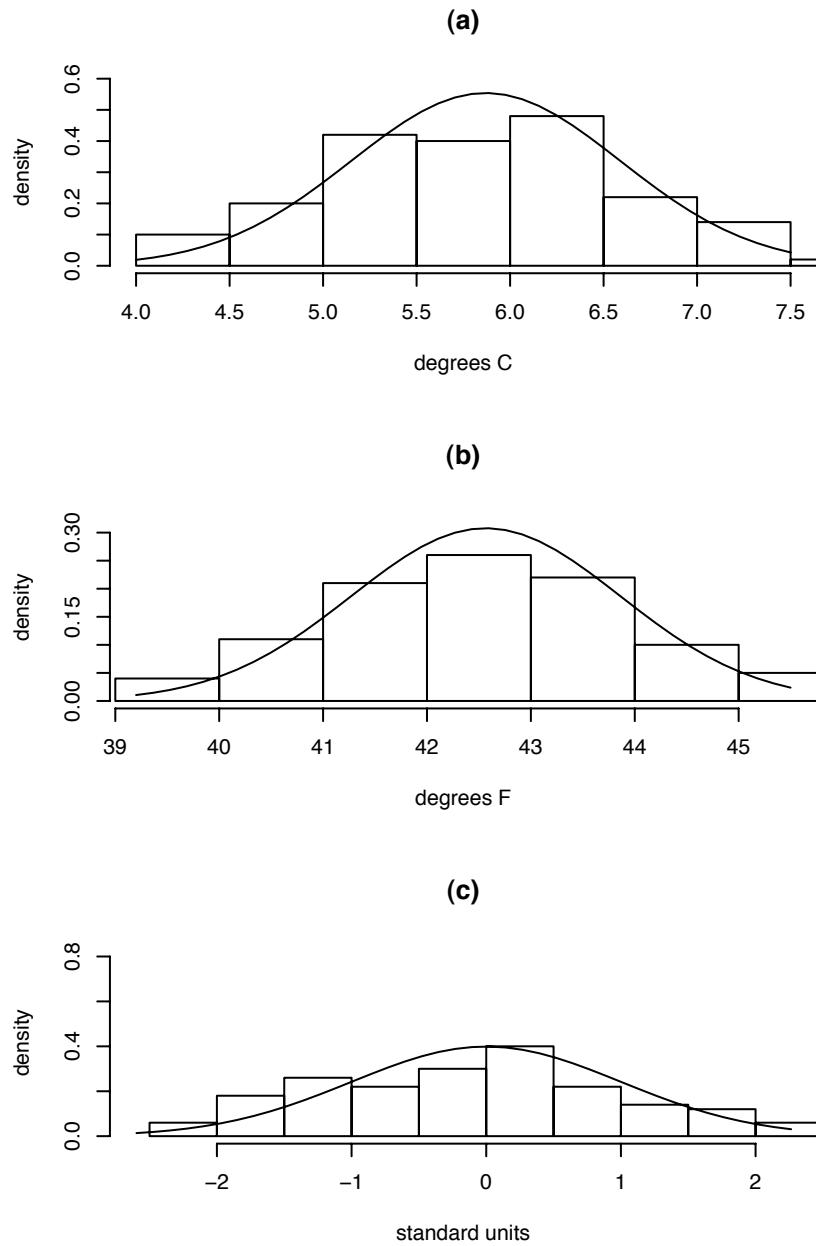


Figure 1.10: **(a)**: A sample of size 100 from $N(5.87, .72)$ and the $N(5.87, .72)$ density. **(b)**: A sample of size 100 from $N(42.566, 1.296)$ and the $N(42.566, 1.296)$ density. **(c)**: A sample of size 100 from $N(0, 1)$ and the $N(0, 1)$ density.

$9/5$ and adds 32, to simulate temperatures measured in °F instead of °C. The histograms in panels **(a)** and **(b)** are slightly different because R has chosen the bin boundaries differently; but the two Normal curves have identical shapes. Now consider some temperatures, say $6.5^\circ\text{C} = 43.7^\circ\text{F}$ and $4.5^\circ\text{C} = 40.1^\circ\text{F}$. Corresponding temperatures occupy corresponding points on the plots. A vertical line at 6.5 in panel **(a)** divides the density into two sections exactly congruent to the two sections created by a vertical line at 43.7 in panel **(b)**. A similar statement holds for 4.5 and 40.1. The point is that the two density curves have exactly the same shape. They are identical except for the scale on the horizontal axis, and that scale is determined by the standard deviation. Standard units are a scale-free way of thinking about the picture.

To continue, we converted the temperatures in panels **(a)** and **(b)** to standard units, and plotted them in panel **(c)**. Once again, R made a slightly different choice for the bin boundaries, but the Normal curves all have the same shape.

Panels **(b)** and **(c)** of Figure 1.10 were produced by

```
y2samp <- samp * 9/5 + 32
y2.vals <- y.vals * 9/5 + 32
hist ( y2samp, prob=T, main="(b)",
       xlim=c(39.2,45.5), xlab="degrees F",
       ylim=c(0,1/3), ylab="density" )
lines ( y2.vals, dnorm(y2.vals,42.566,1.296) )

zsamp <- (samp-5.87) / .72
z.vals <- (y.vals-5.87) / .72
hist ( zsamp, prob=T, main="(c)",
       xlim=c(-2.6,2.26), xlab="standard units",
       ylim=c(0,.833), ylab="density" )
lines ( z.vals, dnorm(z.vals,0,1) )
```

Let $Y \sim N(\mu, \sigma)$ and define a new random variable $Z = (Y - \mu)/\sigma$. Z is in standard units. It tells how many standard units Y is above or below its mean μ . What is the distribution of Z ? The easiest way to find out is to calculate p_Z , the density of Z , and see whether we recognize it. From Theorem 1.1,

$$p_Z(z) = p_Y(\sigma z + \mu)\sigma = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

which we recognize as the $N(0, 1)$ density. I.e., $Z \sim N(0, 1)$. The $N(0, 1)$ distribution is called the *standard Normal* distribution.

1.4 Centers, Spreads, Means, and Moments

Recall Figure 1.3 (pg. 11). In each panel there is a histogram of a data set along with an estimate of the underlying pdf or pmf p . In each case we have found a distribution that matches the data reasonably well, but the distributions we have drawn are not the only ones that match well. We could make modest changes to either distribution and still have a reasonably good match. But whatever pdf we propose for the top panel should be roughly mound shaped with a center around 8° and a spread that ranges from about 6° to about 10° . And in the bottom panel we would want a distribution with a peak around 2 or 3 and a longish right hand tail.

In either case, the details of the distribution matter less than these central features. So statisticians often need to refer to the center, or location, of a sample or a distribution and also to its spread. Section 1.4 gives some of the theoretical underpinnings for talking about centers and spreads of distributions.

Example 1.5

Physical oceanographers study physical properties such as temperature, salinity, pressure, oxygen concentration, and potential vorticity of the world's oceans. Data about the oceans' surface can be collected by satellites' bouncing signals off the surface. But satellites cannot collect data about deep ocean water. Until as recently as the 1970s, the main source of data about deep water came from ships that lower instruments to various depths to record properties of ocean water such as temperature, pressure, salinity, etc. (Since about the 1970s oceanographers have begun to employ neutrally buoyant floats. A brief description and history of the floats can be found on the web at www.soc.soton.ac.uk/JRD/HYDRO/SHB/FLOAT.HISTORY.HTML.) Figure 1.11 shows locations, called *hydrographic stations*, off the coast of Europe and Africa where ship-based measurements were taken between about 1910 and 1990. The outline of the continents is apparent on the right-hand side of the figure due to the lack of measurements over land.

Deep ocean currents cannot be seen but can be inferred from physical properties. Figure 1.12 shows temperatures recorded over time at a depth of 1000 meters at nine different locations. The upper right panel in Figure 1.12 is the same as the top panel of Figure 1.3. Each histogram in Figure 1.12 has a black circle indicating the "center" or "location" of the points that make up the histogram. These centers are good estimates of the centers of the underlying pdf's. The centers range from a low of about 5° at

latitude 45 and longitude -40 to a high of about 9° at latitude 35 and longitude -20. (By convention, longitudes to the west of Greenwich, England are negative; longitudes to the east of Greenwich are positive.) It's apparent from the centers that for each latitude, temperatures tend to get colder as we move from east to west. For each longitude, temperatures are warmest at the middle latitude and colder to the north and south. Data like these allow oceanographers to deduce the presence of a large outpouring of relatively warm water called the *Mediterranean tongue* from the Mediterranean Sea into the Atlantic ocean. The Mediterranean tongue is centered at about 1000 meters depth and 35°N latitude, flows from east to west, and is warmer than the surrounding Atlantic waters into which it flows.

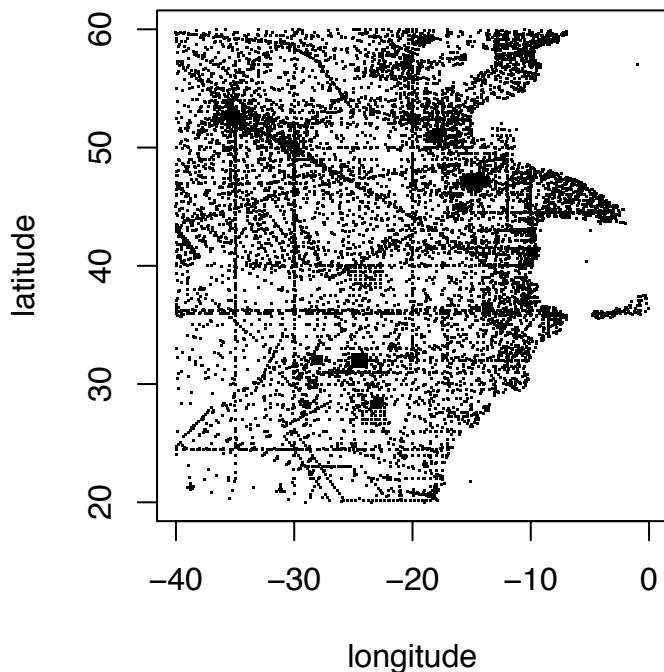


Figure 1.11: hydrographic stations off the coast of Europe and Africa

There are many ways of describing the center of a data sample. But by far the most

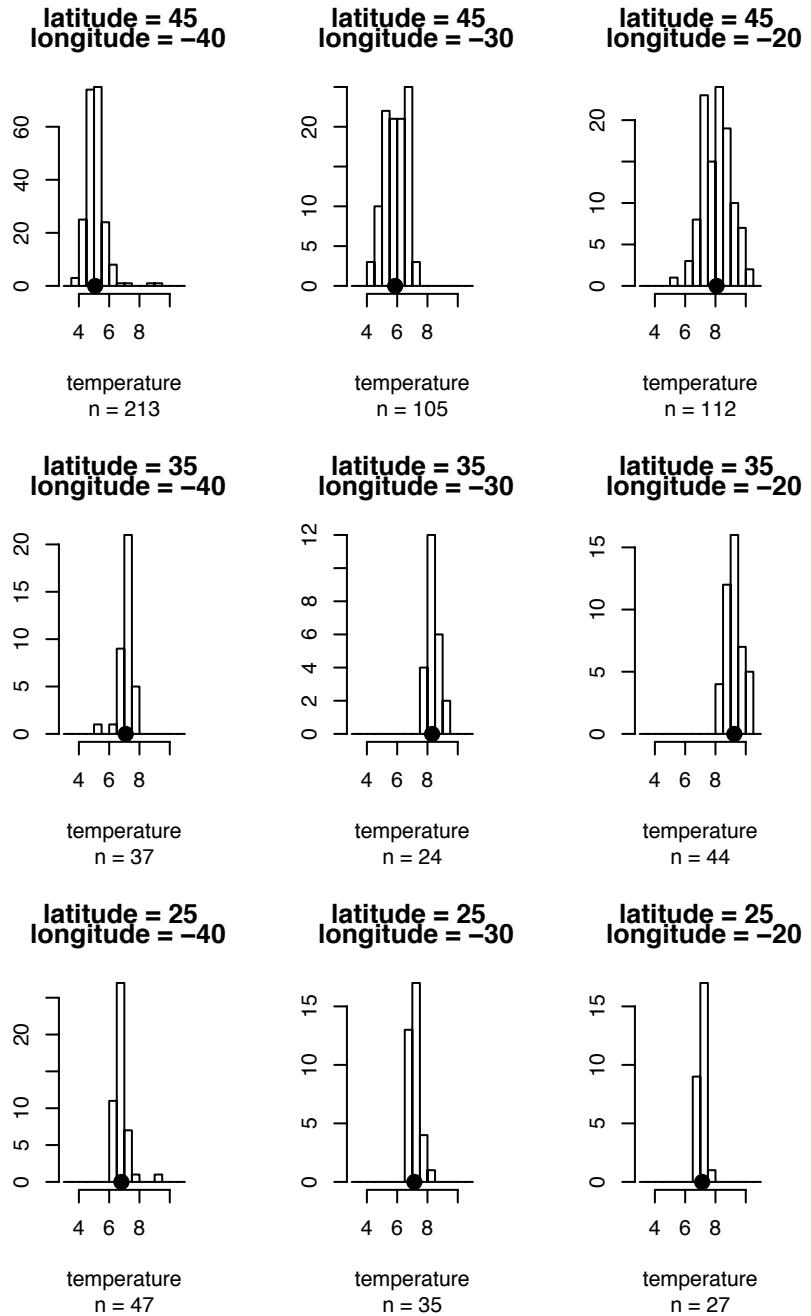


Figure 1.12: Water temperatures ($^{\circ}\text{C}$) at 1000m depth, latitude 25, 35, 45 degrees North longitude 20, 30, 40 degrees West

common is the mean. The *mean* of a sample, or of any list of numbers, is just the average.

Definition 1.1 (Mean of a sample). The *mean* of a sample, or any list of numbers, x_1, \dots, x_n is

$$\text{mean of } x_1, \dots, x_n = \frac{1}{n} \sum x_i. \quad (1.9)$$

The black circles in Figure 1.12 are means. The mean of x_1, \dots, x_n is often denoted \bar{x} . Means are often a good first step in describing data that are unimodal and roughly symmetric.

Similarly, means are often useful in describing distributions. For example, the mean of the pdf in the upper panel of Figure 1.3 is about 8.1, the same as the mean of the data in same panel. Similarly, in the bottom panel, the mean of the Poi(3.1) distribution is 3.1, the same as the mean of the *discoveries* data. Of course we chose the distributions to have means that matched the means of the data.

For some other examples, consider the $\text{Bin}(n, p)$ distributions shown in Figure 1.5. The center of the $\text{Bin}(30, .5)$ distribution appears to be around 15, the center of the $\text{Bin}(300, .9)$ distribution appears to be around 270, and so on. The mean of a distribution, or of a random variable, is also called the *expected value* or *expectation* and is written $\mathbb{E}(X)$.

Definition 1.2 (Mean of a random variable). Let X be a random variable with cdf F_X and pdf p_X . Then the *mean* of X (equivalently, the *mean* of F_X) is

$$\mathbb{E}(X) = \begin{cases} \sum_i i P[X = i] & \text{if } X \text{ is discrete} \\ \int x p_X(x) dx & \text{if } X \text{ is continuous} \end{cases} \quad (1.10)$$

The logic of the definition is that $\mathbb{E}(X)$ is a weighted average of the possible values of X . Each value is weighted by its importance, or probability. In addition to $\mathbb{E}(X)$, another common notation for the mean of a random variable X is μ_X .

Let's look at some of the families of probability distributions that we have already studied and calculate their expectations.

Binomial If $X \sim \text{Bin}(n, p)$ then

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{i=0}^n i P[x = i] \\
&= \sum_{i=0}^n i \binom{n}{i} p^i (1-p)^{n-i} \\
&= \sum_{i=1}^n i \binom{n}{i} p^i (1-p)^{n-i} \\
&= np \sum_{i=1}^n \frac{(n-1)!}{(i-1)!(n-i)!} p^{i-1} (1-p)^{n-i} \\
&= np \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!} p^j (1-p)^{n-1-j} \\
&= np
\end{aligned} \tag{1.11}$$

The first five equalities are just algebra. The sixth is worth remembering. The sum $\sum_{j=0}^{n-1} \dots$ is the sum of the probabilities of the $\text{Bin}(n-1, p)$ distribution. Therefore the sum is equal to 1. You may wish to compare $\mathbb{E}(X)$ to Figure 1.5.

Poisson If $X \sim \text{Poi}(\lambda)$ then

$$\mathbb{E}(X) = \lambda.$$

The derivation is left as Exercise 18.

Exponential If $X \sim \text{Exp}(\lambda)$ then

$$\mathbb{E}(X) = \int_0^\infty x p(x) dx = \lambda^{-1} \int_0^\infty x e^{-x/\lambda} dx = \lambda.$$

Use integration by parts.

Normal If $X \sim \text{N}(\mu, \sigma)$ then $\mathbb{E}(X) = \mu$. The derivation is left as Exercise 18.

Statisticians also need to measure and describe the spread of distributions, random variables and samples. In Figure 1.12, the spread would measure how much variation there is in ocean temperatures at a single location, which in turn would tell us something about how heat moves from place to place in the ocean. Spread could also describe the variation in the annual numbers of “great” discoveries, the range of typical outcomes for a gambler playing a game repeatedly at a casino or an investor in the stock market, or the

uncertain effect of a change in the Federal Reserve Bank's monetary policy, or even why different patches of the same forest have different plants on them.

By far the most common measures of spread are the *variance* and its square root, the *standard deviation*.

Definition 1.3 (Variance). The *variance* of a sample y_1, \dots, y_n is

$$\text{Var}(y_1, \dots, y_n) = n^{-1} \sum (y_i - \bar{y})^2.$$

The *variance* of a random variable Y is

$$\text{Var}(Y) = \mathbb{E}((Y - \mu_Y)^2)$$

Definition 1.4 (Standard deviation). The *standard deviation* of a sample y_1, \dots, y_n is

$$\text{SD}(y_1, \dots, y_n) = \sqrt{n^{-1} \sum (y_i - \bar{y})^2}.$$

The *standard deviation* of a random variable Y is

$$\text{SD}(Y) = \sqrt{\mathbb{E}((Y - \mu_Y)^2)}.$$

The variance (standard deviation) of Y is often denoted σ_Y^2 (σ_Y). The variances of common distributions will be derived later in the book.

Caution: for reasons which we don't go into here, many books define the variance of a sample as $\text{Var}(y_1, \dots, y_n) = (n-1)^{-1} \sum (y_i - \bar{y})^2$. For large n there is no practical difference between the two definitions. And the definition of variance of a random variable remains unchanged.

While the definition of the variance of a random variable highlights its interpretation as deviations away from the mean, there is an equivalent formula that is sometimes easier to compute.

Theorem 1.2. *If Y is a random variable, then $\text{Var}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}Y)^2$.*

Proof.

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}((Y - \mathbb{E}Y)^2) \\ &= \mathbb{E}(Y^2 - 2Y\mathbb{E}Y + (\mathbb{E}Y)^2) \\ &= \mathbb{E}(Y^2) - 2(\mathbb{E}Y)^2 + (\mathbb{E}Y)^2 \\ &= \mathbb{E}(Y^2) - (\mathbb{E}Y)^2 \end{aligned}$$

□

To develop a feel for what the standard deviation measures, Figure 1.13 repeats Figure 1.12 and adds arrows showing ± 1 standard deviation away from the mean. Standard deviations have the same units as the original random variable; variances have squared units. E.g., if Y is measured in degrees, then $SD(Y)$ is in degrees but $Var(Y)$ is in degrees². Because of this, SD is easier to interpret graphically. That's why we were able to depict SD's in Figure 1.13.

Most “mound-shaped” samples, that is, samples that are unimodal and roughly symmetric, follow this rule of thumb:

- about 2/3 of the sample falls within about 1 standard deviation of the mean;
- about 95% of the sample falls within about 2 standard deviations of the mean.

The rule of thumb has implications for predictive accuracy. If x_1, \dots, x_n are a sample from a mound-shaped distribution, then one would predict that future observations will be around \bar{x} with, again, about 2/3 of them within about one SD and about 95% of them within about two SD's.

To illustrate further, we'll calculate the SD of a few mound-shaped random variables and compare the SD's to the pdf's.

Binomial Let $Y \sim \text{Bin}(30, .5)$.

$$\begin{aligned}
 \text{Var}(Y) &= \mathbb{E}(Y^2) - (\mathbb{E}Y)^2 \\
 &= \sum_{y=0}^{30} y^2 \binom{30}{y} \cdot 5^{30} - 15^2 \\
 &= \sum_{y=1}^{30} y \frac{30!}{(y-1)!(30-y)!} \cdot 5^{30} - 15^2 \\
 &= 15 \sum_{v=0}^{29} (v+1) \frac{29!}{v!(29-v)!} \cdot 5^{29} - 15^2 \\
 &= 15 \left(\sum_{v=0}^{29} v \frac{29!}{v!(29-v)!} \cdot 5^{29} + \sum_{v=0}^{29} \frac{29!}{v!(29-v)!} \cdot 5^{29} \right) - 15^2 \\
 &= 15 \left(\frac{29}{2} + 1 - 15 \right) \\
 &= \frac{15}{2}
 \end{aligned} \tag{1.12}$$

and therefore $SD(Y) = \sqrt{15/2} \approx 2.7$. (See Exercises 19 and 20.)

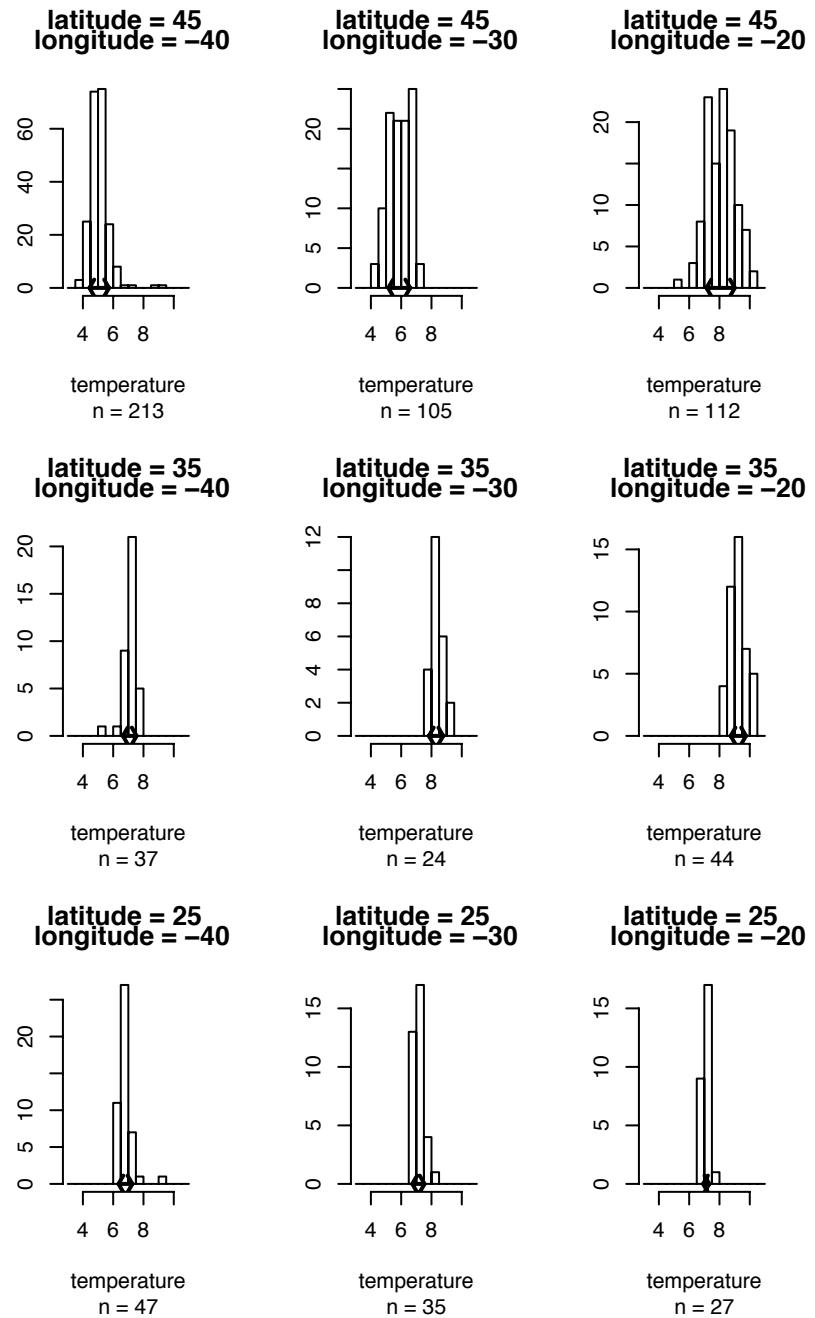


Figure 1.13: Water temperatures ($^{\circ}\text{C}$) at 1000m depth, latitude 25, 35, 45 degrees North, longitude 20, 30, 40 degrees West, with standard deviations

Normal Let $Y \sim N(0, 1)$.

$$\begin{aligned}
 \text{Var}(Y) &= \mathbb{E}(Y^2) \\
 &= \int_{-\infty}^{\infty} \frac{y^2}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
 &= 2 \int_0^{\infty} \frac{y^2}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
 &= 2 \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
 &= 1
 \end{aligned} \tag{1.13}$$

and therefore $\text{SD}(Y) = 1$. (See Exercises 19 and 20.)

Figure 1.14 shows the comparison. The top panel shows the pdf of the $\text{Bin}(30, .5)$ distribution; the bottom panel shows the $N(0, 1)$ distribution.

Figure 1.14 was produced by the following R code.

```

par ( mfrow=c(2,1) )

y <- 0:30
sd <- sqrt ( 15 / 2 )
plot ( y, dbinom(y,30,.5), ylab="p(y)" )
arrows ( 15-2*sd, 0, 15+2*sd, 0, angle=60, length=.1,
         code=3, lwd=2 )
text ( 15, .008, "+/- 2 SD's" )
arrows ( 15-sd, .03, 15+sd, .03, angle=60, length=.1,
         code=3, lwd=2 )
text ( 15, .04, "+/- 1 SD" )

y <- seq(-3,3,length=60)
plot ( y, dnorm(y,0,1), ylab="p(y)", type="l" )
arrows ( -2, .02, 2, .02, angle=60, length=.1, code=3, lwd=2 )
text ( 0, .04, "+/- 2 SD's" )
arrows ( -1, .15, 1, .15, angle=60, length=.1, code=3, lwd=2 )
text ( 0, .17, "+/- 1 SD" )

```

- `arrows(x0, y0, x1, y1, length, angle, code, ...)` adds arrows to a plot. See the documentation for the meaning of the arguments.

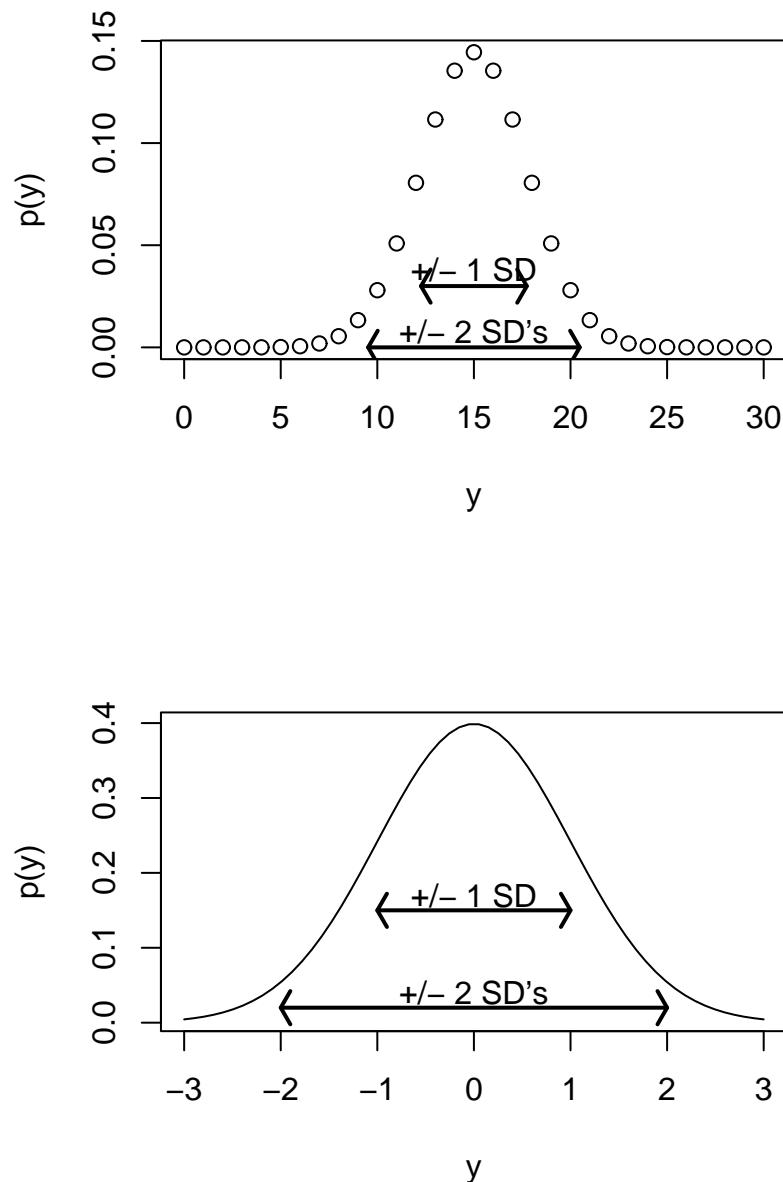


Figure 1.14: Two pdf's with ± 1 and ± 2 SD's. top panel: $\text{Bin}(30, .5)$; bottom panel: $N(0, 1)$.

- `text` adds text to a plot. See the documentation for the meaning of the arguments.

Definition 1.5 (Moment). The r 'th moment of a sample y_1, \dots, y_n or random variable Y is defined as

$$\begin{aligned} n^{-1} \sum (y_i - \bar{y})^r & \quad \text{(for samples)} \\ \mathbb{E}((Y - \mu_Y)^r) & \quad \text{(for random variables)} \end{aligned}$$

Variances are second moments. Moments above the second have little applicability.

R has built-in functions to compute means and variances and can compute other moments easily. Note that R uses the divisor $n - 1$ in its definition of variance.

```
# Use R to calculate moments of the Bin(100,.5) distribution
x <- rbinom ( 5000, 100, .5 )
m <- mean ( x ) # the mean
v <- var ( x ) # the variance
s <- sqrt ( v ) # the SD
mean ( (x-m)^3 ) # the third moment
our.v <- mean ( (x-m)^2 ) # our variance
our.s <- sqrt ( our.v ) # our standard deviation
print ( c ( v, our.v ) ) # not quite equal
print ( c ( s, our.s ) ) # not quite equal
```

- `rbinom(...)` generates random draws from the binomial distribution. The `5000` says how many draws to generate. The `100` and `.5` say that the draws are to be from the $\text{Bin}(100, .5)$ distribution.

Let h be a function. Then $\mathbb{E}[h(Y)] = \int h(y)p(y) dy$ ($\sum h(y)p(y)$ in the discrete case) is the expected value of $h(Y)$ and is called a *generalized moment*. There are sometimes two ways to evaluate $\mathbb{E}[h(Y)]$. One is to evaluate the integral. The other is to let $X = h(Y)$, find p_X , and then evaluate $\mathbb{E}[X] = \int x p_X(x) dx$. For example, let Y have pdf $f_Y(y) = 1$ for $y \in (0, 1)$, and let $X = h(Y) = \exp(Y)$.

Method 1

$$\mathbb{E}[h(Y)] = \int_0^1 \exp(y) dy = \exp(y)|_0^1 = e - 1.$$

Method 2

$$p_X(x) = p_Y(\log(x)) dy/dx = 1/x$$

$$\mathbb{E}[X] = \int_1^e x p_X(x) dx = \int_1^e 1 dx = e - 1$$

If h is a linear function then $\mathbb{E}[h(Y)]$ has a particularly appealing form.

Theorem 1.3. *If $X = a + bY$ then $\mathbb{E}[X] = a + b\mathbb{E}[Y]$.*

Proof. We prove the continuous case; the discrete case is left as an exercise.

$$\begin{aligned}\mathbb{E}X &= \int (a + by)f_Y(y) dy \\ &= a \int f_Y(y) dy + b \int yf_Y(y) dy \\ &= a + b\mathbb{E}Y\end{aligned}$$

□

There is a corresponding theorem for variance.

Theorem 1.4. *If $X = a + bY$ then $\text{Var}(X) = b^2 \text{Var}(Y)$.*

Proof. We prove the continuous case; the discrete case is left as an exercise. Let $\mu = \mathbb{E}[Y]$.

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(a + bY - (a + b\mu))^2] \\ &= \mathbb{E}[b^2(Y - \mu)^2] \\ &= b^2 \text{Var}(Y)\end{aligned}$$

□

1.5 Joint, Marginal and Conditional Probability

Statisticians often have to deal simultaneously with the probabilities of several events, quantities, or random variables. For example, we may classify voters in a city according to political party affiliation and support for a school bond referendum. Let A and S be a voter's affiliation and support, respectively.

$$A = \begin{cases} D & \text{if Democrat} \\ R & \text{if Republican} \end{cases} \quad \text{and} \quad S = \begin{cases} Y & \text{if in favor} \\ N & \text{if opposed.} \end{cases}$$

Suppose a polling organization finds that 80% of Democrats and 35% of Republicans favor the bond referendum. The 80% and 35% are called *conditional* probabilities because they

	For	Against	
Democrat	48%	12%	60%
Republican	14%	26%	40%
	62%	38%	

Table 1.1: Party Affiliation and Referendum Support

are conditional on party affiliation. The notation for conditional probabilities is $p_{S|A}$. As usual, the subscript indicates which random variables we're talking about. Specifically,

$$\begin{aligned} p_{S|A}(Y|D) &= 0.80; & p_{S|A}(N|D) &= 0.20; \\ p_{S|A}(Y|R) &= 0.35; & p_{S|A}(N|R) &= 0.65. \end{aligned}$$

We say “the conditional probability that $S = N$ given $A = D$ is 0.20”, etc.

Suppose further that 60% of voters in the city are Democrats. Then 80% of 60% = 48% of the voters are Democrats who favor the referendum. The 48% is called a *joint* probability because it is the probability of $(A = D, S = Y)$ jointly. The notation is $p_{A,S}(D, Y) = .48$. Likewise, $p_{A,S}(D, N) = .12$; $p_{A,S}(R, Y) = .14$; and $p_{A,S}(R, N) = 0.26$. Table 1.1 summarizes the calculations. The quantities .60, .40, .62, and .38 are called *marginal* probabilities. The name derives from historical reasons, because they were written in the margins of the table. Marginal probabilities are probabilities for one variable alone, the ordinary probabilities that we've been talking about all along.

The event $A = D$ can be partitioned into the two smaller events $(A = D, S = Y)$ and $(A = D, S = N)$. So

$$p_A(D) = .60 = .48 + .12 = p_{A,S}(D, Y) + p_{A,S}(D, N).$$

The event $A = R$ can be partitioned similarly. Too, the event $S = Y$ can be partitioned into $(A = D, S = Y)$ and $(A = R, S = Y)$. So

$$p_S(Y) = .62 = .48 + .14 = p_{A,S}(D, Y) + p_{A,S}(R, Y).$$

These calculations illustrate a general principle: *To get a marginal probability for one variable, add the joint probabilities for all values of the other variable.* The general formulae for working simultaneously with two discrete random variables X and Y are

$$\begin{aligned} f_{X,Y}(x, y) &= f_X(x) \cdot f_{Y|X}(y|x) = f_Y(y) \cdot f_{X|Y}(x|y) & (1.14) \\ f_X(x) &= \sum_y f_{X,Y}(x, y) & f_Y(y) &= \sum_x f_{X,Y}(x, y) \end{aligned}$$

Sometimes we know joint probabilities and need to find marginals and conditionals; sometimes it's the other way around. And sometimes we know f_X and $f_{Y|X}$ and need to find f_Y or $f_{X|Y}$. The following story is an example of the latter. It is a common problem in drug testing, disease screening, polygraph testing, and many other fields.

The participants in an athletic competition are to be randomly tested for steroid use. The test is 90% accurate in the following sense: for athletes who use steroids, the test has a 90% chance of returning a positive result; for non-users, the test has a 10% chance of returning a positive result. Suppose that only 30% of athletes use steroids. An athlete is randomly selected. Her test returns a positive result. What is the probability that she is a steroid user?

This is a problem of two random variables, U , the steroid use of the athlete and T , the test result of the athlete. Let $U = 1$ if the athlete uses steroids; $U = 0$ if not. Let $T = 1$ if the test result is positive; $T = 0$ if not. We want $f_{U|T}(1|1)$. We can calculate $f_{U|T}$ if we know $f_{U,T}$; and we can calculate $f_{U,T}$ because we know f_U and $f_{T|U}$. Pictorially,

$$f_U, f_{T|U} \longrightarrow f_{U,T} \longrightarrow f_{U|T}$$

The calculations are

$$\begin{aligned} f_{U,T}(0,0) &= (.7)(.9) = .63 & f_{U,T}(0,1) &= (.7)(.1) = .07 \\ f_{U,T}(1,0) &= (.3)(.1) = .03 & f_{U,T}(1,1) &= (.3)(.9) = .27 \end{aligned}$$

so

$$f_T(0) = .63 + .03 = .66 \quad f_T(1) = .07 + .27 = .34$$

and finally

$$f_{U|T}(1|1) = f_{U,T}(1,1)/f_T(1) = .27/.34 \approx .80.$$

In other words, even though the test is 90% accurate, the athlete has only an 80% chance of using steroids. If that doesn't seem intuitively reasonable, think of a large number of athletes, say 100. About 30 will be steroid users of whom about 27 will test positive. About 70 will be non-users of whom about 7 will test positive. So there will be about 34 athletes who test positive, of whom about 27, or 80% will be users.

Table 1.2 is another representation of the same problem. It is important to become familiar with the concepts and notation in terms of marginal, conditional and joint distributions, and not to rely too heavily on the tabular representation because in more complicated problems there is no convenient tabular representation.

Example 1.6 is a further illustration of joint, conditional, and marginal distributions.

	$T = 0$	$T = 1$	
$U = 0$.63	.07	.70
$U = 1$.03	.27	.30
	.66	.34	

Table 1.2: Steroid Use and Test Results

Example 1.6 (Seedlings)

Example 1.4 introduced an observational experiment to learn about the rate of seedling production and survival at the Coweeta Long Term Ecological Research station in western North Carolina. For a particular quadrat in a particular year, let N be the number of new seedlings that emerge. Suppose that $N \sim \text{Poi}(\lambda)$ for some $\lambda > 0$. Each seedling either dies over the winter or survives to become an old seedling the next year. Let θ be the probability of survival and X be the number of seedlings that survive. Suppose that the survival of any one seedling is not affected by the survival of any other seedling. Then $X \sim \text{Bin}(N, \theta)$. Figure 1.15 shows the possible values of the pair (N, X) . The probabilities associated with each of the points in Figure 1.15 are denoted $f_{N,X}$ where, as usual, the subscript indicates which variables we're talking about. For example, $f_{N,X}(3, 2)$ is the probability that $N = 3$ and $X = 2$.

The next step is to figure out what the joint probabilities are. Consider, for example, the event $N = 3$. That event can be partitioned into the four smaller events $(N = 3, X = 0)$, $(N = 3, X = 1)$, $(N = 3, X = 2)$, and $(N = 3, X = 3)$. So

$$f_N(3) = f_{N,X}(3, 0) + f_{N,X}(3, 1) + f_{N,X}(3, 2) + f_{N,X}(3, 3)$$

The Poisson model for N says $f_N(3) = \text{P}[N = 3] = e^{-\lambda}\lambda^3/6$. But how is the total $e^{-\lambda}\lambda^3/6$ divided into the four parts? That's where the Binomial model for X comes in. The division is made according to the Binomial probabilities

$$\binom{3}{0}(1 - \theta)^3 \quad \binom{3}{1}\theta(1 - \theta)^2 \quad \binom{3}{2}\theta^2(1 - \theta) \quad \binom{3}{3}\theta^3$$

The $e^{-\lambda}\lambda^3/6$ is a marginal probability like the 60% in the affiliation/support problem. The binomial probabilities above are *conditional* probabilities like the 80% and 20%; they are conditional on $N = 3$. The notation is $f_{X|N}(2|3)$ or $\text{P}[X = 2|N = 3]$. The joint

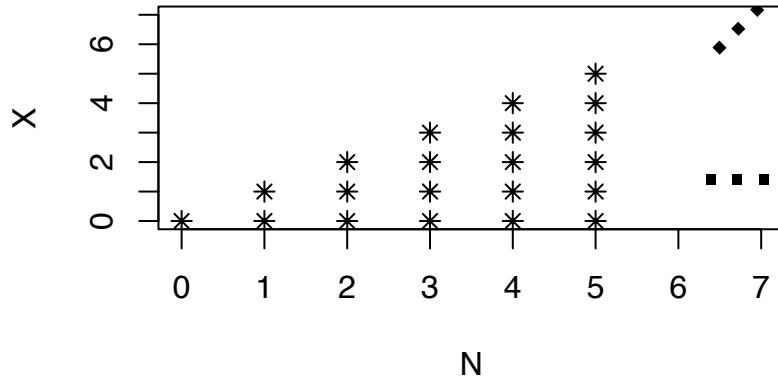


Figure 1.15: Permissible values of N and X , the number of new seedlings and the number that survive.

probabilities are

$$\begin{aligned} f_{N,X}(3,0) &= \frac{e^{-\lambda}\lambda^3}{6}(1-\theta)^3 & f_{N,X}(3,1) &= \frac{e^{-\lambda}\lambda^3}{6}3\theta(1-\theta)^2 \\ f_{N,X}(3,2) &= \frac{e^{-\lambda}\lambda^3}{6}3\theta^2(1-\theta) & f_{N,X}(3,3) &= \frac{e^{-\lambda}\lambda^3}{6}\theta^3 \end{aligned}$$

In general,

$$f_{N,X}(n,x) = f_N(n)f_{X|N}(x|n) = \frac{e^{-\lambda}\lambda^n}{n!} \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

An ecologist might be interested in f_X , the pdf for the number of seedlings that will be recruited into the population in a particular year. For a particular number x , $f_X(x)$ is like looking in Figure 1.15 along the horizontal line corresponding to $X = x$. To get

$f_X(x) \equiv P[X = x]$, we must add up all the probabilities on that line.

$$\begin{aligned} f_X(x) &= \sum_n f_{N,X}(n, x) = \sum_{n=x}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!} \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &= \sum_{n=x}^{\infty} \frac{e^{-\lambda(1-\theta)} (\lambda(1-\theta))^{n-x}}{(n-x)!} \frac{e^{-\lambda\theta} (\lambda\theta)^x}{x!} \\ &= \frac{e^{-\lambda\theta} (\lambda\theta)^x}{x!} \sum_{z=0}^{\infty} \frac{e^{-\lambda(1-\theta)} (\lambda(1-\theta))^z}{z!} \\ &= \frac{e^{-\lambda\theta} (\lambda\theta)^x}{x!} \end{aligned}$$

The last equality follows since $\sum_z \dots = 1$ because it is the sum of probabilities from the $\text{Poi}(\lambda(1-\theta))$ distribution. The final result is recognized as a probability from the $\text{Poi}(\lambda^*)$ distribution where $\lambda^* = \lambda\theta$. So $X \sim \text{Poi}(\lambda^*)$.

In the derivation we used the substitution $z = n - x$. The trick is worth remembering.

For continuous random variables, conditional and joint densities are written $p_{X|Y}(x|y)$ and $p_{X,Y}(x,y)$ respectively and, analogously to Equation 1.14 we have

$$\begin{aligned} p_{X,Y}(x,y) &= p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y) \quad (1.15) \\ p_X(x) &= \int_{-\infty}^{\infty} p_{X,Y}(x,y) dy \quad p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x,y) dx \end{aligned}$$

The logic is the same as for discrete random variables. In order for $(X = x, Y = y)$ to occur we need either of the following.

1. First $X = x$ occurs then $Y = y$ occurs. The “probability” of that happening is just $p_X(x) \times p_{Y|X}(y|x)$, the “probability” that $X = x$ occurs times the “probability” that $Y = y$ occurs under the condition that $X = x$ has already occurred. “Probability” is in quotes because, for continuous random variables, the probability is 0. But “probability” is a useful way to think intuitively.
2. First $Y = y$ occurs then $X = x$ occurs. The reasoning is similar to that in item 1.

Just as for single random variables, probabilities are integrals of the density. If A is a region in the (x,y) plane, $P[(X, Y) \in A] = \int_A p(x,y) dx dy$, where $\int_A \dots$ indicates a double integral over the region A .

Just as for discrete random variables, the unconditional density of a random variable is called its *marginal* density; p_X and p_Y are marginal densities. Let $B \subset \mathbb{R}$ be a set. Since

a density is “the function that must be integrated to calculate a probability”, on one hand, $P[X \in B] = \int_B p_X(x) dx$. On the other hand,

$$P[X \in B] = P[(X, Y) \in B \times \mathbb{R}] = \int_B \left(\int_{\mathbb{R}} p_{X,Y}(x, y) dy \right) dx$$

which implies $p_X(x) = \int_{\mathbb{R}} p_{X,Y}(x, y) dy$.

An example will help illustrate. A customer calls the computer Help Line. Let X be the amount of time he spends on hold and Y be the total duration of the call. The amount of time a consultant spends with him after his call is answered is $W = Y - X$. Suppose the joint density is $p_{X,Y}(x, y) = e^{-y}$ in the region $0 < x < y < \infty$.

1. *What is the marginal density of X ?*

$$p(x) = \int p(x, y) dy = \int_x^\infty e^{-y} dy = -e^{-y} \Big|_x^\infty = e^{-x}$$

2. *What is the marginal density of Y ?*

$$p(y) = \int p(x, y) dx = \int_0^y e^{-y} dx = ye^{-y}$$

3. *What is the conditional density of X given Y ?*

$$p(x|y) = \frac{p(x, y)}{p(y)} = y^{-1}$$

4. *What is the conditional density of Y given X ?*

$$p(y|x) = \frac{p(x, y)}{p(x)} = e^{x-y}$$

5. *What is the marginal density of W ?*

$$\begin{aligned} p(w) &= \frac{d}{dw} P[W \leq w] = \frac{d}{dw} \int_0^\infty \int_x^{x+w} e^{-y} dy dx \\ &= \frac{d}{dw} \int_0^\infty \left(-e^{-y} \Big|_x^{x+w} \right) dx \\ &= \frac{d}{dw} \int_0^\infty e^{-x} (1 - e^{-w}) dx = e^{-w} \end{aligned}$$

Figure 1.16 illustrates the Help Line calculations. For questions 1 and 2, the answer comes from using Equations 1.15. The only part deserving comment is the limits of integration. In question 1, for example, for any particular value $X = x$, Y ranges from x to ∞ , as can be seen from panel (a) of the figure. That's where the limits of integration come from. In question 2, for any particular y , $X \in (0, y)$, which are the limits of integration. Panel (d) shows the conditional density of X given Y for three different values of Y . We see that the density of X is uniform on the interval $(0, y)$. See Section 5.4 for discussion of this density. Panel (d) shows the conditional density of Y given X for three different values of X . It shows, first, that $Y > X$ and second, that the density of Y decays exponentially. See Section 1.3.3 for discussion of this density. Panel (f) shows the region of integration for question 5. Take the time to understand the method being used to answer question 5.

When dealing with a random variable X , sometimes its pdf is given to us and we can calculate its expectation:

$$\mathbb{E}(X) = \int xp(x) dx.$$

(The integral is replaced by a sum if X is discrete.) Other times X arises more naturally as part of a pair (X, Y) and its expectation is

$$\mathbb{E}(X) = \iint xp(x, y) dxdy.$$

The two formulae are, of course, equivalent. But when X does arise as part of a pair, there is still another way to view $p(x)$ and $\mathbb{E}(X)$:

$$p_X(x) = \int (p_{X|Y}(x|y)) p_Y(y) dy = \mathbb{E}(p_{X|Y}(x|y)) \quad (1.16)$$

$$\mathbb{E}(X) = \int \left(\int xp_{X|Y}(x|y) dx \right) p_Y(y) dy = \mathbb{E}(\mathbb{E}(X|Y)). \quad (1.17)$$

The notation deserves some explanation. For any number x , $p_{X|Y}(x|y)$ is a function of y , say $g(y)$. The middle term in Equation 1.16 is $\int g(y)p(y)dy$, which equals $\mathbb{E}(g(y))$, which is the right hand term. Similarly, $\mathbb{E}(X|Y)$ is a function of Y , say $h(Y)$. The middle term in Equation 1.17 is $\int h(y)p(y)dy$, which equals $\mathbb{E}(h(y))$, which is the right hand term.

Example 1.7 (Seedlings, continued)

Examples 1.4 and 1.6 discussed (N, X) , the number of new seedlings in a forest quadrat and the number of those that survived over the winter. Example 1.6 suggested the statistical model

$$N \sim \text{Poi}(\lambda) \quad \text{and} \quad X|N \sim \text{Bin}(N, \theta).$$

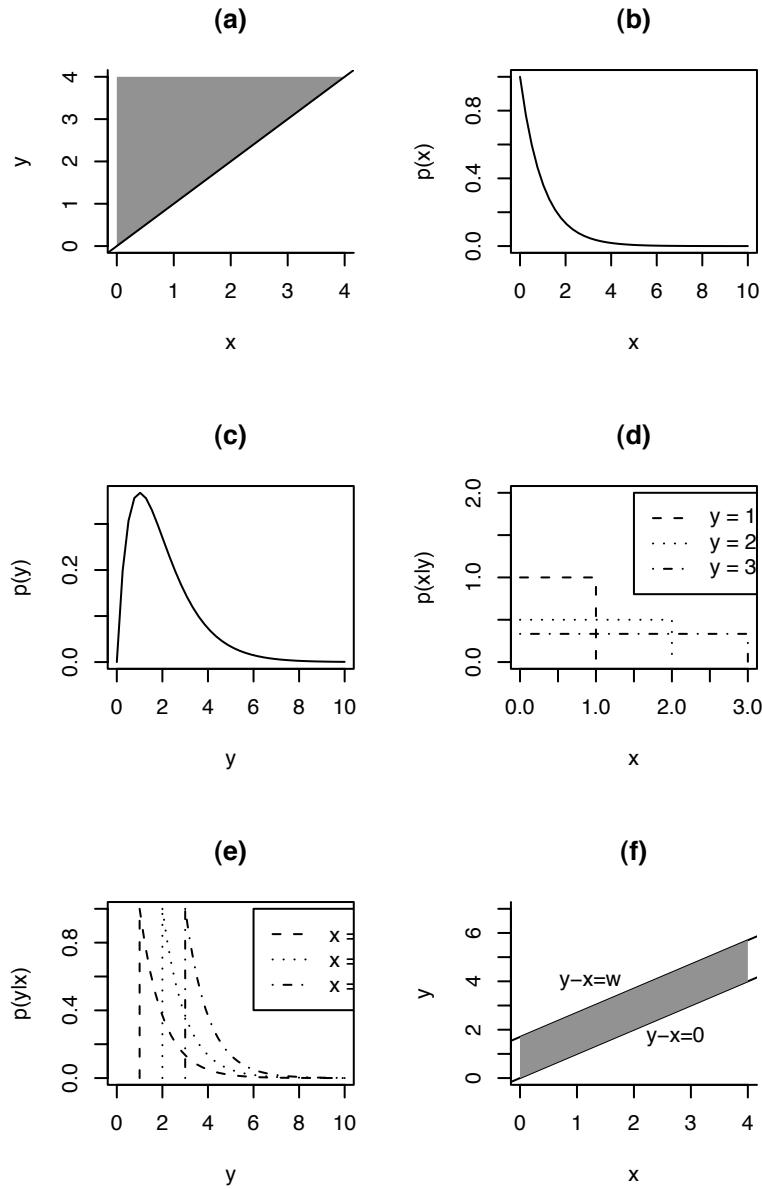


Figure 1.16: (a): the region of \mathbb{R}^2 where (X, Y) live; (b): the marginal density of X ; (c): the marginal density of Y ; (d): the conditional density of X given Y for three values of Y ; (e): the conditional density of Y given X for three values of X ; (f): the region $W \leq w$

Equation 1.17 shows that $\mathbb{E}(X)$ can be computed as

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|N)) = \mathbb{E}(N\theta) = \theta\mathbb{E}(N) = \theta\lambda.$$

Example 1.8 (Craps, continued)

Examples 1.1, 1.2 and 1.3 introduced the game of craps. Example 1.8 calculates the chance of winning.

$$\text{Let } X = \begin{cases} 0 & \text{if shooter loses} \\ 1 & \text{if shooter wins} \end{cases}$$

X has a Bernoulli distribution. We are trying to find

$$P[\text{shooter wins}] = p_X(1) = \mathbb{E}(X).$$

(Make sure you see why $p_X(1) = \mathbb{E}(X)$.) Let Y be the outcome of the Come-out roll. Equation 1.17 says

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(\mathbb{E}(X|Y)) \\ &= \mathbb{E}(X|Y=2)P[Y=2] + \mathbb{E}(X|Y=3)P[Y=3] \\ &\quad + \mathbb{E}(X|Y=4)P[Y=4] + \mathbb{E}(X|Y=5)P[Y=5] \\ &\quad + \mathbb{E}(X|Y=6)P[Y=6] + \mathbb{E}(X|Y=7)P[Y=7] \\ &\quad + \mathbb{E}(X|Y=8)P[Y=8] + \mathbb{E}(X|Y=9)P[Y=9] \\ &\quad + \mathbb{E}(X|Y=10)P[Y=10] + \mathbb{E}(X|Y=11)P[Y=11] \\ &\quad + \mathbb{E}(X|Y=12)P[Y=12] \\ &= 0 \times \frac{1}{36} + 0 \times \frac{2}{36} + \mathbb{E}(X|Y=4)\frac{3}{36} \\ &\quad + \mathbb{E}(X|Y=5)\frac{4}{36} + \mathbb{E}(X|Y=6)\frac{5}{36} \\ &\quad + 1 \times \frac{6}{36} + \mathbb{E}(X|Y=8)\frac{5}{36} + \mathbb{E}(X|Y=9)\frac{4}{36} \\ &\quad + \mathbb{E}(X|Y=10)\frac{3}{36} + 1 \times \frac{2}{36} + 0 \times \frac{1}{36}. \end{aligned}$$

So it only remains to find $\mathbb{E}(X|Y=y)$ for $y = 4, 5, 6, 8, 9, 10$. The calculations are all similar. We will do one of them to illustrate. Let $w = \mathbb{E}(X|Y=5)$ and let z denote the next roll of the dice. Once 5 has been established as the point, then a roll of the dice has three possible outcomes: *win* (if $z=5$), *lose* (if $z=7$), or *roll again* (if z is anything

else). Therefore

$$\begin{aligned} w &= 1 \times 4/36 + 0 \times 6/36 + w \times 26/36 \\ (10/36)w &= 4/36 \\ w &= 4/10. \end{aligned}$$

After similar calculations for the other possible points we find

$$\mathbb{E}(X) = \frac{3}{9} \frac{3}{36} + \frac{4}{10} \frac{4}{36} + \frac{5}{11} \frac{5}{36} + \frac{6}{36} + \frac{5}{11} \frac{5}{36} + \frac{4}{10} \frac{4}{36} + \frac{3}{9} \frac{3}{36} + \frac{2}{36} \approx .493.$$

Craps is a very fair game; the house has only a slight edge.

1.6 Association, Dependence, Independence

It is often useful to describe or measure the degree of association between two random variables X and Y . The R dataset `iris` provides a good example. It contains the lengths and widths of sepals and petals of 150 iris plants. The first several lines of `iris` are

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa

Figure 1.17 shows each variable plotted against every other variable. It is evident from the plot that petal length and petal width are very closely associated with each other, while the relationship between sepal length and sepal width is much weaker. Statisticians need a way to quantify the strength of such relationships.

Figure 1.17 was produced by the following line of R code.

```
pairs ( iris[,1:4] )
```

- `pairs()` produces a *pairs plot*, a matrix of scatterplots of each pair of variables. The names of the variables are shown along the main diagonal of the matrix. The (i, j) th plot in the matrix is a plot of variable i versus variable j . For example, the upper right plot has sepal length on the vertical axis and petal width on the horizontal axis.

By far the most common measures of association are *covariance* and *correlation*.

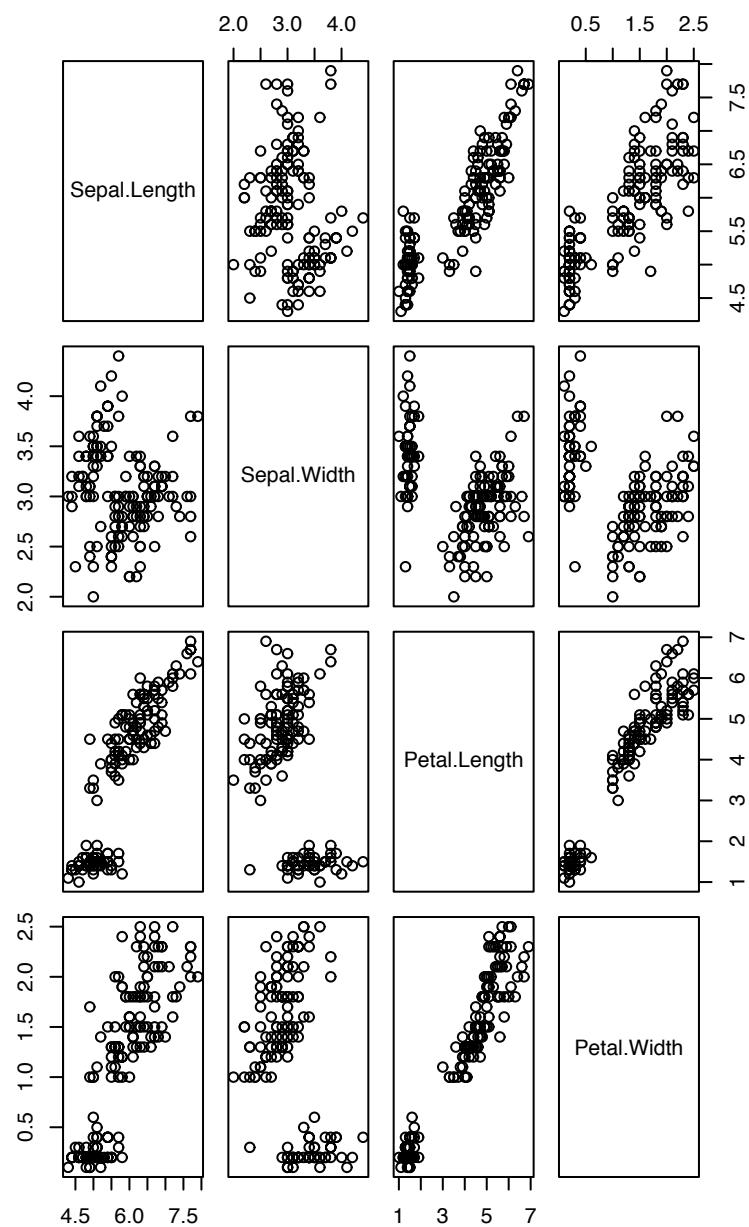


Figure 1.17: Lengths and widths of sepals and petals of 150 iris plants

Definition 1.6. The *covariance* of X and Y is

$$\text{Cov}(X, Y) \equiv \mathbb{E}((X - \mu_X)(Y - \mu_Y))$$

In R, `cov` measures the covariance in a sample. Thus, `cov(iris[, 1:4])` produces the matrix

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.68569351	-0.04243400	1.2743154	0.5162707
Sepal.Width	-0.04243400	0.18997942	-0.3296564	-0.1216394
Petal.Length	1.27431544	-0.32965638	3.1162779	1.2956094
Petal.Width	0.51627069	-0.12163937	1.2956094	0.5810063

in which the diagonal entries are variances and the off-diagonal entries are covariances.

The measurements in `iris` are in centimeters. To change to millimeters we would multiply each measurement by 10. Here's how that affects the covariances.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	68.569351	-4.243400	127.43154	51.62707
Sepal.Width	-4.243400	18.997942	-32.96564	-12.16394
Petal.Length	127.431544	-32.965638	311.62779	129.56094
Petal.Width	51.627069	-12.163937	129.56094	58.10063

Each covariance has been multiplied by 100 because each variable has been multiplied by 10. In fact, this rescaling is a special case of the following theorem.

Theorem 1.5. Let X and Y be random variables. Then $\text{Cov}(aX+b, cY+d) = ac \text{Cov}(X, Y)$.

Proof.

$$\begin{aligned} \text{Cov}(aX + b, cY + d) &= \mathbb{E}((aX + b - (a\mu_X + b))(cY + d - (c\mu_Y + d))) \\ &= \mathbb{E}(ac(X - \mu_X)(Y - \mu_Y)) = ac \text{Cov}(X, Y) \end{aligned}$$

□

Theorem 1.5 shows that $\text{Cov}(X, Y)$ depends on the scales in which X and Y are measured. A scale-free measure of association would also be useful. Correlation is the most common such measure.

Definition 1.7. The *correlation* between X and Y is

$$\text{Cor}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\text{SD}(X) \text{SD}(Y)}$$

`cor` measures correlation. The correlations in `iris` are

```
> cor(iris[,1:4])
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000 -0.1175698   0.8717538   0.8179411
Sepal.Width     -0.1175698  1.0000000  -0.4284401  -0.3661259
Petal.Length    0.8717538 -0.4284401   1.0000000   0.9628654
Petal.Width     0.8179411 -0.3661259   0.9628654   1.0000000
```

which confirms the visual impression that sepal length, petal length, and petal width are highly associated with each other, but are only loosely associated with sepal width.

Theorem 1.6 tells us that correlation is unaffected by linear changes in measurement scale.

Theorem 1.6. *Let X and Y be random variables. Then $\text{Cor}(aX + b, cY + d) = \text{Cor}(X, Y)$.*

Proof. See Exercise 41. □

Correlation doesn't measure all types of association; it only measures clustering around a straight line. The first two columns of Figure 1.18 show data sets that cluster around a line, but with some scatter above and below the line. These data sets are all well described by their correlations, which measure the extent of the clustering; the higher the correlation, the tighter the points cluster around the line and the less they scatter. Negative values of the correlation correspond to lines with negative slopes. The last column of the figure shows some other situations. The first panel of the last column is best described as having two isolated clusters of points. Despite the correlation of .96, the panel does not look at all like the last panel of the second column. The second and third panels of the last column show data sets that follow some nonlinear pattern of association. Again, their correlations are misleading. Finally, the last panel of the last column shows a data set in which most of the points are tightly clustered around a line but in which there are two outliers. The last column demonstrates that correlations are not good descriptors of nonlinear data sets or data sets with outliers.

Correlation measures linear association between random variables. But sometimes we want to say whether two random variables have any association at all, not just linear.

Definition 1.8. Two two random variables, X and Y , are said to be *independent* if $p(X | Y) = p(X)$, for all values of Y . If X and Y are not independent then they are said to be *dependent*.

If X and Y are independent then it is also true that $p(Y | X) = p(Y)$. The interpretation is that knowing one of the random variables does not change the probability distribution of the other. If X and Y are independent (dependent) we write $X \perp Y$ ($X \not\perp Y$). If X and

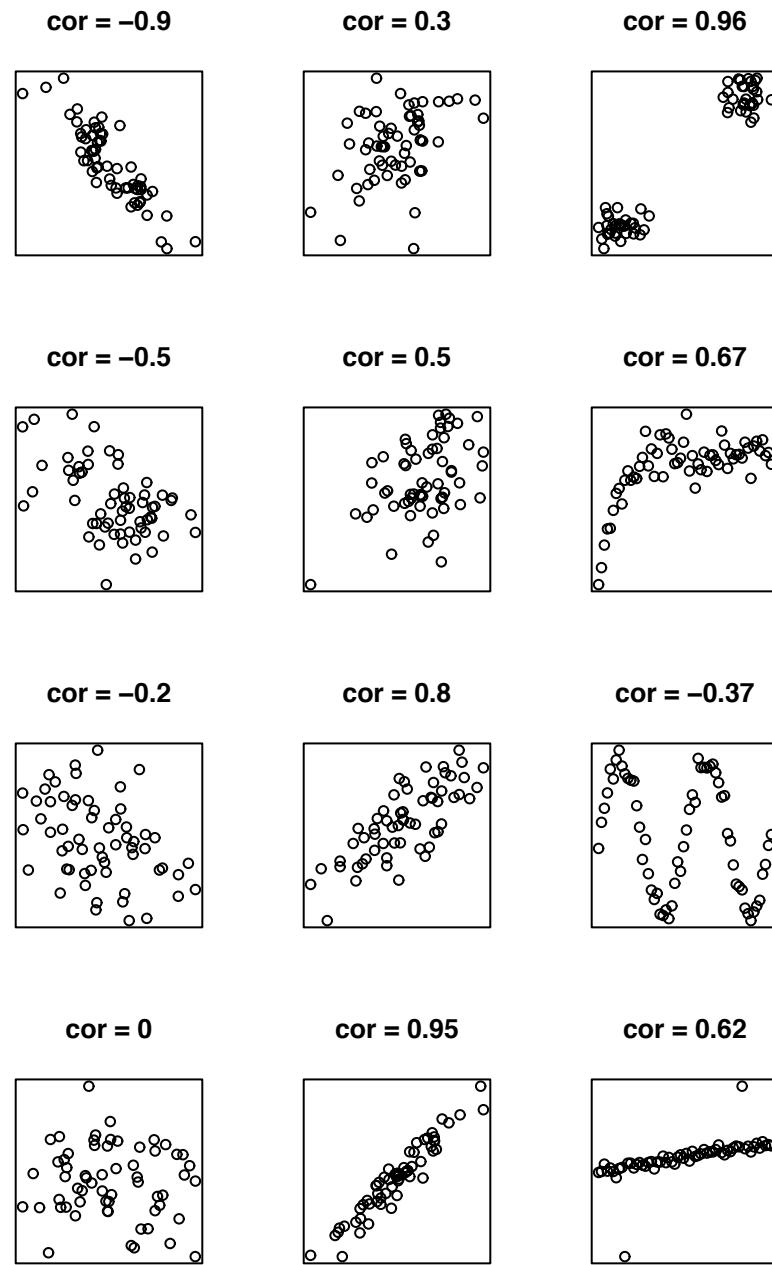


Figure 1.18: correlations

X and Y are independent then $\text{Cov}(X, Y) = \text{Cor}(X, Y) = 0$. The converse is not true. Also, if $X \perp Y$ then $p(x, y) = p(x)p(y)$. This last equality is usually taken to be the definition of independence.

Do not confuse independent with mutually exclusive. Let X denote the outcome of a die roll and let $A = 1$ if $X \in \{1, 2, 3\}$ and $A = 0$ if $X \in \{4, 5, 6\}$. A is called an *indicator* variable because it indicates the occurrence of a particular event. There is a special notation for indicator variables:

$$A = \mathbf{1}_{\{1,2,3\}}(X).$$

$\mathbf{1}_{\{1,2,3\}}$ is an *indicator function*. $\mathbf{1}_{\{1,2,3\}}(X)$ is either 1 or 0 according to whether X is in the subscript. Let $B = \mathbf{1}_{\{4,5,6\}}(X)$, $C = \mathbf{1}_{\{1,3,5\}}(X)$, $D = \mathbf{1}_{\{2,4,6\}}(X)$ and $E = \mathbf{1}_{\{1,2,3,4\}}(X)$. A and B are dependent because $P[A] = .5$ but $P[A|B] = 0$. D and E are independent because $P[D] = P[D|E] = .5$. You can also check that $P[E] = P[E|D] = 2/3$. Do not confuse dependence with causality. A and B are dependent, but neither causes the other.

For an example, recall the Help Line story on page 46. X and Y were the amount of time on hold and the total length of the call, respectively. The difference was $W = Y - X$. We found $p(x|y) = y^{-1}$. Because $p(x|y)$ depends on y , $X \not\perp Y$. Similarly, $p(y|x)$ depends on x . Does that make sense? Would knowing something about X tell us anything about Y ?

What about X and W ? Are they independent?

$$p(w|x) = \frac{d}{dw} P[W \leq w | X = x] = \frac{d}{dw} P[Y \leq w + x | X = x] = e^{-w}$$

which does not depend on x . Therefore $X \perp W$. Does that make sense? Would knowing something about X tell us anything about W ?

Example 1.9 (Seedlings, continued)

Examples 1.4, 1.6, and 1.7 were about new seedlings in forest quadrats. Suppose that ecologists observe the number of new seedlings in a quadrat for k successive years; call the observations N_1, \dots, N_k . If the seedling arrival rate is the same every year, then we could adopt the model $N_i \sim \text{Poi}(\lambda)$. I.e., λ is the same for every year. If λ is known, or if we condition on λ , then the number of new seedlings in one year tells us nothing about the number of new seedlings in another year, we would model the N_i 's as being independent, and their joint density, conditional on λ , would be

$$p(n_1, \dots, n_k | \lambda) = \prod_{i=1}^k p(n_i | \lambda) = \prod_{i=1}^k \frac{e^{-\lambda} \lambda^{n_i}}{n_i!} = \frac{e^{-k\lambda} \lambda^{\sum n_i}}{\prod n_i!}$$

But if λ is unknown then we might treat it like a random variable. It would be a random variable if, for instance, different years had different λ 's, we chose a year at random or

if we thought of Nature as randomly choosing λ for our particular year. In that case the data from early years, N_1, \dots, N_m , say, yield information about λ and therefore about likely values of N_{m+1}, \dots, N_k , so the N_i 's are dependent. In fact,

$$p(n_1, \dots, n_k) = \int p(n_1, \dots, n_k | \lambda) p(\lambda) d\lambda = \int \frac{e^{-k\lambda} \lambda^{\sum n_i}}{\prod n_i!} p(\lambda) d\lambda$$

So, whether the N_i 's are independent is not a question with a single right answer. Instead, it depends on our perspective. But in either case, we would say the N_i 's are conditionally independent given λ .

1.7 Simulation

We have already seen, in Example 1.2, an example of computer simulation to estimate a probability. More broadly, simulation can be helpful in several types of problems: calculating probabilities, assessing statistical procedures, and evaluating integrals. These are explained and exemplified in the next several subsections.

1.7.1 Calculating Probabilities

Probabilities can often be estimated by computer simulation. Simulations are especially useful for events so complicated that their probabilities cannot be easily calculated by hand, but composed of smaller events that are easily mimicked on computer. For instance, in Example 1.2 we wanted to know the probability that the shooter in a craps game rolls either 7 or 11 on the Come Out roll. Although it's easy enough to calculate this probability exactly, we did it by simulation in the Example.

Expected values can also be estimated by simulations. Let Y be a random variable and suppose we want to estimate the expected value of some function $\mu_g \equiv \mathbb{E}(g(Y))$. We can write a computer program to simulate Y many times. To keep track of the simulations we use the notation $y^{(j)}$ for the j 'th simulated value of Y . Let n be the number of simulations. Then

$$\hat{\mu}_g = n^{-1} \sum g(y^{(i)})$$

is a sensible estimate of μ_g . In fact, the Law of Large Numbers tells us that

$$\lim_{n \rightarrow \infty} n^{-1} \sum g(y^{(i)}) = \mu_g.$$

So as we do a larger and larger simulation we get a more and more accurate estimate of μ_g .

Probabilities can be computed as special cases of expectations. Suppose we want to calculate $P[Y \in S]$ for some set S . Define $X \equiv \mathbf{1}_S(Y)$. Then $P[Y \in S] = \mathbb{E}(X)$ and is sensibly estimated by

$$\frac{\text{number of occurrences of } S}{\text{number of trials}} = n^{-1} \sum x^{(i)}.$$

Example 1.10 illustrates with the game of Craps.

Example 1.10 (Craps, continued)

Example 1.8 calculated the chance of winning the game of Craps. Here is the R code to calculate the same probability by simulation.

```
makepoint <- function ( point ) {
  determined <- F
  while ( !determined ) { # roll until outcome is determined
    roll <- sum ( sample ( 6, 2, replace=T ) )
    if ( roll == point ) {
      made <- T
      determined <- T
    } else if ( roll == 7 ) {
      made <- F
      determined <- T
    }
  } # end while
  return ( made )
} # end makepoint

sim.craps <- function () {
  roll <- sum ( sample ( 6, 2, replace=T ) )
  if ( roll==7 || roll==11 )
    win <- T
  else if ( roll==2 || roll==3 || roll==12 )
    win <- F
  else
    win <- makepoint ( roll )

  return ( win )
}
```

```

n.sim <- 1000
wins <- 0
for ( i in 1:n.sim )
  wins <- wins + sim.craps()
print ( wins/n.sim )

```

- “!” is R’s symbol for *not*. If `determined` is T then `!determined` is F.
- `while(!determined)` begins a loop. The loop will repeat as many times as necessary as long as `!determined` is T.

Try the example code a few times. See whether you get about 49% as Example 1.8 suggests.

Along with the estimate itself, it is useful to estimate the accuracy of $\hat{\mu}_g$ as an estimate of μ_g . If the simulations are independent then $\text{Var}(\hat{\mu}_g) = n^{-1} \text{Var}(g(Y))$; if there are many of them then $\text{Var}(g(Y))$ can be well estimated by $n^{-1} \sum((g(y) - \hat{\mu}_g)^2)$ and $\text{SD}(g(Y))$ can be well estimated by

$n^{-1/2} \sqrt{\sum((g(y) - \hat{\mu}_g)^2)}$. Because SD’s decrease in proportion to $n^{-1/2}$ (See the Central Limit Theorem.), it takes a 100 fold increase in n to get, for example, a 10 fold increase in accuracy.

Similar reasoning applies to probabilities, but when we are simulating the occurrence or nonoccurrence of an event, then the simulations are Bernoulli trials, so we have a more explicit formula for the variance and SD.

Example 1.11 (Craps, continued)

How accurate is the simulation in Example 1.10?

The simulation keeps track of X , the number of successes in `n.sim` trials. Let θ be the true probability of success. (Example 1.8 found $\theta \approx .49$, but in most practical applications we won’t know θ .)

$$\begin{aligned} X &\sim \text{Bin}(n.\text{sim}, \theta) \\ \text{Var}(X) &= n.\text{sim}(\theta)(1 - \theta) \\ \text{SD}(X/n.\text{sim}) &= \sqrt{(\theta)(1 - \theta)/n.\text{sim}} \end{aligned}$$

and, by the Central Limit Theorem if `n.sim` is large,

$$\hat{\theta} = X/n.\text{sim} \sim N(\theta, (\theta(1 - \theta)/n.\text{sim})^{1/2})$$

What does this mean in practical terms? How accurate is the simulation when $n.sim = 50$, or 200 , or 1000 , say? To illustrate we did 1000 simulations with $n.sim = 50$, then another 1000 with $n.sim = 200$, and then another 1000 with $n.sim = 1000$.

The results are shown as a boxplot in Figure 1.19. In Figure 1.19 there are three boxes, each with *whiskers* extending vertically. The box for $n.sim = 50$ shows that the median of the $1000 \hat{\theta}$'s was just about $.50$ (the horizontal line through the box), that 50% of the $\hat{\theta}$'s fell between about $.45$ and $.55$ (the upper and lower ends of the box), and that almost all of the $\hat{\theta}$'s fell between about $.30$ and $.68$ (the extent of the whiskers). In comparison, the $1000 \hat{\theta}$'s for $n.sim = 200$ are spread out about half as much, and the $1000 \hat{\theta}$'s for $n.sim = 1000$ are spread out about half as much again. The factor of about a half comes from the $n.sim^5$ in the formula for $SD(\hat{\theta})$. When $n.sim$ increases by a factor of about 4, the SD decreases by a factor of about 2. See the notes for Figure 1.19 for a further description of boxplots.

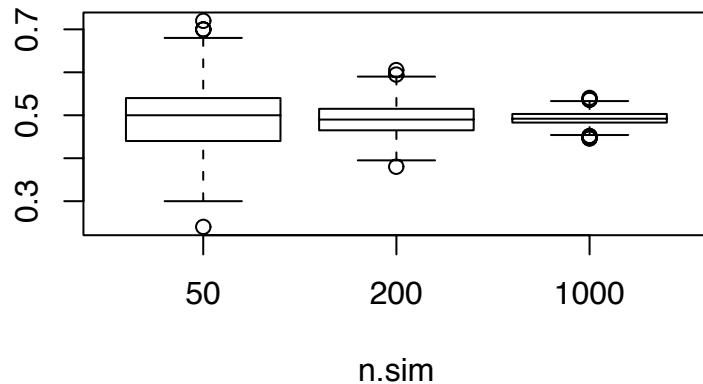


Figure 1.19: 1000 simulations of $\hat{\theta}$ for $n.sim = 50, 200, 1000$

Here is the R code for the simulations and Figure 1.19.

```
N <- 1000

n.sim <- c(50, 200, 1000)
```

```
theta.hat <- matrix ( NA, N, length(n.sim) )

for ( i in seq(along=n.sim) ) {
  for ( j in 1:N ) {
    wins <- 0
    for ( k in 1:n.sim[i] )
      wins <- wins + sim.craps()
    theta.hat[j,i] <- wins / n.sim[i]
  }
}

boxplot ( theta.hat ~ col(theta.hat), names=n.sim,
          xlab="n.sim" )
```

- `matrix` forms a matrix. The form is `matrix(x,nrows,ncols)` where `x` are the entries in the matrix and `nrows` and `ncols` are the numbers of rows and columns.
- `seq(along=n.sim)` is the same as `1:length(n.sim)` except that it behaves more sensibly in case `length(n.sim)` is 0.
- A `boxplot` is one way to display a data set. It produces a rectangle, or *box*, with a line through the middle. The rectangle contains the central 50% of the data. The line indicates the median of the data. Extending vertically above and below the box are dashed lines called *whiskers*. The whiskers contain most of the outer 50% of the data. A few extreme data points are plotted singly. See Example 2.3 for another use of boxplots and a fuller explanation.

1.7.2 Evaluating Statistical Procedures

Simulation can sometimes be useful in deciding whether a particular experiment is worthwhile or in choosing among several possible experiments or statistical procedures. For a fictitious example, consider ABC College, where, of the 10,000 students, 30% of the students are members of a sorority or fraternity (greeks) and 70% are independents. There is an upcoming election for Head of the student governance organization. Two candidates

are running, D and E. Let

$$\begin{aligned}\theta_G &= \text{proportion of greeks supporting D} \\ \theta_I &= \text{proportion of independents supporting D} \\ \theta &= \text{proportion of all students supporting D}\end{aligned}$$

A poll is commisioned to estimate θ and it is agreed that the pollster will sample 100 students. Three different procedures are proposed.

1. Randomly sample 100 students. Estimate

$$\hat{\theta}_1 = \text{proportion of polled students who favor D}$$

2. Randomly sample 100 students. Estimate

$$\begin{aligned}\hat{\theta}_G &= \text{proportion of polled greeks supporting D} \\ \hat{\theta}_I &= \text{proportion of polled independents supporting D} \\ \hat{\theta}_2 &= .3\hat{\theta}_G + .7\hat{\theta}_I\end{aligned}$$

3. Randomly sample 30 greeks and 70 independents. Estimate

$$\begin{aligned}\hat{\theta}_G &= \text{proportion of polled greeks supporting D} \\ \hat{\theta}_I &= \text{proportion of polled independents supporting D} \\ \hat{\theta}_3 &= .3\hat{\theta}_G + .7\hat{\theta}_I\end{aligned}$$

Which procedure is best? One way to answer the question is by exact calculation, but another way is by simulation. In the simulation we try each procedure many times to see how accurate it is, on average. We must choose some “true” values of θ_G , θ_I and θ under which to do the simulation. Here is some R code for the simulation.

```
# choose "true" theta.g and theta.i
theta.g <- .8
theta.i <- .4
prop.g <- .3
prop.i <- 1 - prop.g
theta <- prop.g * theta.g + prop.i * theta.i

sampszie <- 100
```

```

n.times <- 1000 # should be enough
theta.hat <- matrix ( NA, n.times, 3 )
for ( i in 1:n.times ) {
  theta.hat[i,1] <- sim.1()
  theta.hat[i,2] <- sim.2()
  theta.hat[i,3] <- sim.3()
}

print ( apply(theta.hat,2,mean) )
boxplot ( theta.hat ~ col(theta.hat) )

sim.1 <- function() {
  x <- rbinom(1,sampsize,theta)
  return ( x / sampsize )
}

sim.2 <- function() {
  n.g <- rbinom ( 1, sampsize, prop.g )
  n.i <- sampsize - n.g
  x.g <- rbinom ( 1, n.g, theta.g )
  x.i <- rbinom ( 1, n.i, theta.i )
  t.hat.g <- x.g / n.g
  t.hat.i <- x.i / n.i
  return ( prop.g * t.hat.g + (1-prop.g) * t.hat.i )
}

sim.3 <- function() {
  n.g <- sampsize * prop.g
  n.i <- sampsize * prop.i
  x.g <- rbinom ( 1, n.g, theta.g )
  x.i <- rbinom ( 1, n.i, theta.i )
  t.hat.g <- x.g / n.g
  t.hat.i <- x.i / n.i
  return ( prop.g * t.hat.g + (1-prop.g) * t.hat.i )
}

```

- `apply` applies a function to a matrix. In the code above, `apply(...)` applies the

function `mean` to dimension 2 of the matrix `theta.hat`. That is, it returns the mean of each column of `theta.hat`

The boxplot, shown in Figure 1.20 shows little practical difference between the three procedures.

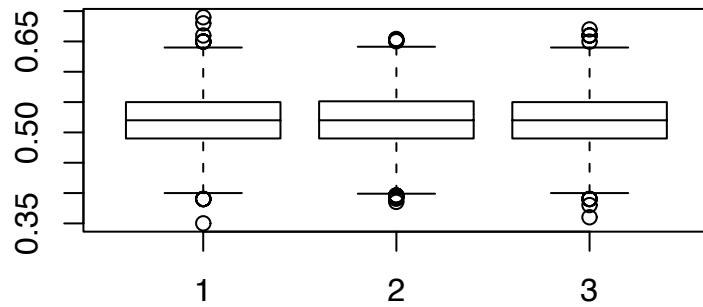


Figure 1.20: 1000 simulations of $\hat{\theta}$ under three possible procedures for conducting a poll

The next example shows how simulation was used to evaluate whether an experiment was worth carrying out.

Example 1.12 (FACE)

The amount of carbon dioxide, or CO_2 , in the Earth's atmosphere has been steadily increasing over the last century or so. You can see the increase yourself in the `co2` data set that comes with R. Typing `ts.plot(co2)` makes a time series plot, reproduced here as Figure 1.21. Typing `help(co2)` gives a brief explanation. The data are the concentrations of CO_2 in the atmosphere measured at Mauna Loa each month from 1959 to 1997. The plot shows a steadily increasing trend imposed on a regular annual cycle. The primary reason for the increase is burning of fossil fuels. CO_2 is a greenhouse gas that traps heat in the atmosphere instead of letting it radiate out, so an increase in atmospheric CO_2 will eventually result in an increase in the Earth's temperature. But what is harder to predict is the effect on the Earth's plants. Carbon

is a nutrient needed by plants. It is possible that an increase in CO₂ will cause an increase in plant growth which in turn will partly absorb some of the extra carbon.

To learn about plant growth under elevated CO₂, ecologists began by conducting experiments in greenhouses. In a greenhouse, two sets of plants could be grown under conditions that are identical except for the amount of CO₂ in the atmosphere. But the controlled environment of a greenhouse is quite unlike the uncontrolled natural environment, so, to gain verisimilitude, experiments soon moved to open-top chambers. An open-top chamber is a space, typically a few meters in diameter, enclosed by a solid, usually plastic, wall and open at the top. CO₂ can be added to the air inside the chamber. Because the chamber is mostly enclosed, not much CO₂ will escape, and more can be added as needed. Some plants can be grown in chambers with excess CO₂ others in chambers without and their growth compared. But as with greenhouses, open-top chambers are not completely satisfactory. Ecologists wanted to conduct experiments under even more natural conditions.

To that end, in the late 1980's the Office of Biological and Environmental Research in the U.S. Department of Energy (DOE) began supporting research using a technology called FACE, or *Free Air CO₂ Enrichment* developed at the Brookhaven National Laboratory. As the lab's webpage www.FACE.BNL.GOV/FACE1.HTM explains

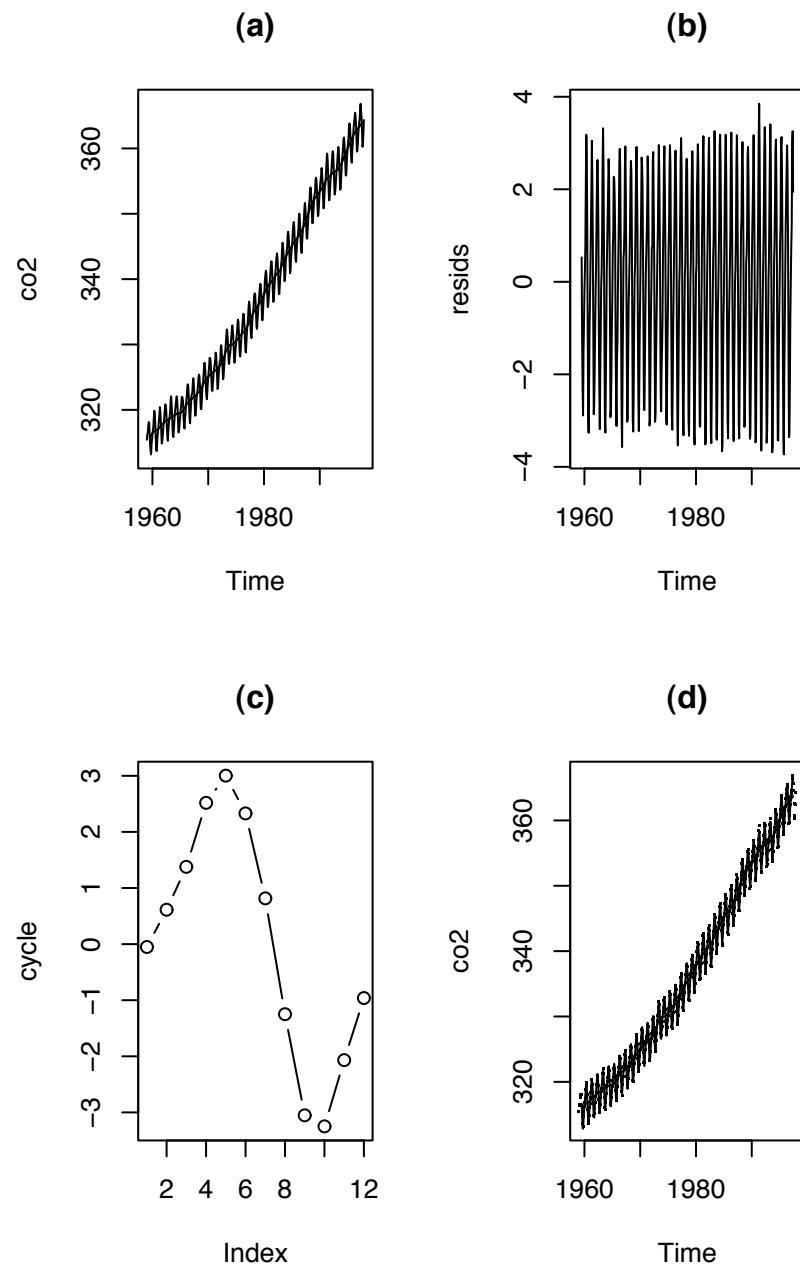
“FACE provides a technology by which the microclimate around growing plants may be modified to simulate climate change conditions. Typically CO₂-enriched air is released from a circle of vertical pipes into plots up to 30m in diameter, and as tall as 20 m.

“Fast feedback control and pre-dilution of CO₂ provide stable, elevated [CO₂] simulating climate change conditions.

“No containment is required with FACE equipment and there is no significant change in natural air-flow. Large FACE plots reduce effects of plot edge and capture fully-functioning, integrated ecosystem-scale processes. FACE Field data represent plant and ecosystems responses to concentrations of atmospheric CO₂ expected in the mid-twenty-first century.”

See the website for pictures and more information. In a FACE experiment, CO₂ is released into some treatment plots. The level of CO₂ inside the plot is continually monitored. More CO₂ is released as needed to keep the amount of CO₂ in the atmosphere at some prespecified level, typically the level that is expected in the mid-21st century. Other plots are reserved as control plots. Plant growth in the treatment plots is compared to that in the control plots.

Because a FACE site is not enclosed, CO₂ continually drifts out of the site and needs to be replenished. Keeping enough CO₂ in the air is very costly and is, in fact,

Figure 1.21: Monthly concentrations of CO₂ at Mauna Loa

the major expense in conducting FACE experiments.

The first several FACE sites were in Arizona (sorghum, wheat, cotton), Switzerland (rye grass, clover) and California (native chaparral). All of these contained low-growing plants. By the early 1990's, ecologists wanted to conduct a FACE experiment in a forest, and such an experiment was proposed by investigators at Duke University, to take place in Duke Forest. But before the experiment could be funded the investigators had to convince the Department of Energy (DOE) that it would be worthwhile. In particular, they wanted to demonstrate that the experiment would have a good chance of uncovering whatever growth differences would exist between treatment and control. The demonstration was carried out by computer simulation. The code for that demonstration, slightly edited for clarity, is given at the end of this Example and explained below.

1. The experiment would consist of 6 sites, divided into 3 pairs. One site in each pair would receive the CO₂ treatment; the other would be a control. The experiment was planned to run for 10 years. Investigators had identified 16 potential sites in Duke Forest. The above ground biomass of those sites, measured before the experiment began, is given in the line `b.mass <- c (...)`.
2. The code simulates 1000 repetitions of the experiment. That's the meaning of `nreps <- 1000`.
3. The above ground biomass of each site is stored in `M.actual.control` and `M.actual.treatment`. There must be room to store the biomass of each site for every combination of (pair,year,repetition). The `array(...)` command creates a multidimensional matrix, or `array`, filled with NA's. The dimensions are given by `c(npairs,nyears+1,nreps)`.
4. A site's actual biomass is not known exactly but is measured with error. The simulated measurements are stored in `M.observed.control` and `M.observed.treatment`.
5. Each repetition begins by choosing 6 sites from among the 16 available. Their observed biomass goes into `temp`. The first three values are assigned to `M.observed.control` and the last three to `M.observed.treatment`. All this happens in a loop `for(i in 1:nreps)`.
6. Investigators expected that control plots would grow at an average rate of 2% per year and treatment plots at an average of something else. Those values are called `betaC` and `betaT`. The simulation was run with `betaT = 1.04, 1.06, 1.08`

(shown below) and 1.10. Each site would have its own growth rate which would be slightly different from betaC or betaT. For control sites, those rates are drawn from the $N(\text{betaC}, 0.1 * (\text{betaC} - 1))$ distribution and stored in `beta.control`, and similarly for the treatment sites.

7. Measurement errors of biomass were expected to have an SD around 5%. That's `sigmaE`. But at each site in each year the measurement error would be slightly different. The measurement errors are drawn from the $N(1, \text{sigmaE})$ distribution and stored in `errors.control` and `errors.treatment`.
8. Next we simulate the actual biomass of the sites. For the first year where we already have measurements that's

```
M.actual.control [ , 1, ] <- ...
M.actual.treatment [ , 1, ] <- ...
```

For subsequent years the biomass in year i is the biomass in year $i-1$ multiplied by the growth factor `beta.control` or `beta.treatment`. Biomass is simulated in the loop `for(i in 2:(nyears+1))`.

9. Measured biomass is the actual biomass multiplied by measurement error. It is simulated by

```
M.observed.control <- ...
M.observed.treatment <- ...
```

10. The simulations for each year were analyzed each year by a *two-sample t-test* which looks at the ratio

$$\frac{\text{biomass in year } i}{\text{biomass in year 1}}$$

to see whether it is significantly larger for treatment sites than for control sites. See Section xyz for details about t-tests. For our purposes here, we have replaced the t-test with a plot, Figure 1.22, which shows a clear separation between treatment and control sites after about 5 years.

The DOE did decide to fund the proposal for a FACE experiment in Duke Forest, at least partly because of the demonstration that such an experiment would have a reasonably large chance of success.

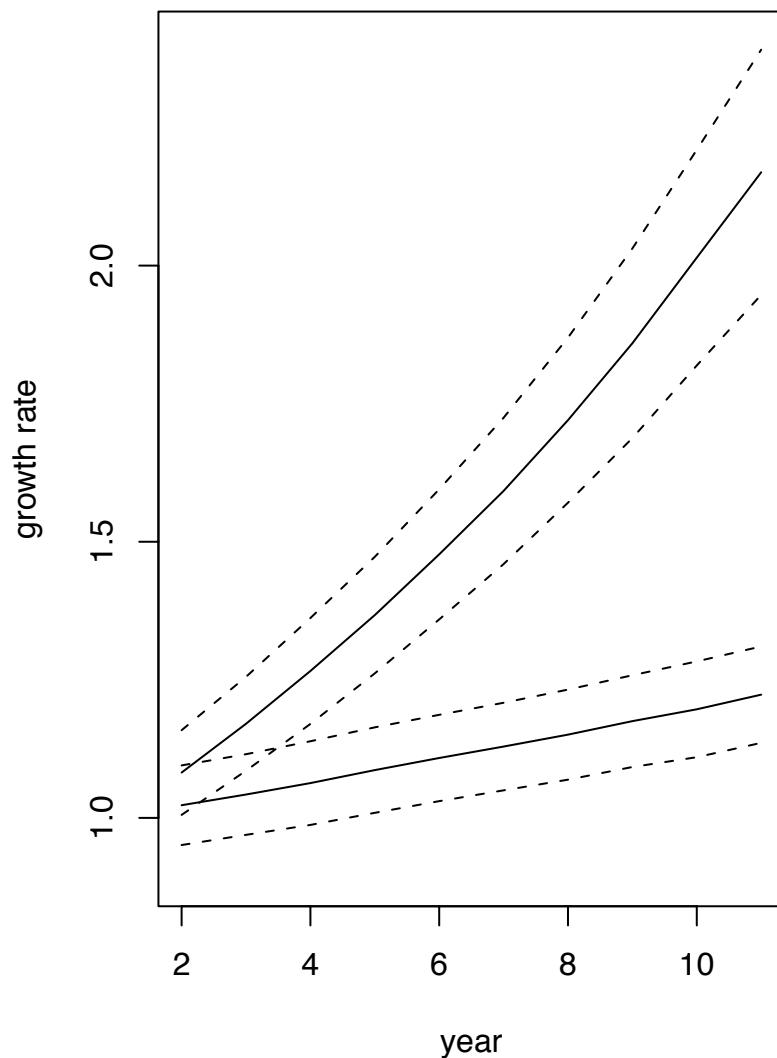


Figure 1.22: 1000 simulations of a FACE experiment. The x -axis is years. The y -axis shows the mean growth rate (biomass in year i / biomass in year 1) of control plants (lower solid line) and treatment plants (upper solid line). Standard deviations are shown as dashed lines.

```
#####
# A power analysis of the FACE experiment
#
# Initial measured biomass of potential FACE sites in g/m2:
b.mass <- c ( 17299.1, 17793.1, 23211.7, 23351.8, 24278,
                25335.9, 27001.5, 27113.6, 30184.3, 30625.5,
                33496.2, 33733.76, 35974.3, 38490.8, 40319.6,
                44903 )

npairs <- 3
nyears <- 10
nreps <- 1000

M.observed.control <- array ( NA, c(npairs,nyears+1,nreps) )
M.actual.control <- array ( NA, c(npairs,nyears+1,nreps) )
M.observed.treatment <- array ( NA, c(npairs,nyears+1,nreps) )
M.actual.treatment <- array ( NA, c(npairs,nyears+1,nreps) )

# Specify the initial levels of biomass
for ( i in 1:nreps ) {
    temp <- sample ( b.mass, size=2*npairs )
    M.observed.control [,1,i] <- temp [1:npairs]
    M.observed.treatment [,1,i] <- temp [(npairs+1):(2*npairs)]
}

# Specify the betas

betaC <- 1.02
betaT <- 1.08

beta.control <- matrix ( rnorm ( npairs*nreps, betaC,
                                    0.1*(betaC-1) ),
                           npairs, nreps )
beta.treatment <- matrix ( rnorm ( npairs*nreps, betaT,
                                       0.1*(betaT-1) ),
                            npairs, nreps )
```

```
#####
# measurement errors in biomass
sigmaE <- 0.05

errors.control <- array ( rnorm ( npairs*(nyears+1)*nreps,
                                1, sigmaE ),
                           c(npairs,nyears+1,nreps) )
errors.treatment <- array ( rnorm ( npairs*(nyears+1)*nreps,
                                    1, sigmaE ),
                            c(npairs,nyears+1,nreps) )
#####
#####
# Generate 10 years of data. The model for generation is:
# M.actual[i,j,] : above ground biomass in ring i, year j
# M.actual[i,j+1,] = beta[i] * M.actual[i,j,]

# We actually observe M.observed[i,j] = M.actual[i,j] * error
# Start with M.observed[i,1] and generate M.actual[i,1]

M.actual.control [ , 1, ] <-
  M.observed.control [ , 1, ] / errors.control[ , 1, ]

M.actual.treatment [ , 1, ] <-
  M.observed.treatment [ , 1, ] / errors.treatment[ , 1, ]

# Now, for the rest of the years, generate M.actual
for ( i in 2:(nyears+1) ) {
  M.actual.control [ , i, ] <-
    M.actual.control [ , i-1, ] * beta.control

  M.actual.treatment [ , i, ] <-
    M.actual.treatment [ , i-1, ] * beta.treatment
}

# The initial observed data, corresponding to the [,1,],
# doesn't need to be recomputed. But the following two
# statements are cleaner without subscripts.
M.observed.control <- M.actual.control * errors.control
```

```

M.observed.treatment <- M.actual.treatment * errors.treatment
#####
# two-sample t-test on (M.observed[j]/M.observed[1]) removed
# plot added

ratio.control <- matrix ( NA, nyears, npairs*nreps )
ratio.treatment <- matrix ( NA, nyears, npairs*nreps )

for ( i in 2:(nyears+1) ) {
  ratio.control [ i-1, ] <-
    as.vector ( M.observed.control[,i,]
                / M.observed.control[,1,] )
  ratio.treatment [ i-1, ] <-
    as.vector ( M.observed.treatment[,i,]
                / M.observed.treatment[,1,] )
}
}

mean.control <- apply ( ratio.control, 1, mean )
mean.treatment <- apply ( ratio.treatment, 1, mean )
sd.control <- sqrt ( apply ( ratio.control, 1, var ) )
sd.treatment <- sqrt ( apply ( ratio.treatment, 1, var ) )

plot ( 2:11, mean.control, type="l", ylim=c(.9,2.4),
       xlab="year", ylab="growth rate" )
lines ( 2:11, mean.treatment )
lines ( 2:11, mean.control - sd.control, lty=2 )
lines ( 2:11, mean.control + sd.control, lty=2 )
lines ( 2:11, mean.treatment - sd.treatment, lty=2 )
lines ( 2:11, mean.treatment + sd.treatment, lty=2 )

```

1.8 R

This section introduces a few more of the R commands we will need to work fluently with the software. They are introduced in the context of studying a dataset on the percent bodyfat of 252 men. You should download the data onto your own computer and try out the

analysis in R to develop your familiarity with what will prove to be a very useful tool. The data can be found at **StatLib**, an on-line repository of statistical data and software. The data were originally contributed by Roger Johnson of the Department of Mathematics and Computer Science at the South Dakota School of Mines and Technology. The **StatLib** website is lib.stat.cmu.edu. If you go to **StatLib** and follow the links to datasets and then **bodyfat** you will find a file containing both the data and an explanation. Copy just the data to a text file named **bodyfat.dat** on your own computer. The file should contain just the data; the first few lines should look like this:

```
1.0708    12.3      23 ...
1.0853     6.1      22 ...
1.0414    25.3      22 ...
```

The following snippet shows how to read the data into R and save it into **bodyfat**.

```
bodyfat <- read.table ("bodyfat.dat",
  col.names = c ( "density", "percent.fat", "age", "weight",
    "height", "neck.circum", "chest.circum", "abdomen.circum",
    "hip.circum", "thigh.circum", "knee.circum", "ankle.circum",
    "bicep.circum", "forearm.circum", "wrist.circum" ) )
dim   ( bodyfat )    # how many rows and columns in the dataset?
names ( bodyfat )    # names of the columns
```

- `read.table(...)` reads data from a text file into a **dataframe**. A **dataframe** is a flexible way to represent data because it can be treated as either a matrix or a list. Type `help(read.table)` to learn more. The first argument, `"bodyfat.dat"`, tells R what file to read. The second argument, `col.names = c ("density", ...)`, tells R the names of the columns.
- `dim` gives the dimension of a matrix, a **dataframe**, or anything else that has a dimension. For a matrix or **dataframe**, `dim` tells how many rows and columns.
- `names` gives the names of things. `names(bodyfat)` should tell us the names `density`, `percent.fat`, It's used here to check that the data were read the way we intended.

Individual elements of matrices can be accessed by two-dimensional subscripts such as `bodyfat[1,1]` or `bodyfat[3,7]` in which the subscripts refer to the row and column of the matrix. (Try this out to make sure you know how two dimensional subscripts work.) If the columns of the matrix have names, then the second subscript can be a name, as in

`bodyfat[1,"density"]` or `bodyfat[3,"chest.circum"]`. Often we need to refer to an entire column at once, which can be done by omitting the first subscript. For example, `bodyfat[,2]` refers to the entire set of 252 measurements of percent body fat.

A `dataframe` is a list of columns. Because `bodyfat` has 15 columns its length, `length(bodyfat)`, is 15. Members of a list can be accessed by subscripts with double brackets, as in `bodyfat[[1]]`. Each member of `bodyfat` is a vector of length 252. Individual measurements can be accessed as in `bodyfat[[1]][1]` or `bodyfat[[3]][7]`. If the list members have names, then they can be accessed as in `bodyfat$percent.fat`. Note the quotation marks used when treating `bodyfat` as a matrix and the lack of quotation marks when treating `bodyfat` as a list. The name following the "\$" can be abbreviated, as long as the abbreviation is unambiguous. Thus `bodyfat$ab` works, but `bodyfat$a` fails to distinguish between `age` and `abdomen.circum`.

Begin by displaying the data.

```
par ( mfrow=c(5,3) ) # establish a 5 by 3 array of plots

for ( i in 1:15 ) {
  hist ( bodyfat[[i]], xlab="", main=names(bodyfat)[i] )
}
```

Although it's not our immediate purpose, it's interesting to see what the relationships are among the variables. Try `pairs(bodyfat)`.

To illustrate some of R's capabilities and to explore the concepts of marginal, joint and conditional densities, we'll look more closely at percent fat and its relation to abdomen circumference. Begin with a histogram of percent fat.

```
fat <- bodyfat$per # give these two variables short names
abd <- bodyfat$abd # so we can refer to them easily
par ( mfrow=c(1,1) ) # just need one plot now, not 15
hist ( fat )
```

We'd like to rescale the vertical axis to make the area under the histogram equal to 1, as for a density. R will do that by drawing the histogram on a "density" scale instead of a "frequency" scale. While we're at it, we'll also make the labels prettier. We also want to draw a Normal curve approximation to the histogram, so we'll need the mean and standard deviation.

```
hist ( fat, xlab="", main="percent fat", freq=F )
mu <- mean ( fat )
sigma <- sqrt ( var ( fat ) ) # standard deviation
```

```

lo <- mu - 3*sigma
hi <- mu + 3*sigma
x <- seq ( lo, hi, length=50 )
lines ( x, dnorm ( x, mu, sigma ) )

```

That looks better, but we can do better still by slightly enlarging the axes. Redraw the picture, but use

```

hist ( fat, xlab="", main="percent fat", freq=F,
      xlim=c(-10, 60), ylim=c(0,.06) )

```

The Normal curve fits the data reasonably well. A good summary of the data is that it is distributed approximately $N(19.15, 8.37)$.

Now examine the relationship between abdomen circumference and percent body fat. Try the following command.

```

plot ( abd, fat, xlab="abdomen circumference",
       ylab="percent body fat" )

```

The scatter diagram shows a clear relationship between abdomen circumference and body fat in this group of men. One man doesn't fit the general pattern; he has a circumference around 148 but a body fat only around 35%, relatively low for such a large circumference. To quantify the relationship between the variables, let's divide the men into groups according to circumference and estimate the conditional distribution of fat given circumference. If we divide the men into twelfths we'll have 21 men per group.

```

cut.pts <- quantile ( abd, (0:12)/12 )
groups <- cut ( abd, cut.pts, include.lowest=T, labels=1:12 )
boxplot ( fat ~ groups,
           xlab="quantiles of abdomen circumference",
           ylab="percent body fat" )

```

Note:

- A *quantile* is a generalization of *median*. For example, the 1/12-th quantile of *abd* is the number q such that 1/12 of all the *abd* measurements are less than q and 11/12 are greater than q . (A more careful definition would say what to do in case of ties.) The median is the .5 quantile. We have cut our data according to the 1/12, 2/12, ..., 12/12 quantiles of *abd*.
- If you don't see what the `cut(abd, ...)` command does, print out `cut.pts` and `groups`, then look at them until you figure it out.

- *Boxplots* are a convenient way to compare different groups of data. In this case there are 12 groups. Each group is represented on the plot by a box with whiskers. The box spans the first and third quartiles (.25 and .75 quantiles) of `fat` for that group. The line through the middle of the box is the median `fat` for that group. The whiskers extend to cover most of the rest of the data. A few outlying `fat` values fall outside the whiskers; they are indicated as individual points.
- "`fat ~ groups`" is R's notation for a *formula*. It means to treat `fat` as a function of `groups`. Formulas are extremely useful and will arise repeatedly.

The medians increase in not quite a regular pattern. The irregularities are probably due to the vagaries of sampling. We can find the mean, median and variance of `fat` for each group with

```
mu.fat <- tapply ( fat, groups, mean )
me.fat <- tapply ( fat, groups, median )
sd.fat <- sqrt ( tapply ( fat, groups, var ) )
cbind ( mu.fat, me.fat, sd.fat )
```

- `tapply` means "apply to every element of a table." In this case, the table is `fat`, grouped according to `groups`.
- `cbind` means "bind together in columns". There is an analogous command `rbind`.

Finally, let's make a figure similar to Figure 1.12.

```
x <- seq ( 0, 50, by=1 )
par ( mfrow=c(4,3) )
for ( i in 1:12 ) {
  good <- groups == i
  hist ( fat[good], xlim=c(0,50), ylim=c(0,.1),
         breaks=seq(0,50,by=5), freq=F,
         xlab="percent fat", main="" )
  y <- dnorm ( x, mu.fat[i], sd.fat[i] )
  lines ( x, y )
}
```

The Normal curves seem to fit well. We saw earlier that the marginal (*Marginal* means unconditional.) distribution of percent body fat is well approximated by $N(19.15, 8.37)$. Here we see that the conditional distribution of percent body fat, given that abdomen circumference is in between the $(i - 1)/12$ and $i/12$ quantiles is $N(\text{mu.fat}[i], \text{sd.fat}[i])$.

If we know a man's abdomen circumference even approximately then (1) we can estimate his percent body fat more accurately and (2) the typical estimation error is smaller. [add something about estimation error in the sd section]

1.9 Some Results for Large Samples

It is intuitively obvious that large samples are better than small, that more data is better than less, and, less obviously, that with enough data one is eventually led to the right answer. These intuitive observations have precise mathematical statements in the form of Theorems 1.12, 1.13 and 1.14. We state those theorems here so we can use them throughout the rest of the book. They are examined in more detail in Section 8.4.

Definition 1.9 (Random Sample). A collection y_1, \dots, y_n of random variables is said to be a *random sample* of size n from pdf or pmf f if and only if

1. $y_i \sim f$ for each $i = 1, 2, \dots, n$ and
2. the y_i 's are mutually independent, i.e. $f(y_1, \dots, y_n) = \prod_{i=1}^n f(y_i)$.

The collection (y_1, \dots, y_n) is called a *data set*. We write

$$y_1, \dots, y_n \sim \text{i.i.d. } f$$

where i.i.d. stands for *independent and identically distributed*.

Many introductory statistics texts describe how to collect random samples, many pitfalls that await, and many pratfalls taken in the attempt. We omit that discussion here and refer the interested reader to our favorite introductory text on the subject, FREEDMAN ET AL. [1998] which has an excellent description of random sampling in general as well as detailed discussion of the US census and the Current Population Survey.

Suppose $y_1, y_2, \dots, \sim \text{i.i.d. } f$. Let $\mu = \int y f(y) dy$ and $\sigma^2 = \int (y - \mu)^2 f(y) dy$ be the mean and variance of f . Typically μ and σ are unknown and we take the sample in order to learn about them. We will often use $\bar{y}_n \equiv (y_1 + \dots + y_n)/n$, the mean of the first n observations, to estimate μ . Some questions to consider are

- For a sample of size n , how accurate is \bar{y}_n as an estimate of μ ?
- Does \bar{y}_n get closer to μ as n increases?
- How large must n be in order to achieve a desired level of accuracy?

Theorems 1.12 and 1.14 provide answers to these questions. Before stating them we need some preliminary results about the mean and variance of \bar{y}_n .

Theorem 1.7. Let x_1, \dots, x_n be random variables with means μ_1, \dots, μ_n . Then $\mathbb{E}[x_1 + \dots + x_n] = \mu_1 + \dots + \mu_n$.

Proof. It suffices to prove the case $n = 2$.

$$\begin{aligned}\mathbb{E}[x_1 + x_2] &= \int \int (x_1 + x_2) f(x_1, x_2) dx_1 dx_2 \\ &= \int \int x_1 f(x_1, x_2) dx_1 dx_2 + \int \int x_2 f(x_1, x_2) dx_1 dx_2 \\ &= \mu_1 + \mu_2\end{aligned}$$

□

Corollary 1.8. Let y_1, \dots, y_n be a random sample from f with mean μ . Then $\mathbb{E}[\bar{y}_n] = \mu$.

Proof. The corollary follows from Theorems 1.3 and 1.7. □

Theorem 1.9. Let x_1, \dots, x_n be independent random variables with means μ_1, \dots, μ_n and SDs $\sigma_1, \dots, \sigma_n$. Then $\text{Var}[x_1 + \dots + x_n] = \sigma_1^2 + \dots + \sigma_n^2$.

Proof. It suffices to prove the case $n = 2$. Using Theorem 1.2,

$$\begin{aligned}\text{Var}(X_1 + X_2) &= \mathbb{E}((X_1 + X_2)^2) - (\mu_1 + \mu_2)^2 \\ &= \mathbb{E}(X_1^2) + 2\mathbb{E}(X_1 X_2) + \mathbb{E}(X_2^2) - \mu_1^2 - 2\mu_1\mu_2 - \mu_2^2 \\ &= (\mathbb{E}(X_1^2) - \mu_1^2) + (\mathbb{E}(X_2^2) - \mu_2^2) + 2(\mathbb{E}(X_1 X_2) - \mu_1\mu_2) \\ &= \sigma_1^2 + \sigma_2^2 + 2(\mathbb{E}(X_1 X_2) - \mu_1\mu_2).\end{aligned}$$

But if $X_1 \perp X_2$ then

$$\begin{aligned}\mathbb{E}(X_1 X_2) &= \int \int x_1 x_2 f(x_1, x_2) dx_1 dx_2 \\ &= \int x_1 \int x_2 f(x_2) dx_2 f(x_1) dx_1 \\ &= \mu_2 \int x_1 f(x_1) dx_1 = \mu_1\mu_2.\end{aligned}$$

So $\text{Var}(X_1 + X_2) = \sigma_1^2 + \sigma_2^2$. □

Note that Theorem 1.9 requires independence while Theorem 1.7 does not.

Corollary 1.10. Let y_1, \dots, y_n be a random sample from f with variance σ^2 . Then $\text{Var}(\bar{y}_n) = \sigma^2/n$.

Proof. The corollary follows from Theorems 1.4 and 1.9. \square

Theorem 1.11 (Chebychev's Inequality). *Let X be a random variable with mean μ and SD σ . Then for any $\epsilon > 0$,*

$$P[|X - \mu| \geq \epsilon] \leq \sigma^2/\epsilon^2.$$

Proof.

$$\begin{aligned} \sigma^2 &= \int (x - \mu)^2 f(x) dx \\ &= \int_{-\infty}^{\mu-\epsilon} (x - \mu)^2 f(x) dx + \int_{\mu-\epsilon}^{\mu+\epsilon} (x - \mu)^2 f(x) dx + \int_{\mu+\epsilon}^{\infty} (x - \mu)^2 f(x) dx \\ &\geq \int_{-\infty}^{\mu-\epsilon} (x - \mu)^2 f(x) dx + \int_{\mu+\epsilon}^{\infty} (x - \mu)^2 f(x) dx \\ &\geq \epsilon^2 \int_{-\infty}^{\mu-\epsilon} f(x) dx + \epsilon^2 \int_{\mu+\epsilon}^{\infty} f(x) dx \\ &= \epsilon^2 P[|X - \mu| \geq \epsilon]. \end{aligned}$$

\square

Theorems 1.12 and 1.14 are the two main limit theorems of statistics. They provide answers, at least probabilistically, to the questions on page 76.

Theorem 1.12 (Weak Law of Large Numbers). *Let y_1, \dots, y_n be a random sample from a distribution with mean μ and variance σ^2 . Then for any $\epsilon > 0$,*

$$\lim_{n \rightarrow \infty} P[|\bar{y}_n - \mu| < \epsilon] = 1. \quad (1.18)$$

Proof. Apply Chebychev's Inequality (Theorem 1.11) to \bar{y}_n .

$$\lim_{n \rightarrow \infty} P[|\bar{y}_n - \mu| < \epsilon] = \lim_{n \rightarrow \infty} 1 - P[|\bar{y}_n - \mu| \geq \epsilon] \geq \lim_{n \rightarrow \infty} 1 - \sigma^2/n\epsilon^2 = 1.$$

\square

Another version of Theorem 1.12 is called the Strong Law of Large Numbers.

Theorem 1.13 (Strong Law of Large Numbers). *Let y_1, \dots, y_n be a random sample from a distribution with mean μ and variance σ^2 . Then for any $\epsilon > 0$,*

$$P[\lim_{n \rightarrow \infty} |\bar{Y}_n - \mu| < \epsilon] = 1;$$

i.e.,

$$P[\lim_{n \rightarrow \infty} \bar{Y}_n = \mu] = 1.$$

It is beyond the scope of this section to explain the difference between the WLLN and the SLLN. See Section 8.4.

Theorem 1.14 (Central Limit Theorem). *Let y_1, \dots, y_n be a random sample from f with mean μ and variance σ^2 . Let $z_n = (\bar{y}_n - \mu)/(\sigma/\sqrt{n})$. Then, for any numbers $a < b$,*

$$\lim_{n \rightarrow \infty} P[z_n \in [a, b]] = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw.$$

I.e. The limiting distribution of z_n is $N(0, 1)$.

The Law of Large Numbers is what makes simulations work and why large samples are better than small. It says that as the number of simulation grows or as the sample size grows, ($n \rightarrow \infty$), the average of the simulations or the average of the sample gets closer and closer to the true value ($\bar{X}_n \rightarrow \mu$). For instance, in Example 1.11, where we used simulation to estimate $P[\text{Shooter wins}]$ in Craps, the estimate became more and more accurate as the number of simulations increased from 50, to 200, and then to 1000. The Central Limit Theorem helps us look at those simulations more closely.

Colloquially, the Central Limit Theorem says that

$$\lim_{n \rightarrow \infty} p_{z_n} = N(0, 1).$$

Its practical application is that, for large n ,

$$p_{z_n} \approx N(0, 1),$$

which in turn means

$$p_{\bar{y}_n} \approx N(\mu, \sigma/\sqrt{n}).$$

In Example 1.11 we simulated the game of Craps for $n.\text{sim} = 50, 200$, and 1000 . Those simulations are depicted in Figure 1.23. The upper panel shows a histogram of 1000 simulations, all using $n.\text{sim} = 50$. For a single simulation with $n.\text{sim} = 50$ let X_1, \dots, X_{50} be the outcomes of those simulations. Each $X_i \sim \text{Bern}(.493)$, so $\mu = .493$ and $\sigma = \sqrt{.493 * .507} \approx .5$. Therefore, according to the Central Limit Theorem, when $n.\text{sim} = 50$

$$\bar{X}_n \sim N(\mu, \sigma/\sqrt{n}) \approx N(.493, .071)$$

This is the Normal density plotted in the upper panel of Figure 1.23. We see that the $N(.493, .071)$ is a good approximation to the histogram. And that's because $\hat{\theta} = \bar{X}_{50} \sim N(.493, .071)$, approximately. The Central Limit Theorem says that the approximation will be good for “large” n . In this case $n = 50$ is large enough. (Section ?? will discuss the question of when n is “large enough”).

Similarly,

$$\begin{aligned}\bar{X}_n &\sim N(.493, .035) && \text{when } n.\text{sim} = 200 \\ \bar{X}_n &\sim N(.493, .016) && \text{when } n.\text{sim} = 1000.\end{aligned}$$

These densities are plotted in the middle and lower panels of Figure 1.23.

The Central Limit Theorem makes three statements about the distribution of \bar{y}_n (z_n) in large samples:

1. $E[\bar{y}_n] = \mu$ ($E[z_n] = 0$),
2. $SD(\bar{y}_n) = \sigma / \sqrt{n}$ ($SD(z_n) = 1$), and
3. \bar{y}_n (z_n) has, approximately, a Normal distribution.

The first two of these are already known from Theorems 1.7 and 1.9. It's the third point that is key to the Central Limit Theorem. Another surprising implication from the Central Limit Theorem is that the distributions of \bar{y}_n and z_n in large samples are determined solely by μ and σ ; no other features of f matter.

1.10 Exercises

1. Show: if μ is a probability measure then for any integer $n \geq 2$, and disjoint sets A_1, \dots, A_n

$$\mu\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mu(A_i).$$
2. **Simulating Dice Rolls**
 - (a) simulate 6000 dice rolls. Count the number of 1's, 2's, ..., 6's.
 - (b) You expect about 1000 of each number. How close was your result to what you expected?
 - (c) About how often would you expect to get more than 1030 1's? Run an R simulation to estimate the answer.
3. **The Game of Risk** In the board game *Risk* players place their armies in different countries and try eventually to control the whole world by capturing countries one at a time from other players. To capture a country, a player must attack it from an adjacent country. If player A has $A \geq 2$ armies in country A, she may attack adjacent

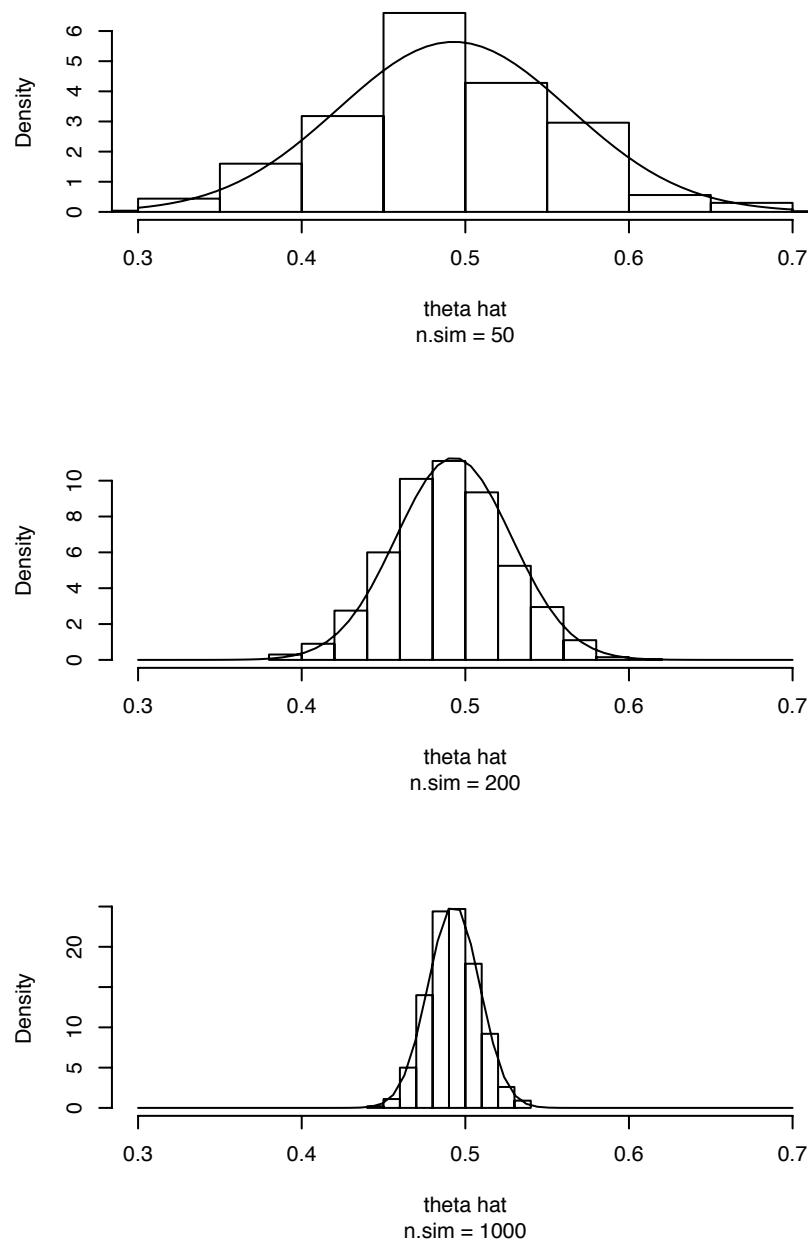


Figure 1.23: Histograms of craps simulations. Solid curves are Normal approximations according to the Central Limit Theorem.

country D . Attacks are made with from 1 to 3 armies. Since at least 1 army must be left behind in the attacking country, A may choose to attack with a minimum of 1 and a maximum of $\min(3, A - 1)$ armies. If player D has $D \geq 1$ armies in country D , he may defend himself against attack using a minimum of 1 and a maximum of $\min(2, D)$ armies. It is almost always best to attack and defend with the maximum permissible number of armies.

When player A attacks with a armies she rolls a dice. When player D defends with d armies he rolls d dice. A's highest die is compared to D's highest. If both players use at least two dice, then A's second highest is also compared to D's second highest. For each comparison, if A's die is higher than D's then A wins and D removes one army from the board; otherwise D wins and A removes one army from the board. When there are two comparisons, a total of two armies are removed from the board.

- If A attacks with one army (she has two armies in country A, so may only attack with one) and D defends with one army (he has only one army in country D) what is the probability that A will win?
 - Suppose that Player 1 has two armies each in countries C_1, C_2, C_3 and C_4 , that Player 2 has one army each in countries B_1, B_2, B_3 and B_4 , and that country C_i attacks country B_i . What is the chance that Player 1 will be successful in at least one of the four attacks?
4. (a) Justify the last step of Equation 1.2.
 (b) Justify the last step of the proof of Theorem 1.1.
 (c) Prove Theorem 1.1 when g is a decreasing function.
 5. Y is a random variable. $Y \in (-1, 1)$. The pdf is $p(y) = ky^2$ for some constant, k .
 - (a) Find k .
 - (b) Use R to plot the pdf.
 - (c) Let $Z = -Y$. Find the pdf of Z . Plot it.
 6. U is a random variable on the interval $[0, 1]$; $p(u) = 1$.
 - (a) $V = U^2$. On what interval does V live? Plot V as a function of U . Find the pdf of V . Plot $p_V(v)$ as a function of v .
 - (b) $W = 2U$. On what interval does W live? Plot W as a function of U . Find the pdf of W . Plot $p_W(w)$ as a function of w .

- (c) $X = -\log(U)$. On what interval does X live? Plot X as a function of U . Find the pdf of X . Plot $p_X(x)$ as a function of x .
7. Let $X \sim \text{Exp}(\lambda)$ and let $Y = cX$ for some constant c .
- Write down the density of X .
 - Find the density of Y .
 - Name the distribution of Y .
8. A teacher randomly selects a student from a Sta 103 class. Let X be the number of math courses the student has completed. Let $Y = 1$ if the student is female and $Y = 0$ if the student is male. Fifty percent of the class is female. Among the women, thirty percent have completed one math class, forty percent have completed two math classes and thirty percent have completed three. Among the men, thirty percent have completed one math class, fifty percent have completed two math classes and twenty percent have completed three.
- True or False:** X and Y are independent.
 - Find $E[X|Y = 1]$.
9. Sue is studying the $\text{Bin}(25,.4)$ distribution. In R she types
- ```
y <- rbinom(50, 25, .4)
m1 <- mean(y)
m2 <- sum(y) / 25
m3 <- sum ((y-m1)^2) / 50
```
- Is  $y$  a number, a vector or a matrix?
  - What is the approximate value of  $m1$ ?
  - What is the approximate value of  $m2$ ?
  - What was Sue trying to estimate with  $m3$ ?
10. The random variables  $X$  and  $Y$  have joint pdf  $f_{X,Y}(x,y) = 1$  in the triangle of the  $XY$ -plane determined by the points  $(-1,0)$ ,  $(1,0)$ , and  $(0,1)$ . **Hint: Draw a picture.**
- Find  $f_X(.5)$ .
  - Find  $f_Y(y)$ .

- (c) Find  $f_{Y|X}(y|X = .5)$ .  
 (d) Find  $E[Y|X = .5]$ .  
 (e) Find  $f_Y(.5)$ .  
 (f) Find  $f_X(x)$ .  
 (g) Find  $f_{X|Y}(x|Y = .5)$ .  
 (h) Find  $E[X|Y = .5]$ .
11.  $X$  and  $Y$  are uniformly distributed in the unit disk. I.e., the joint density  $p(x, y)$  is constant on the region of  $\mathbb{R}^2$  such that  $x^2 + y^2 \leq 1$ .
- (a) Find  $p(x, y)$ .  
 (b) Are  $X$  and  $Y$  independent?  
 (c) Find the marginal densities  $p(x)$  and  $p(y)$ .  
 (d) Find the conditional densities  $p(x|y)$  and  $p(y|x)$ .  
 (e) Find  $E[X]$ ,  $E[X|Y = .5]$ , and  $E[X|Y = -.5]$ .
12. Verify the claim in Example 1.4 that  $\operatorname{argmax}_\lambda P[x = 3|\lambda] = 3$ . Hint: differentiate Equation 1.6.
13. (a)  $p$  is the pdf of a continuous random variable  $w$ . Find  $\int_{\mathbb{R}} p(s) ds$ .  
 (b) Find  $\int_{\mathbb{R}} p(s) ds$  for the pdf in Equation 1.7.
14. Page 7 says “Every pdf must satisfy two properties ...” and that one of them is “ $p(y) \geq 0$  for all  $y$ .” Explain why that’s not quite right.
15. 
$$p(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$$
  
 is the pdf of a continuous random variable  $y$ . Find  $\int_{-\infty}^0 p(s) ds$ .
16. When spun, an unbiased spinner points to some number  $y \in (0, 1]$ . What is  $p(y)$ ?
17. Some exercises on the densities of tranformed variables. One of them should illustrate the need for the absolute value of the Jacobian.
18. (a) Prove: if  $X \sim \text{Poi}(\lambda)$  then  $E(X) = \lambda$ . Hint: use the same trick we used to derive the mean of the Binomial distribution.

- (b) Prove: if  $X \sim N(\mu, \sigma)$  then  $\mathbb{E}(X) = \mu$ . Hint: change variables in the integral.
19. (a) Prove: if  $X \sim \text{Bin}(n, p)$  then  $\text{Var}(X) = np(1 - p)$ . Hint: use Theorem 1.9.  
 (b) Prove: if  $X \sim \text{Poi}(\lambda)$  then  $\text{Var}(X) = \lambda$ . Hint: use the same trick we used to derive the mean of the Binomial distribution and Theorem 1.2.  
 (c) If  $X \sim \text{Exp}(\lambda)$ , find  $\text{Var}(X)$ . Hint: use Theorem 1.2.  
 (d) If  $X \sim N(\mu, \sigma)$ , find  $\text{Var}(X)$ . Hint: use Theorem 1.2.
20. (a) Justify each step of Equation 1.12.  
 (b) Justify each step of Equation 1.13. (Hint: integrate by parts.)
21. Let  $X_1 \sim \text{Bin}(50, .1)$ ,  $X_2 \sim \text{Bin}(50, .9)$  and  $X_1 \perp X_2$ . Define  $Y = X_1 + X_2$ . Does  $Y$  have the  $\text{Bin}(100, .5)$  distribution? Why or why not?
22. Let  $X_1 \sim \text{Bin}(50, .5)$ ,  $X_2 \sim \text{Bin}(50, .5)$  and  $X_1 \perp X_2$ . Define  $Y_1 = X_1 + X_2$  and  $Y_2 = 2X_1$ . Which has the bigger mean:  $Y_1$  or  $Y_2$ ? Which has the bigger variance:  $Y_1$  or  $Y_2$ ? Justify your answer.
23. Consider customers arriving at a service counter. Interarrival times often have a distribution that is approximately exponential with a parameter  $\lambda$  that depends on conditions specific to the particular counter. I.e.,  $p(t) = \lambda e^{-\lambda t}$ . Assume that successive interarrival times are independent of each other. Let  $T_1$  be the arrival time of the next customer and  $T_2$  be the additional time until the arrival of the second customer.
- (a) What is the joint density of  $(T_1, T_2)$ ?
  - (b) Let  $S = T_1 + T_2$ , the time until the next two customers arrive. What is  $P[S \leq 5]$ ; i.e. the probability that at least 2 customers arrive within the next 5 minutes?
  - (c) What is  $\mathbb{E}(S)$ ?
24. A gambler plays at a roulette table for two hours betting on Red at each spin of the wheel. There are 60 spins during the two hour period. What is the distribution of
- (a)  $z$ , the number of times the gambler wins,
  - (b)  $y$ , the number of times the gambler loses,
  - (c)  $x$ , the number of times the gambler is jostled by the person standing behind,
  - (d)  $w$ , the gambler's net gain?

25. If human DNA contains  $xxx$  bases, and if each base mutates with probability  $p$  over the course of a lifetime, what is the average number of mutations per person? What is the variance of the number of mutations per person?
26. Isaac is in 5th grade. Each sentence he writes for homework has a 90% chance of being grammatically correct. The correctness of one sentence does not affect the correctness of any other sentence. He recently wrote a 10 sentence paragraph for a writing assignment. Write a formula for the chance that no more than two sentences are grammatically incorrect.
27. Teams A and B play each other in the World Series of baseball. Team A has a 60% chance of winning each game. What is the chance that B wins the series? (The winner of the series is the first team to win 4 games.)
28. A basketball player shoots ten free throws in a game. She has a 70% chance of making each shot. If she misses the shot, her team has a 30% chance of getting the rebound.
  - (a) Let  $m$  be the number of shots she makes. What is the distribution of  $m$ ? What are its expected value and variance? What is the chance that she makes somewhere between 5 and 9 shots, inclusive?
  - (b) Let  $r$  be the number of rebounds her team gets from her free throws. What is the distribution of  $r$ ? What are its expected value and variance? What is the chance that  $r \geq 1$ ?
29. Let  $(x, y)$  have joint density function  $f_{x,y}$ . There are two ways to find  $\mathbb{E}y$ . One way is to evaluate  $\iint y f_{x,y}(x, y) dx dy$ . The other is to start with the joint density  $f_{x,y}$ , find the marginal density  $f_y$ , then evaluate  $\int y f_y(y) dy$ . Show that these two methods give the same answer.
30. Prove Theorem 1.3 (pg. 40) in the discrete case.
31. Prove Theorem 1.7 (pg. 77) in the continuous case.
32. A researcher randomly selects mother-daughter pairs. Let  $x_i$  and  $y_i$  be the heights of the  $i$ 'th mother and daughter, respectively. True or False:
  - (a)  $x_i$  and  $x_j$  are independent
  - (b)  $x_i$  and  $y_j$  are independent
  - (c)  $y_i$  and  $y_j$  are independent

- (d)  $x_i$  and  $y_i$  are independent
33. As part of his math homework Isaac had to roll two dice and record the results. Let  $X_1$  be the result of the first die and  $X_2$  be the result of the second. What is the probability that  $X_1=1$  given that  $X_1 + X_2 = 5$ ?
34. A doctor suspects a patient has the rare medical condition DS, or disstaticularia, the inability to learn statistics. DS occurs in .01% of the population, or one person in 10,000. The doctor orders a diagnostic test. The test is quite accurate. Among people who have DS the test yields a positive result 99% of the time. Among people who do not have DS the test yields a positive result only 5% of the time.
- For the patient in question, the test result is positive. Calculate the probability that the patient has DS.
35. Ecologists are studying salamanders in a forest. There are two types of forest. Type A is conducive to salamanders while type B is not. They are studying one forest but don't know which type it is. Types A and B are equally likely.
- During the study, they randomly sample quadrats. (A quadrat is a square-meter plot.) In each quadrat they count the number of salamanders. Some quadrats have poor salamander habitat. In those quadrats the number of salamanders is 0. Other quadrats have good salamander habitat. In those quadrats the number of salamanders is either 0, 1, 2, or 3, with probabilities 0.1, 0.3, 0.4, and 0.2, respectively. (Yes, there might be no salamanders in a quadrat with good habitat.) In a type A forest, the probability that a quadrat is good is 0.8 and the probability that it is poor is 0.2. In a type B forest the probability that a quadrat is good is 0.3 and the probability that it is poor is 0.7.
- (a) On average, what is the probability that a quadrat is good?
  - (b) On average, what is the probability that a quadrat has 0 salamanders, 1 salamander, 2 salamanders, 3 salamanders?
  - (c) The ecologists sample the first quadrat. It has 0 salamanders. What is the probability that the quadrat is good?
  - (d) Given that the quadrat had 0 salamanders, what is the probability that the forest is type A?
  - (e) Now the ecologists prepare to sample the second quadrat. Given the results from the first quadrat, what is the probability that the second quadrat is good?
  - (f) Given the results from the first quadrat, what is the probability that they find no salamanders in the second quadrat?

36. For various reasons, researchers often want to know the number of people who have participated in embarrassing activities such as illegal drug use, cheating on tests, robbing banks, etc. An opinion poll which asks these questions directly is likely to elicit many untruthful answers. To get around the problem, researchers have devised the method of randomized response. The following scenario illustrates the method.

A pollster identifies a respondent and gives the following instructions. "Toss a coin, but don't show it to me. If it lands Heads, answer question (a). If it lands tails, answer question (b). Just answer 'yes' or 'no'. Do not tell me which question you are answering.

Question (a): Does your telephone number end in an even digit?

Question (b): Have you ever used cocaine?"

Because the respondent can answer truthfully without revealing his or her cocaine use, the incentive to lie is removed. Researchers hope respondents will tell the truth.

You may assume that respondents are truthful and that telephone numbers are equally likely to be odd or even. Let  $p$  be the probability that a randomly selected person has used cocaine.

- (a) What is the probability that a randomly selected person answers "yes"?
  - (b) Suppose we survey 100 people. Let  $X$  be the number who answer "yes". What is the distribution of  $X$ ?
37. In a 1991 article (See Utts [1991] and discussants.) Jessica Utts reviews some of the history of probability and statistics in ESP research. This question concerns a particular series of *autoganzfeld* experiments in which a sender looking at a picture tries to convey that picture telepathically to a receiver. Utts explains:

"... 'autoganzfeld' experiments require four participants. The first is the Receiver (R), who attempts to identify the target material being observed by the Sender (S). The Experimenter (E) prepares R for the task, elicits the response from R and supervises R's judging of the response against the four potential targets. (Judging is double blind; E does not know which is the correct target.) The fourth participant is the lab assistant (LA) whose only task is to instruct the computer to randomly select the target. No one involved in the experiment knows the identity of the target.

"Both R and S are sequestered in sound-isolated, electrically shielded rooms. R is prepared as in earlier ganzfeld studies, with white noise and a field of red light. In a nonadjacent room, S watches the target material

on a television and can hear R's target description ('mentation') as it is being given. The mentation is also tape recorded.

"The judging process takes place immediately after the 30-minute sending period. On a TV monitor in the isolated room, R views the four choices from the target pack that contains the actual target. R is asked to rate each one according to how closely it matches the ganzfeld mentation. The ratings are converted to ranks and, if the correct target is ranked first, a direct hit is scored. The entire process is automatically recorded by the computer. The computer then displays the correct choice to R as feedback."

In the series of autoganzfeld experiments analyzed by Utts, there were a total of 355 trials. Let  $X$  be the number of direct hits.

- (a) What are the possible values of  $X$ ?
  - (b) Assuming there is no ESP, and no cheating, what is the distribution of  $X$ ?
  - (c) Plot the pmf of the distribution in part (b).
  - (d) Find  $\mathbb{E}[X]$  and  $\text{SD}(X)$ .
  - (e) Add a Normal approximation to the plot in part (c).
  - (f) Judging from the plot in part (c), approximately what values of  $X$  are consistent with the "no ESP, no cheating" hypothesis?
  - (g) In fact, the total number of hits was  $x = 122$ . What do you conclude?
38. This exercise is based on a computer lab that another professor uses to teach the Central Limit Theorem. It was originally written in MATLAB but here it's translated into R.

Enter the following R commands:

```
u <- matrix (runif(250000), 1000, 250)
y <- apply (u, 2, mean)
```

These create a 1000x250 (a thousand rows and two hundred fifty columns) matrix of random draws, called  $u$  and a 250-dimensional vector  $y$  which contains the means of each column of  $U$ .

Now enter the command `hist(u[,1])`. This command takes the first column of  $u$  (a column vector with 1000 entries) and makes a histogram. Print out this histogram

and describe what it looks like. What distribution is the `rnorm` command drawing from?

Now enter the command `hist(y)`. This command makes a histogram from the vector `y`. Print out this histogram. Describe what it looks like and how it differs from the one above. Based on the histogram, what distribution do you think `y` follows?

You generated `y` and `u` with the *same* random draws, so how can they have different distributions? What's going on here?

39. Suppose that extensive testing has revealed that people in Group A have IQ's that are well described by a  $N(100, 10)$  distribution while the IQ's of people in Group B have a  $N(105, 10)$  distribution. *What is the probability that a randomly chosen individual from Group A has a higher IQ than a randomly chosen individual from Group B?*
  - (a) Write a formula to answer the question. You don't need to evaluate the formula.
  - (b) Write some R code to answer the question.
40. The so-called *Monty Hall* or *Let's Make a Deal* problem has caused much consternation over the years. It is named for an old television program. A contestant is presented with three doors. Behind one door is a fabulous prize; behind the other two doors are virtually worthless prizes. The contestant chooses a door. The host of the show, Monty Hall, then opens one of the remaining two doors, revealing one of the worthless prizes. Because Monty is the host, he knows which doors conceal the worthless prizes and always chooses one of them to reveal, but never the door chosen by the contestant. Then the contestant is offered the choice of keeping what is behind her original door or trading for what is behind the remaining unopened door. What should she do?

There are two popular answers.

- There are two unopened doors, they are equally likely to conceal the fabulous prize, so it doesn't matter which one she chooses.
- She had a  $1/3$  probability of choosing the right door initially, a  $2/3$  chance of getting the prize if she trades, so she should trade.
  - (a) Create a simulation in R to discover which answer is correct.
  - (b) Show using formal arguments of conditional probability which answer is correct.

Make sure your answers to (a) and (b) agree!

41. Prove Theorem 1.6 (pg. 53).
42. Theorem 1.11, Chebychev's Inequality, gives an upper bound for how far a random variable can vary from its mean. Specifically, the theorem states that for any random variable  $X$ , with mean  $\mu$  and SD  $\sigma$  and for any  $\epsilon > 0$ ,  $P[|X - \mu| \geq \epsilon] \leq \sigma^2/\epsilon^2$ . This exercise investigates the conditions under which the upper bound is tight, i.e., the conditions under which  $P[|X - \mu| \geq \epsilon] = \sigma^2/\epsilon^2$ .
  - (a) Let  $\mu = 0$  and  $\sigma = 1$ . Consider the value  $\epsilon = 1$ . Are there any random variables  $X$  with mean 0 and SD 1 such that  $P[|X| \geq 1] = 1$ ? If not, explain why. If so, construct such a distribution.
  - (b) Now consider values of  $\epsilon$  other than 1. Are there any random variables  $X$  with mean 0 and SD 1, and values of  $\epsilon \neq 1$  such that  $P[|X - \mu| \geq \epsilon] = \sigma^2/\epsilon^2$ ?
  - (c) Now let  $\mu$  and  $\sigma$  be arbitrary and let  $\epsilon = \sigma$ . Are there any continuous random variables  $X$  with mean  $\mu$  and SD  $\sigma$  such that  $P[|X - \mu| \geq \epsilon] = \sigma^2/\epsilon^2$ ? If not, explain. If so, show how to construct one.
  - (d) Are there any discrete random variables  $X$  with mean  $\mu$  and SD  $\sigma$  such that  $P[|X - \mu| \geq \epsilon] = \sigma^2/\epsilon^2$ ? If not, explain. If so, show how to construct one.
  - (e) Let  $\mu$  and  $\sigma$  be arbitrary and let  $\epsilon \neq \sigma$ . Are there any random variables  $X$  with mean  $\mu$  and SD  $\sigma$  such that  $P[|X - \mu| \geq \epsilon] = \sigma^2/\epsilon^2$ ? If not, explain. If so, show how to construct one.

## CHAPTER 2

# MODES OF INFERENCE

## 2.1 Data

This chapter takes up the heart of statistics: making inferences, quantitatively, from data. The data,  $y_1, \dots, y_n$  are assumed to be a random sample from a population.

In Chapter 1 we reasoned from  $f$  to  $Y$ . That is, we made statements like “If the experiment is like ..., then  $f$  will be ..., and  $(y_1, \dots, y_n)$  will look like ...” or “ $\mathbb{E}(Y)$  must be ...”, etc. In Chapter 2 we reason from  $Y$  to  $f$ . That is, we make statements such as “Since  $(y_1, \dots, y_n)$  turned out to be ... it seems that  $f$  is likely to be ...”, or “ $\int yf(y) dy$  is likely to be around ...”, etc. This is a basis for knowledge: learning about the world by observing it. Its importance cannot be overstated. The field of statistics illuminates the type of thinking that allows us to learn from data and contains the tools for learning quantitatively.

Reasoning from  $Y$  to  $f$  works because samples are usually like the populations from which they come. For example, if  $f$  has a mean around 6 then most reasonably large samples from  $f$  also have a mean around 6, and if our sample has a mean around 6 then we infer that  $f$  likely has a mean around 6. If our sample has an SD around 10 then we infer that  $f$  likely has an SD around 10, and so on. So much is obvious. But can we be more precise? If our sample has a mean around 6, then can we infer that  $f$  likely has a mean somewhere between, say, 5.5 and 6.5, or can we only infer that  $f$  likely has a mean between 4 and 8, or even worse, between about -100 and 100? When can we say anything quantitative at all about the mean of  $f$ ? The answer is not obvious, and that’s where statistics comes in. Statistics provides the quantitative tools for answering such questions.

This chapter presents several generic modes of statistical analysis.

**Data Description** Data description can be visual, through graphs, charts, etc., or numerical, through calculating sample means, SD’s, etc. Displaying a few simple features

of the data  $y_1, \dots, y_n$  can allow us to visualize those same features of  $f$ . Data description requires few *a priori* assumptions about  $f$ .

**Likelihood** In likelihood inference we assume that  $f$  is a member of a parametric family of distributions  $\{f_\theta : \theta \in \Theta\}$ . Then inference about  $f$  is the same as inference about the parameter  $\theta$ , and different values of  $\theta$  are compared according to how well  $f_\theta$  explains the data.

**Estimation** The goal of estimation is to estimate various aspects of  $f$ , such as its mean, median, SD, etc. Along with the estimate, statisticians try to give quantitative measures of how accurate the estimates are.

**Bayesian Inference** Bayesian inference is a way to account not just for the data  $y_1, \dots, y_n$ , but also for other information we may have about  $f$ .

**Prediction** Sometimes the goal of statistical analysis is not to learn about  $f$  *per se*, but to make predictions about  $y$ 's that we will see in the future. In addition to the usual problem of not knowing  $f$ , we have the additional problem that even if we knew  $f$ , we still wouldn't be able to predict future  $y$ 's exactly.

**Hypothesis Testing** Sometimes we want to test hypotheses like *Head Start is good for kids* or *lower taxes are good for the economy* or *the new treatment is better than the old*.

**Decision Making** Often, decisions have to be made on the basis of what we have learned about  $f$ . In addition, making good decisions requires accounting for the potential gains and losses of each decision.

## 2.2 Data Description

There are many ways, both graphical and numerical, to describe data sets. Sometimes we're interested in means, sometimes variations, sometimes trends through time, and there are good ways to describe and display all these aspects and many more. Simple data description is often enough to shed light on an underlying scientific problem. The subsections of Section 2.2 show some basic ways to describe various types of data.

### 2.2.1 Summary Statistics

One of the simplest ways to describe a data set is by a low dimensional summary. For instance, in Example 1.5 on ocean temperatures there were multiple measurements of

temperatures from each of 9 locations. The measurements from each location were summarized by the sample mean  $\bar{y} = n^{-1} \sum y_i$ ; comparisons of the 9 sample means helped oceanographers deduce the presence of the Mediterranean tongue. Similarly, the essential features of many data sets can be captured in a one-dimensional or low-dimensional summary. Such a summary is called a *statistic*. The examples below refer to a data set  $y_1, \dots, y_n$  of size  $n$ .

**Definition 2.1** (Statistic). A *statistic* is any function, possibly vector valued, of the data.

The most important statistics are measures of location and dispersion. Important examples of location statistics include

**mean** The mean of the data is  $\bar{y} \equiv n^{-1} \sum y_i$ . R can compute means:

```
y <- 1:10
mean(y)
```

**median** A *median* of the data is any number  $m$  such that at least half of the  $y_i$ 's are less than or equal to  $m$  and at least half of the  $y_i$ 's are greater than or equal to  $m$ . We say “a” median instead of “the” median because a data set with an even number of observations has an interval of medians. For example, if  $y <- 1:10$ , then every  $m \in [5, 6]$  is a median. When R computes a median it computes a single number by taking the midpoint of the interval of medians. So `median(y)` yields 5.5.

**quantiles** For any  $p \in [0, 1]$ , the  $p$ -th *quantile* of the data should be, roughly speaking, the number  $q$  such that  $pn$  of the data points are less than  $q$  and  $(1 - p)n$  of the data points are greater than  $q$ .

Figure 2.1 illustrates the idea. Panel a shows a sample of 100 points plotted as a stripchart (page 107). The black circles on the abscissa are the .05, .5, and .9 quantiles; so 5 points (open circles) are to the left of the first vertical line, 50 points are on either side of the middle vertical line, and 10 points are to the right of the third vertical line. Panel b shows the empirical cdf of the sample. The values .05, .5, and .9 are shown as squares on the vertical axis; the quantiles are found by following the horizontal lines from the vertical axis to the cdf, then the vertical lines from the cdf to the horizontal axis. Panels c and d are similar, but show the distribution from which the sample was drawn instead of showing the sample itself. In panel c, 5% of the mass is to the left of the first black circle; 50% is on either side of the middle black circle; and 10% is to the right of the third black dot. In panel d, the open

squares are at .05, .5, and .9 on the vertical axis; the quantiles are the circles on the horizontal axis.

Denote the  $p$ -th quantile as  $q_p(y_1, \dots, y_n)$ , or simply as  $q_p$  if the data set is clear from the context. With only a finite sized sample  $q_p(y_1, \dots, y_n)$  cannot be found exactly. So the algorithm for finding quantiles works as follows.

1. Sort the  $y_i$ 's in ascending order. Label them  $y_{(1)}, \dots, y_{(n)}$  so that

$$y_{(1)} \leq \dots \leq y_{(n)}.$$

2. Set  $q_0 \equiv y_{(1)}$  and  $q_1 \equiv y_{(n)}$ .
3.  $y_{(2)}$  through  $y_{(n-1)}$  determine  $n-1$  subintervals in  $[y_{(1)}, y_{(n)}]$ . So, for  $i = 1, \dots, n-2$ , set  $q_{\frac{i}{n-1}} \equiv y_{(i+1)}$ .
4. For  $p \in (\frac{i}{n-1}, \frac{i+1}{n-1})$  let  $q_p$  be any number in the interval  $(q_{\frac{i}{n-1}}, q_{\frac{i+1}{n-1}})$ .

If  $p$  is a “nice” number then  $q_p$  is often given a special name. For example,  $q_{.5}$  is the median;  $(q_{.25}, q_{.5}, q_{.75})$ , the first, second and third quartiles, is a vector-valued statistic of dimension 3;  $q_{.1}, q_{.2}, \dots$  are the deciles;  $q_{.78}$  is the 78'th percentile.

R can compute quantiles. When faced with  $p \in (\frac{i}{n-1}, \frac{i+1}{n-1})$  R does linear interpolation. E.g. `quantile(y, c(.25, .75))` yields (3.25, 7.75).

The vector  $(y_{(1)}, \dots, y_{(n)})$  defined in step 1 of the algorithm for quantiles is an  $n$ -dimensional statistic called the *order statistic*.  $y_{(i)}$  by itself is called the  $i$ 'th order statistic.

Figure 2.1 was created with the following R code.

```

par (mffrow=c(2,2))
quant <- c (.05, .5, .9)
nquant <- length(quant)

y <- rgamma (100, 3)
stripchart(y, method="jitter", pch=1, xlim=c(0,10),
 xlab="y", main="a")
abline (v=quantile(y,quant))
points (x=quantile(y,quant), y=rep(.5,nquant), pch=19)

plot.ecdf(y, xlab="y", ylab="F(y)", xlim=c(0,10),

```

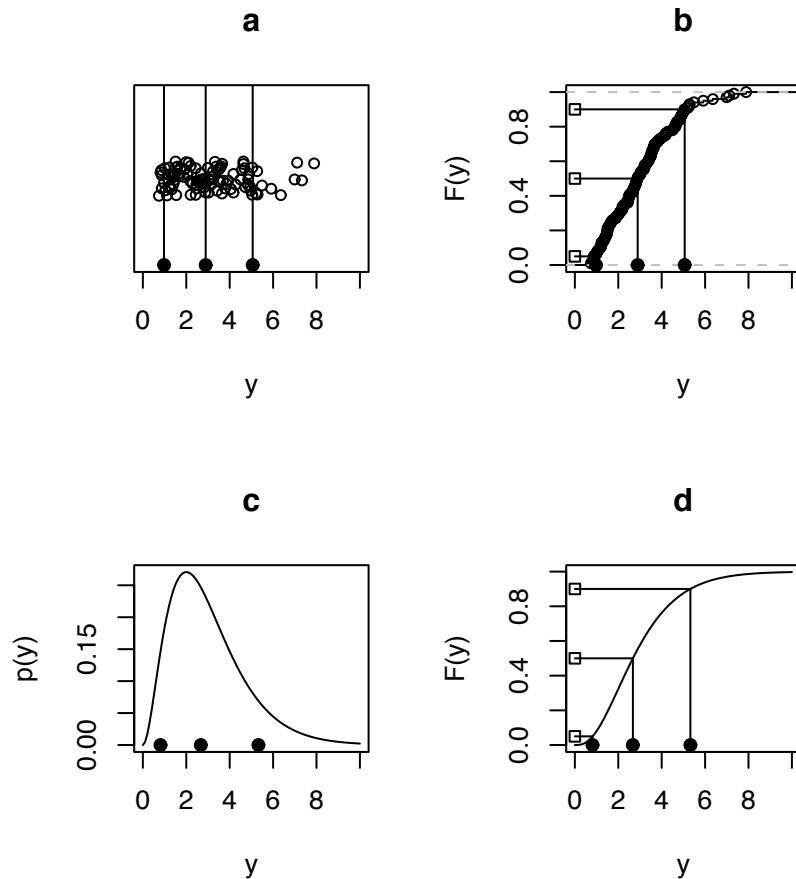


Figure 2.1: Quantiles. The black circles are the .05, .5, and .9 quantiles. The open squares are the numbers .05, .5, and .9 on the vertical axis. Panels a and b are for a sample; panels c and d are for a distribution.

```

main="b")
for (q in quant)
 segments (c(0,quantile(y,q)), c(q,0),
 rep(quantile(y,q),2), rep(q,2))
points (x=quantile(y,quant), y=rep(0,nquant), pch=19)
points (x=rep(0,nquant), y=quant, pch=22)

y <- seq(0,10,length=100)
plot (y, dgamma(y,3), type="l", xlim=c(0,10), ylab="p(y)",
 main="c")
points (x=qgamma(quant,3), y=rep(0,nquant), pch=19)

plot (y, pgamma(y,3), type="l", ylab="F(y)", main="d")
for (q in quant)
 segments (c(0,qgamma(q,3)), c(q,0), rep(qgamma(q,3),2),
 rep(q,2))
points (x=qgamma(quant,3), y=rep(0,nquant), pch=19)
points (x=rep(0,nquant), y=quant, pch=22)

```

- **plot.ecdf** plots the empirical cumulative distribution function. Here the word “empirical” means that the cdf comes from a sample, as opposed to theoretical probability calculations.

Dispersion statistics measure how spread out the data are. Since there are many ways to measure dispersion there are many dispersion statistics. Important dispersion statistics include

**standard deviation** The sample standard deviation or SD of a data set is

$$s \equiv \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}}$$

Note: some statisticians prefer

$$s \equiv \sqrt{\frac{\sum(y_i - \bar{y})^2}{n - 1}}$$

for reasons which do not concern us here. If  $n$  is large there is little difference between the two versions of  $s$ .

**variance** The sample variance is

$$s^2 \equiv \frac{\sum(y_i - \bar{y})^2}{n}$$

Note: some statisticians prefer

$$s^2 \equiv \frac{\sum(y_i - \bar{y})^2}{n - 1}$$

for reasons which do not concern us here. If  $n$  is large there is little difference between the two versions of  $s^2$ .

**interquartile range** The interquartile range is  $q_{.75} - q_{.25}$

Presenting a low dimensional statistic is useful if we believe that the statistic is representative of the whole population. For instance, in Example 1.5, oceanographers believe the data they have collected is representative of the long term state of the ocean. Therefore the sample means at the nine locations in Figure 1.12 are representative of the long term state of the ocean at those locations. More formally, for each location we can imagine a population of temperatures, one temperature for each moment in time. That population has an unknown pdf  $f$ . Even though the data are not really a random sample from  $f$  (The sampling times were not chosen randomly, among other problems.) we can think of them that way without making too serious an error. The histograms in Figure 1.12 are estimates of the  $f$ 's for the nine locations. The mean of each  $f$  is what oceanographers call a *climatological mean*, or an average which, because it is taken over a long period of time, represents the climate. The nine sample means are estimates of the nine climatological mean temperatures at those nine locations. Simply presenting the sample means reveals some interesting structure in the data, and hence an interesting facet of physical oceanography.

Often, more than a simple data description or display is necessary; the statistician has to do a bit of exploring the data set. This activity is called *exploratory data analysis* or simply *eda*. It is hard to give general rules for *eda*, although displaying the data in many different ways is often a good idea. The statistician must decide what displays and *eda* are appropriate for each data set and each question that might be answered by the data set. That is one thing that makes statistics interesting. It cannot be reduced to a set of rules and procedures. A good statistician must be attuned to the potentially unique aspects of each analysis. We now present several examples to show just a few of the possible ways to explore data sets by displaying them graphically. The examples reveal some of the power of graphical display in illuminating data and teasing out what it has to say.

### 2.2.2 Displaying Distributions

Instead of reducing a data set to just a few summary statistics, it is often helpful to display the full data set. But reading a long list of numbers is usually not helpful; humans are not good at assimilating data in that form. We can learn a lot more from a graphical representation of the data.

**Histograms** The next examples use histograms to display the full distribution of some data sets. Visual comparison of the histograms reveals structure in the data.

**Example 2.1** (Tooth Growth)

The R statistical language comes with many data sets. Type `data()` to see what they are. This example uses the data set `ToothGrowth` on the effect of vitamin C on tooth growth in guinea pigs. You can get a description by typing `help(ToothGrowth)`. You can load the data set into your R session by typing `data(ToothGrowth)`. `ToothGrowth` is a `dataframe` of three columns. The first few rows look like this:

|   | len  | supp | dose |
|---|------|------|------|
| 1 | 4.2  | VC   | 0.5  |
| 2 | 11.5 | VC   | 0.5  |
| 3 | 7.3  | VC   | 0.5  |

Column 1, or `len`, records the amount of tooth growth. Column 2, `supp`, records whether the guinea pig was given vitamin C in ascorbic acid or orange juice. Column 3, `dose`, records the dose, either 0.5, 1.0 or 2.0 mg. Thus there are six groups of guinea pigs in a two by three layout. Each group has ten guinea pigs, for a total of sixty observations. Figure 2.2 shows histograms of growth for each of the six groups. From Figure 2.2 it is clear that dose affects tooth growth.

Figure 2.2 was produced by the following R code.

```

supp <- unique (ToothGrowth$supp)
dose <- unique (ToothGrowth$dose)
par (mfcoll=c(3,2))
for (i in 1:2)
 for (j in 1:3) {
 good <- (ToothGrowth$supp == supp[i]
 & ToothGrowth$dose == dose[j])
 hist (ToothGrowth$len[good], breaks=seq(0,34,by=2),

```

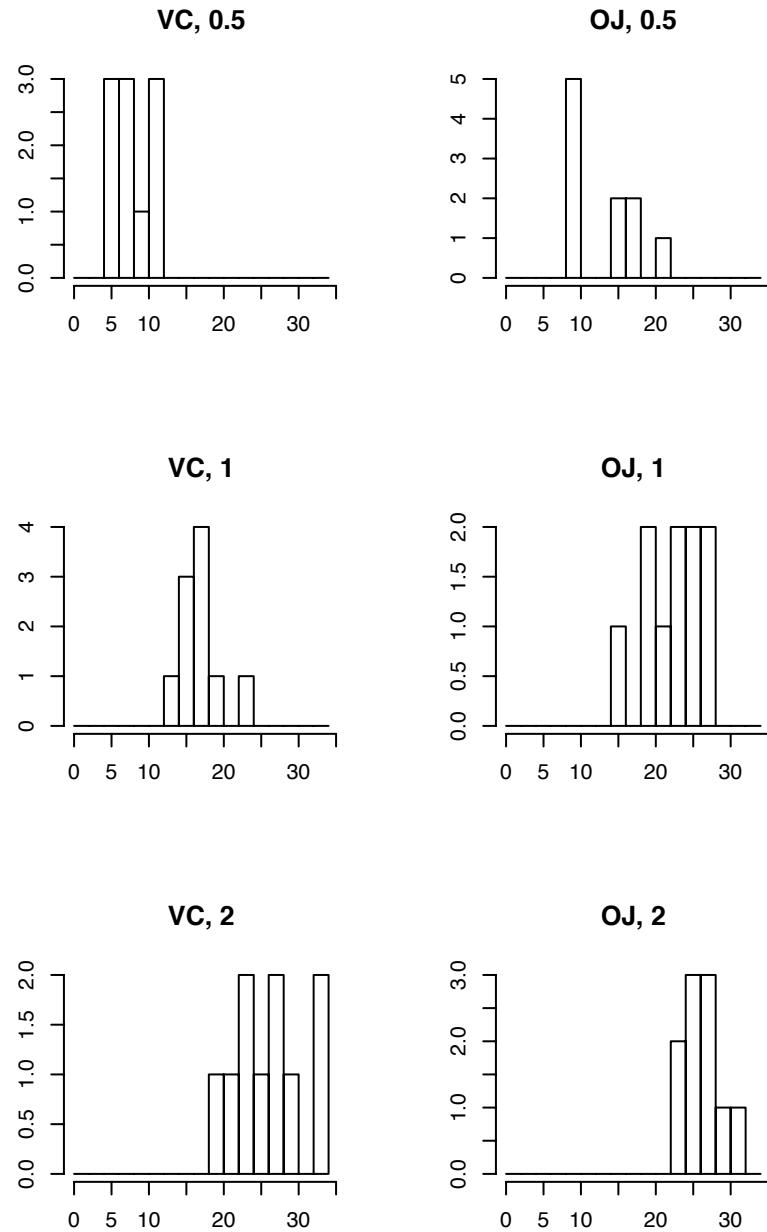


Figure 2.2: Histograms of tooth growth by delivery method (VC or OJ) and dose (0.5, 1.0 or 2.0).

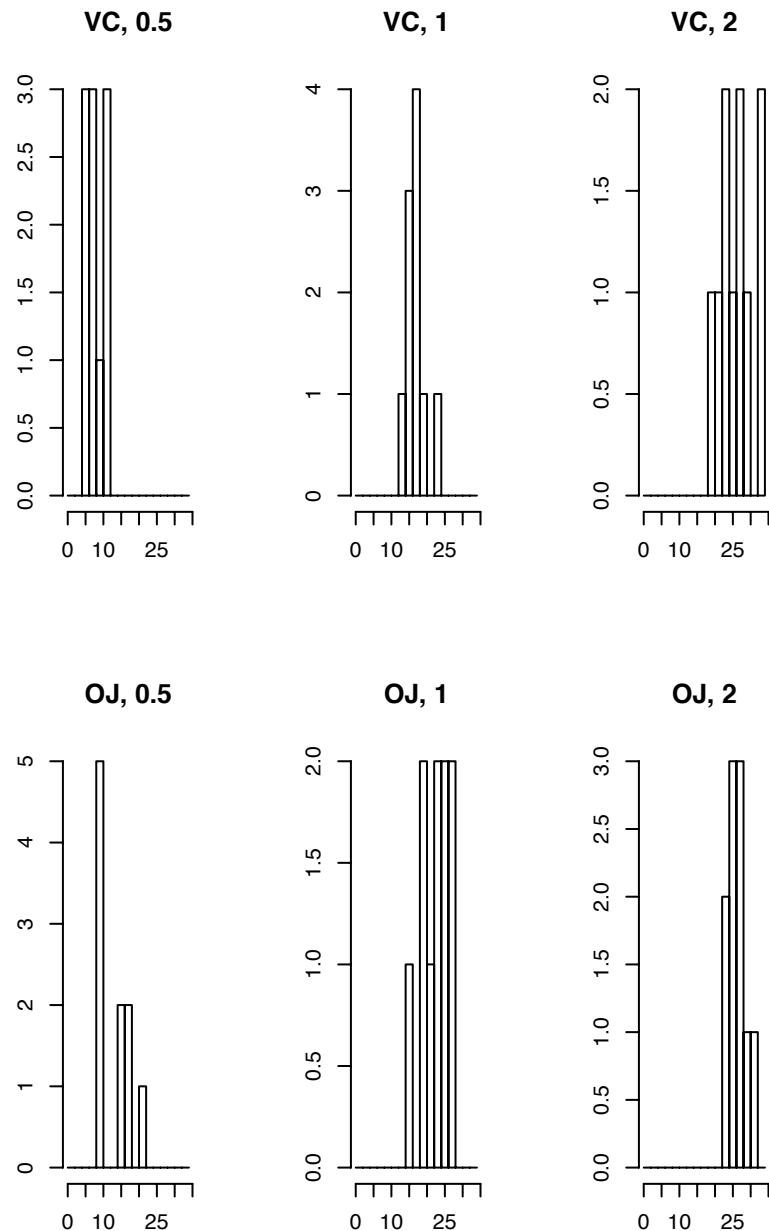


Figure 2.3: Histograms of tooth growth by delivery method (VC or OJ) and dose (0.5, 1.0 or 2.0).

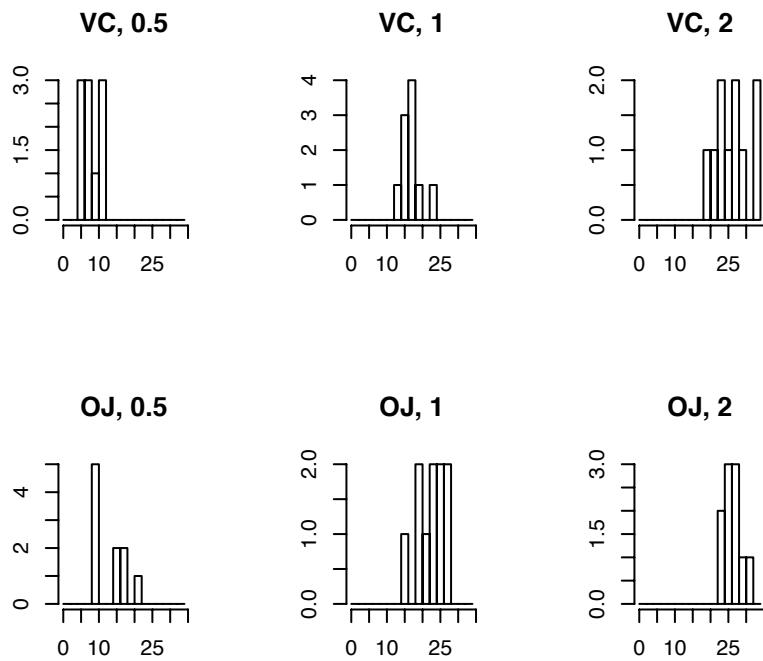


Figure 2.4: Histograms of tooth growth by delivery method (VC or OJ) and dose (0.5, 1.0 or 2.0).

```

 xlab="", ylab="",
 main=paste(supp[i], " ", " ", dose[j], sep=""))
}
```

- `unique(x)` returns the unique values in `x`. For example, if `x <- c(1, 1, 2)` then `unique(x)` would be `1 2`.

Figure 2.3 is similar to Figure 2.2 but laid out in the other direction. (Notice that it's easier to compare histograms when they are arranged vertically rather than horizontally.) The figures suggest that delivery method does have an effect, but not as strong as the dose effect. Notice also that Figure 2.3 is more difficult to read than Figure 2.2 because the histograms are too tall and narrow. Figure 2.4 repeats Figure 2.3 but using less vertical distance; it is therefore easier to read. Part of good statistical practice is displaying figures in a way that makes them easiest to read and interpret.

The figures alone have suggested that dose is the most important effect, and delivery method less so. A further analysis could try to be more quantitative: what is the typical size of each effect, how sure can we be of the typical size, and how much does the effect vary from animal to animal. The figures already suggest answers, but a more formal analysis is deferred to Section 2.7.

Figures 1.12, 2.2, and 2.3 are *histograms*. The abscissa has the same scale as the data. The data are divided into bins. The ordinate shows the number of data points in each bin. (`hist(..., prob=T)` plots the ordinate as probability rather than counts.) Histograms are a powerful way to display data because they give a strong visual impression of the main features of a data set. However, details of the histogram can depend on both the number of bins and on the cut points between bins. For that reason it is sometimes better to use a display that does not depend on those features, or at least not so strongly. Example 2.2 illustrates.

## Density Estimation

### Example 2.2 (Hot Dogs)

In June of 1986, Consumer Reports published a study of hot dogs. The data are available at DASL, the *Data and Story Library*, a collection of data sets for free use by statistics students. DASL says the data are

"Results of a laboratory analysis of calories and sodium content of major hot dog brands. Researchers for Consumer Reports analyzed three types

of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat)."

You can download the data from [HTTP://LIB.STAT.CMU.EDU/DASL/DATAFILES/HOTDOGS.HTML](http://lib.stat.cmu.edu/DASL/Datafiles/HOTDOGS.html). The first few lines look like this:

| Type | Calories | Sodium |
|------|----------|--------|
| Beef | 186      | 495    |
| Beef | 181      | 477    |

This example looks at the calorie content of beef hot dogs. (Later examples will compare the calorie contents of different types of hot dogs.)

Figure 2.5(a) is a histogram of the calorie contents of beef hot dogs in the study. From the histogram one might form the impression that there are two major varieties of beef hot dogs, one with about 130–160 calories or so, another with about 180 calories or so, and a rare outlier with fewer calories. Figure 2.5(b) is another histogram of the same data but with a different bin width. It gives a different impression, that calorie content is evenly distributed, approximately, from about 130 to about 190 with a small number of lower calorie hot dogs. Figure 2.5(c) gives much the same impression as 2.5(b). It was made with the same bin width as 2.5(a), but with cut points starting at 105 instead of 110. These histograms illustrate that one's impression can be influenced by both bin width and cut points.

*Density estimation* is a method of reducing dependence on cut points. Let  $x_1, \dots, x_{20}$  be the calorie contents of beef hot dogs in the study. We think of  $x_1, \dots, x_{20}$  as a random sample from a density  $f$  representing the population of all beef hot dogs. Our goal is to estimate  $f$ . For any fixed number  $x$ , how shall we estimate  $f(x)$ ? The idea is to use information local to  $x$  to estimate  $f(x)$ . We first describe a basic version, then add two refinements to get *kernel density estimation* and the `density()` function in R.

Let  $n$  be the sample size (20 for the hot dog data). Begin by choosing a number  $h > 0$ . For any number  $x$  the estimate  $\hat{f}_{\text{basic}}(x)$  is defined to be

$$\hat{f}_{\text{basic}}(x) \equiv \frac{1}{2nh} \sum_{i=1}^n \mathbf{1}_{(x-h, x+h)}(x_i) = \frac{\text{fraction of sample points within } h \text{ of } x}{2h}$$

$\hat{f}_{\text{basic}}$  has at least two apparently undesirable features.

1.  $\hat{f}_{\text{basic}}(x)$  gives equal weight to all data points in the interval  $(x - h, x + h)$  and has abrupt cutoffs at the ends of the interval. It would be better to give the most weight to data points closest to  $x$  and have the weights decrease gradually for points increasingly further away from  $x$ .

2.  $\hat{f}_{\text{basic}}(x)$  depends critically on the choice of  $h$ .

We deal with these problems by introducing a weight function that depends on distance from  $x$ . Let  $g_0$  be a probability density function. Usually  $g_0$  is chosen to be symmetric and unimodal, centered at 0. Define

$$\hat{f}(x) \equiv \frac{1}{n} \sum g_0(x - x_i)$$

Choosing  $g_0$  to be a probability density ensures that  $\hat{f}$  is also a probability density because

$$\int_{-\infty}^{\infty} \hat{f}(x) dx = \frac{1}{n} \sum_i \int_{-\infty}^{\infty} g_0(x - x_i) dx = 1 \quad (2.1)$$

When  $g_0$  is chosen to be a continuous function it deals nicely with problem 1 above. In fact,  $\hat{f}_{\text{basic}}$  comes from taking  $g_0$  to be the uniform density on  $(-h, h)$ .

To deal with problem 2 we rescale  $g_0$ . Choose a number  $h > 0$  and define a new density  $g(x) = h^{-1}g_0(x/h)$ . A little thought shows that  $g$  differs from  $g_0$  by a rescaling of the horizontal axis; the factor  $h^{-1}$  compensates to make  $\int g = 1$ . Now define the density estimate to be

$$\hat{f}_h(x) \equiv \frac{1}{n} \sum g(x - x_i) = \frac{1}{nh} \sum g_0((x - x_i)/h)$$

$h$  is called the *bandwidth*. Of course  $\hat{f}_h$  still depends on  $h$ . It turns out that dependence on bandwidth is not really a problem. It is useful to view density estimates for several different bandwidths. Each reveals features of  $f$  at different scales. Figures 2.5(d), (e), and (f) are examples. Panel (d) was produced by the default bandwidth; panels (e) and (f) were produced with 1/4 and 1/2 the default bandwidth. Larger bandwidth makes a smoother estimate of  $f$ ; smaller bandwidth makes it rougher. None is exactly right. It is useful to look at several.

Figure 2.5 was produced with

```
hotdogs <- read.table ("data/hotdogs/data", header=T)
cal.beef <- hotdogs$Calories [hotdogs>Type == "Beef"]
par (mfrow=c(3,2))
hist (cal.beef, main="(a)", xlab="calories", ylab="")
hist (cal.beef, breaks=seq(110,190,by=20), main="(b)",
 xlab="calories", ylab="")
hist (cal.beef, breaks=seq(105,195,by=10), main="(c)" ,
```

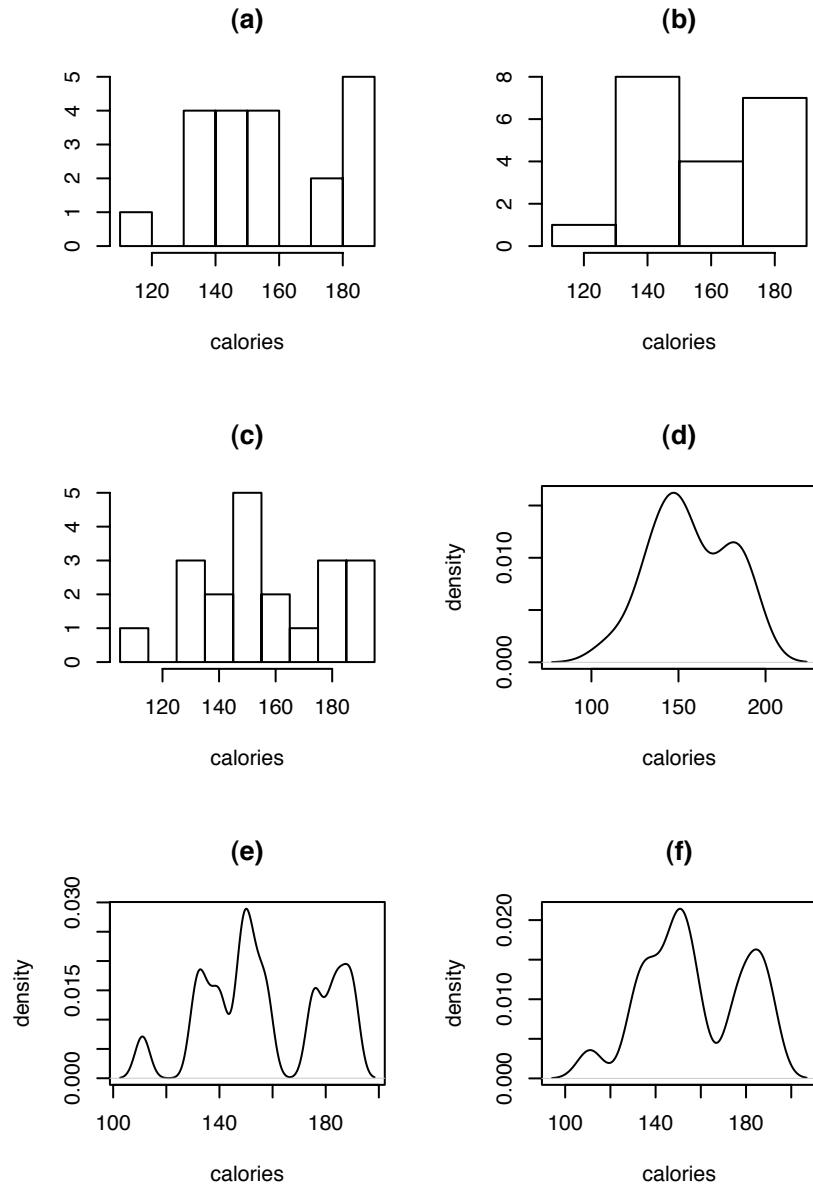


Figure 2.5: **(a), (b), (c)**: histograms of calorie contents of beef hot dogs; **(d), (e), (f)**: density estimates of calorie contents of beef hot dogs.

```

xlab="calories", ylab="")

plot (density (cal.beef), main="(d)", xlab="calories",
 ylab="density")
plot (density (cal.beef, adjust=1/4), main="(e)",
 xlab="calories", ylab="density")
plot (density (cal.beef, adjust=1/2), main="(f)",
 xlab="calories", ylab="density")

```

- In panel **(a)** R used its default method for choosing histogram bins.
- In panels **(b)** and **(c)** the histogram bins were set by  
`hist ( ... , breaks=seq(...))`.
- `density()` produces a kernel density estimate.
- R uses a Gaussian kernel by default which means that  $g_0$  above is the  $N(0, 1)$  density.
- In panel **(d)** R used its default method for choosing bandwidth.
- In panels **(e)** and **(f)** the bandwidth was set to 1/4 and 1/2 the default by  
`density(..., adjust=...)`.

**Stripcharts and Dotplots** Figure 2.6 uses the ToothGrowth data to illustrate *stripcharts*, also called *dotplots*, an alternative to histograms. Panel **(a)** has three rows of points corresponding to the three doses of ascorbic acid. Each point is for one animal. The abscissa shows the amount of tooth growth; the ordinate shows the dose. The panel is slightly misleading because points with identical coordinates are plotted directly on top of each other. In such situations statisticians often add a small amount of jitter to the data, to avoid overplotting. The middle panel is a repeat of the top, but with jitter added. The bottom panel shows tooth growth by delivery method. Compare Figure 2.6 to Figures 2.2 and 2.3. Which is a better display for this particular data set?

Figure 2.6 was produced with the following R code.

```
par (mfrow=c(3,1))
```

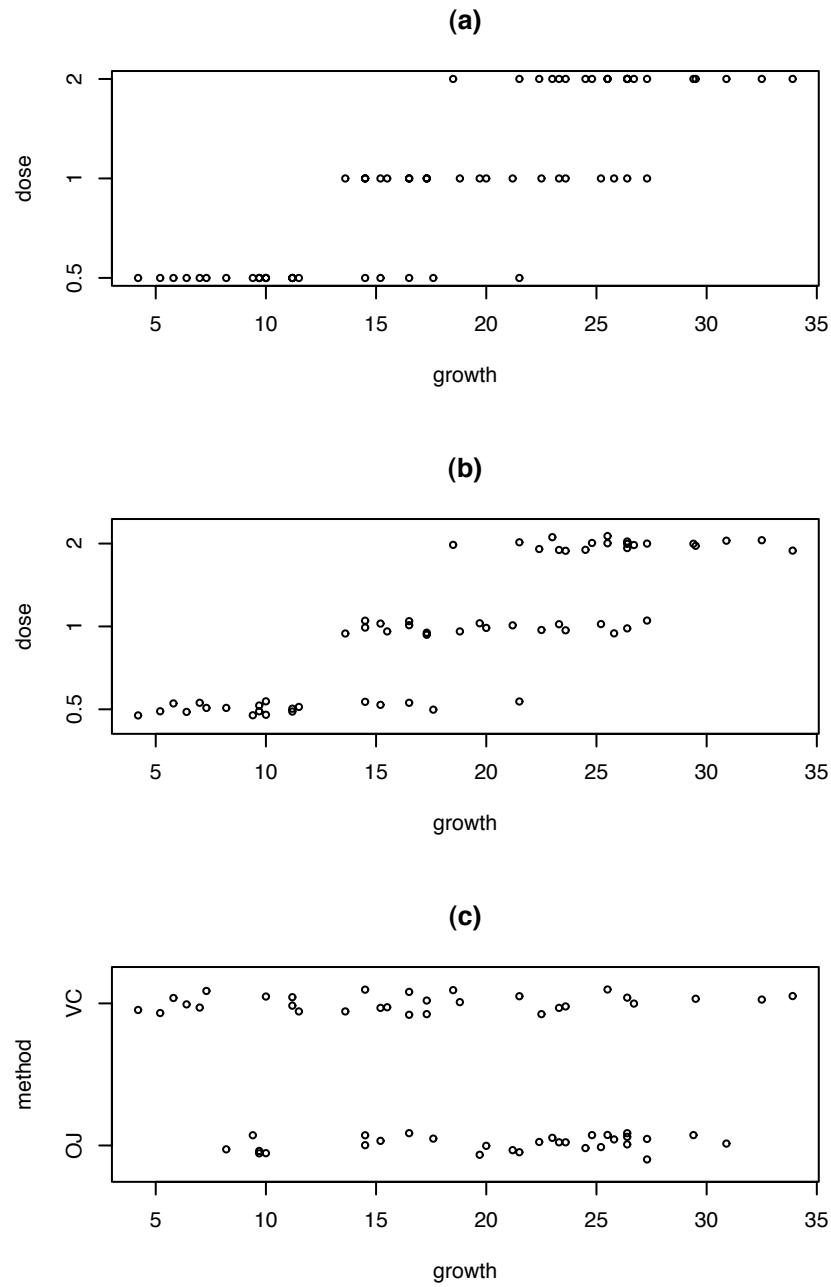


Figure 2.6: (a) Tooth growth by dose, no jittering; (b) Tooth growth by dose with jittering; (c) Tooth growth by delivery method with jittering

```
stripchart (ToothGrowth$len ~ ToothGrowth$dose, pch=1,
 main="(a)", xlab="growth", ylab="dose")
stripchart (ToothGrowth$len ~ ToothGrowth$dose,
 method="jitter", main="(b)", xlab="growth",
 ylab="dose", pch=1)
stripchart (ToothGrowth$len ~ ToothGrowth$supp,
 method="jitter", main="(c)", xlab="growth",
 ylab="method", pch=1)
```

**Boxplots** An alternative, useful for comparing many distributions simultaneously, is the *boxplot*. Example 2.3 uses boxplots to compare scores on 24 quizzes in a statistics course.

### Example 2.3 (Quiz Scores)

In the spring semester of 2003, 58 students completed Statistics 103 at Duke University. Figure 2.7 displays their grades.

There were 24 quizzes during the semester. Each was worth 10 points. The upper panel of the figure shows the distribution of scores on each quiz. The abscissa is labelled 1 through 24, indicating the quiz number. For each quiz, the figure shows a *boxplot*. For each quiz there is a box. The horizontal line through the center of the box is the median grade for that quiz. We can see that the median score on Quiz 2 is around 7, while the median score on Quiz 3 is around 4. The upper end of the box is the 75th percentile (3rd quartile) of scores; the lower end of the box is the 25th percentile (1st quartile). We can see that about half the students scored between about 5 and 8 on Quiz 2, while about half the students scored between about 2 and 6 on Quiz 3. Quiz 3 was tough.

Each box may have *whiskers*, or dashed lines which extend above and below the box. The exact definition of the whiskers is not important, but they are meant to include most of the data points that don't fall inside the box. (In R, by default, the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range away from the median.) Finally, there may be some individual points plotted above or below each boxplot. These indicate *outliers*, or scores that are extremely high or low relative to other scores on that quiz. Many quizzes had low outliers; only Quiz 5 had a high outlier.

Box plots are extremely useful for comparing many sets of data. We can easily see, for example, that Quiz 5 was the most difficult (75% of the class scored 3 or less.) while Quiz 1 was the easiest (over 75% of the class scored 10.)

There were no exams or graded homeworks. Students' grades were determined by their best 20 quizzes. To compute grades, each student's scores were sorted, the first 4 were dropped, then the others were averaged. Those averages are displayed in a stripchart in the bottom panel of the figure. It's easy to see that most of the class had quiz averages between about 5 and 9 but that 4 averages were much lower.

Figure 2.7 was produced by the following R code.

```
... # read in the data
colnames(scores) <- paste("Q", 1:24, sep="")
define column names
boxplot (data.frame(scores), main="Individual quizzes")

scores[is.na(scores)] <- 0 # replace missing scores
 # with 0's
temp <- apply (scores, 1, sort) # sort
temp <- temp[5:24,] # drop the 4 lowest
scores.ave <- apply (temp, 2, mean) # find the average

stripchart (scores.ave, "jitter", pch=1, xlab="score",
 xlim=c(0,10), main="Student averages")
```

**QQ plots** Sometimes we want to assess whether a data set is well modelled by a Normal distribution and, if not, how it differs from Normal. One obvious way to assess Normality is by looking at histograms or density estimates. But the answer is often not obvious from the figure. A better way to assess Normality is with *QQ plots*. Figure 2.8 illustrates for the nine histograms of ocean temperatures in Figure 1.12.

Each panel in Figure 2.8 was created with the ocean temperatures near a particular (latitude, longitude) combination. Consider, for example, the upper left panel which was constructed from the  $n = 213$  points  $x_1, \dots, x_{213}$  taken near (45, -40). Those points are sorted, from smallest to largest, to create the order statistic  $(x_{(1)}, \dots, x_{(213)})$ . Then they are plotted against  $\mathbb{E}[(z_{(1)}, \dots, z_{(213)})]$ , the expected order statistic from a Normal distribution. If the  $x_i$ s are approximately Normal then the QQ plot will look approximately linear. The slope of the line indicates the standard deviation.

In Figure 2.8 most of the panels do look approximately linear, indicating that a Normal model is reasonable. But some of the panels show departures from Normality. In the upper left and lower left panels, for example, the plots looks roughly linear except for the upper

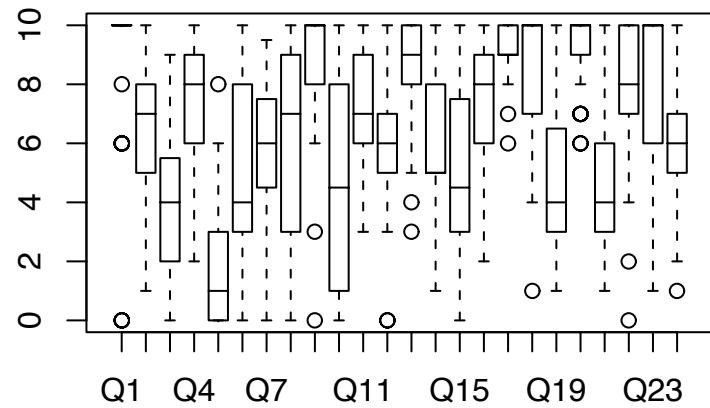
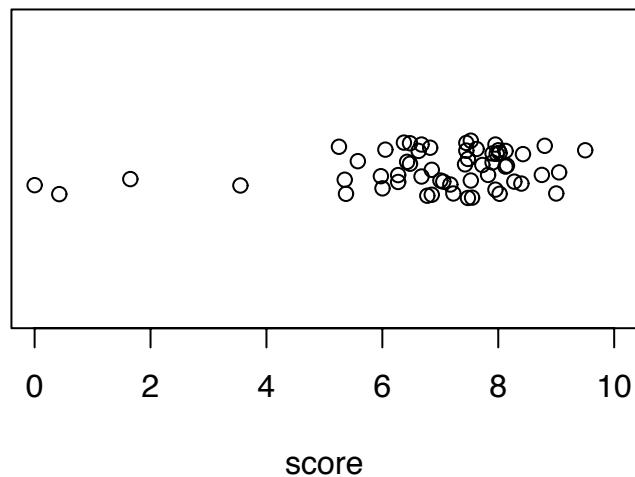
**Individual quizzes****Student averages**

Figure 2.7: Quiz scores from Statistics 103

right corners which show some data points much warmer than expected if they followed a Normal distribution. In contrast, the coolest temperatures in the lower middle panel are not quite as cool as expected from a Normal distribution.

Figure 2.8 was produced with

```

lats <- c (45, 35, 25)
lons <- c (-40, -30, -20)
par (mfrow=c(3,3))
for (i in 1:3)
for (j in 1:3) {
 good <- abs (med.1000$lon - lons[j]) < 1 &
 abs (med.1000$lat - lats[i]) < 1
 qqnorm (med.1000$temp[good], xlab="", ylab="",
 sub=paste("n = ", sum(good), sep=""),
 main = paste ("latitude =", lats[i], "\n longitude =",
 lons[j]))
}

```

### 2.2.3 Exploring Relationships

Sometimes it is the relationships between several random variables that are of interest. For example, in discrimination cases the focus is on the relationship between race or gender on one hand and employment or salary on the other hand. Subsection 2.2.3 shows several graphical ways to display relationships.

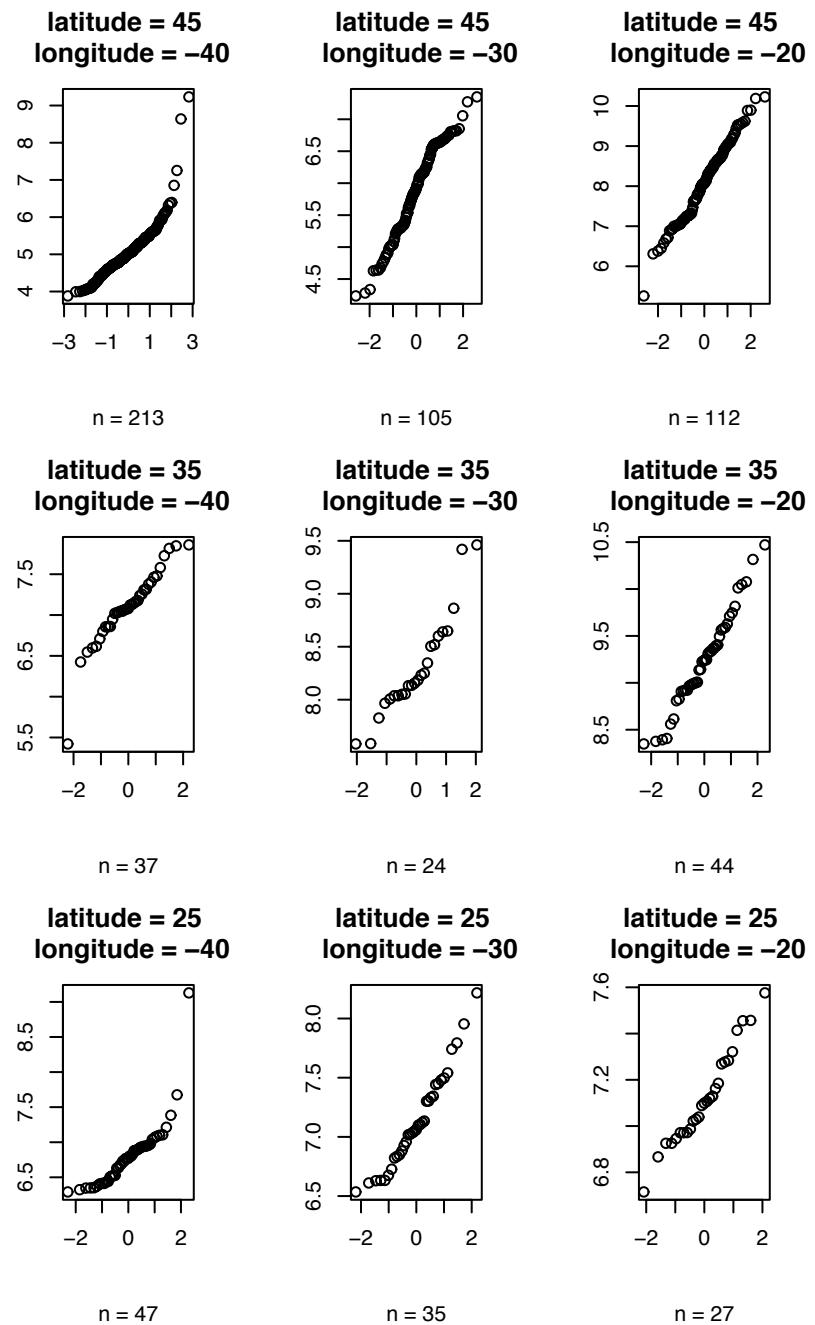
We begin with Example 2.4, an analysis of potential discrimination in admission to UC Berkeley graduate school.

#### **Example 2.4**

In 1973 UC Berkeley investigated its graduate admissions rates for potential sex bias. Apparently women were more likely to be rejected than men. The data set UCBAdmissions gives the acceptance and rejection data from the six largest graduate departments on which the study was based. Typing `help(UCBAdmissions)` tells more about the data. It tells us, among other things:

...

Format:

Figure 2.8: QQ plots of water temperatures ( $^{\circ}\text{C}$ ) at 1000m depth

A 3-dimensional array resulting from cross-tabulating 4526 observations on 3 variables. The variables and their levels are as follows:

| No | Name   | Levels             |
|----|--------|--------------------|
| 1  | Admit  | Admitted, Rejected |
| 2  | Gender | Male, Female       |
| 3  | Dept   | A, B, C, D, E, F   |

...

The major question at issue is whether there is sex bias in admissions. To investigate we ask whether men and women are admitted at roughly equal rates.

Typing UCBAdmissions gives the following numerical summary of the data.

, , Dept = A

| Gender   |      |        |
|----------|------|--------|
| Admit    | Male | Female |
| Admitted | 512  | 89     |
| Rejected | 313  | 19     |

, , Dept = B

| Gender   |      |        |
|----------|------|--------|
| Admit    | Male | Female |
| Admitted | 353  | 17     |
| Rejected | 207  | 8      |

, , Dept = C

| Gender   |      |        |
|----------|------|--------|
| Admit    | Male | Female |
| Admitted | 120  | 202    |
| Rejected | 205  | 391    |

, , Dept = D

| Gender |      |        |
|--------|------|--------|
| Admit  | Male | Female |

```
Admitted 138 131
Rejected 279 244
```

```
, , Dept = E
```

|          |      | Gender |  |
|----------|------|--------|--|
| Admit    | Male | Female |  |
| Admitted | 53   | 94     |  |
| Rejected | 138  | 299    |  |

```
, , Dept = F
```

|          |      | Gender |  |
|----------|------|--------|--|
| Admit    | Male | Female |  |
| Admitted | 22   | 24     |  |
| Rejected | 351  | 317    |  |

For each department, the two-way table of admission status versus sex is displayed. Such a display, called a *crosstabulation*, simply tabulates the number of entries in each cell of a multiway table. It's hard to tell from the crosstabulation whether there is a sex bias and, if so, whether it is systemic or confined to just a few departments. Let's continue by finding the marginal (aggregated by department as opposed to conditional given department) admissions rates for men and women.

```
> apply(UCBAdmissions, c(1, 2), sum)
 Gender
Admit Male Female
 Admitted 1198 557
 Rejected 1493 1278
```

The admission rate for men is  $1198/(1198 + 1493) = 44.5\%$  while the admission rate for women is  $557/(557 + 1493) = 30.4\%$ , much lower. A *mosaic plot*, created with

```
mosaicplot(apply(UCBAdmissions, c(1, 2), sum),
 main = "Student admissions at UC Berkeley")
```

is a graphical way to display the discrepancy. (A beautiful example of a mosaic plot is on the cover of *CHANCE* magazine in Spring, 2007.) The left column is for admitted students; the heights of the rectangles show how many admitted students were male and how many were female. The right column is for rejected students; the heights of

the rectangles show how many were male and female. If sex and admission status were independent, i.e., if there were no sex bias, then the proportion of men among admitted students would equal the proportion of men among rejected students and the heights of the left rectangles would equal the heights of the right rectangles. The apparent difference in heights is a visual representation of the discrepancy in sex ratios among admitted and rejected students. The same data can be viewed as discrepant admission rates for men and women by transposing the matrix:

```
mosaicplot(t(apply(UCBAdmissions, c(1, 2), sum)),
 main = "Student admissions at UC Berkeley")
```

The existence of discrepant sex ratios for admitted and rejected students is equivalent to the existence of discrepant admission rates for males and females and to dependence of sex and admission rates. The lack of discrepant ratios is equivalent to independence of sex and admission rates.

Evidently UC Berkeley admitted men and women at different rates. But graduate admission decisions are not made by a central admissions office; they are made by the individual departments to which students apply. So our next step is to look at admission rates for each department separately. We can look at the crosstabulation on page 114 or make mosaic plots for each department separately (not shown here) with

```
Mosaic plots for individual departments
for(i in 1:6)
 mosaicplot(UCBAdmissions[,,i],
 xlab = "Admit", ylab = "Sex",
 main = paste("Department", LETTERS[i]))
```

The plots show that in each department men and women are admitted at roughly equal rates. The following snippet calculates and prints the rates. It confirms the rough equality except for department A which admitted women at a higher rate than men.

```
for (i in 1:6) { # for each department
 temp <- UCBAdmissions[,,i] # that department's data
 m <- temp[1,1] / (temp[1,1]+temp[2,1]) # Men's admission rate
 w <- temp[1,2] / (temp[1,2]+temp[2,2]) # Women's admission rate
 print (c (m, w)) # print them
}
```

Note that departments A and B which had high admission rates also had large numbers of male applicants while departments C, D, E and F which had low admission

## Student admissions at UC Berkeley

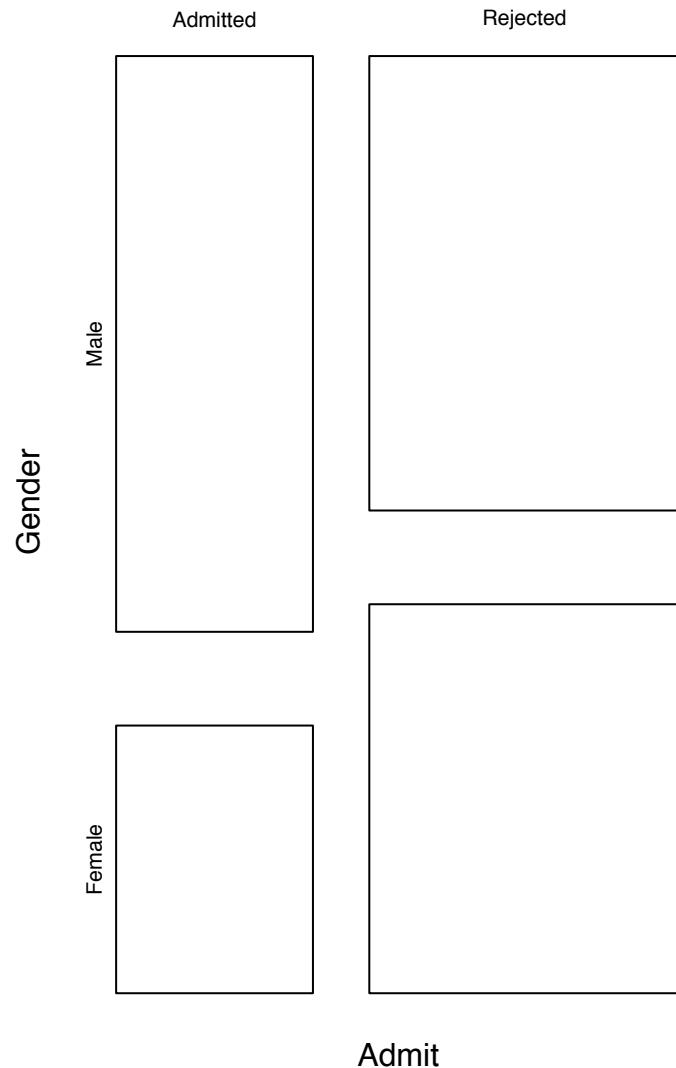


Figure 2.9: Mosaic plot of UCBAdmissions

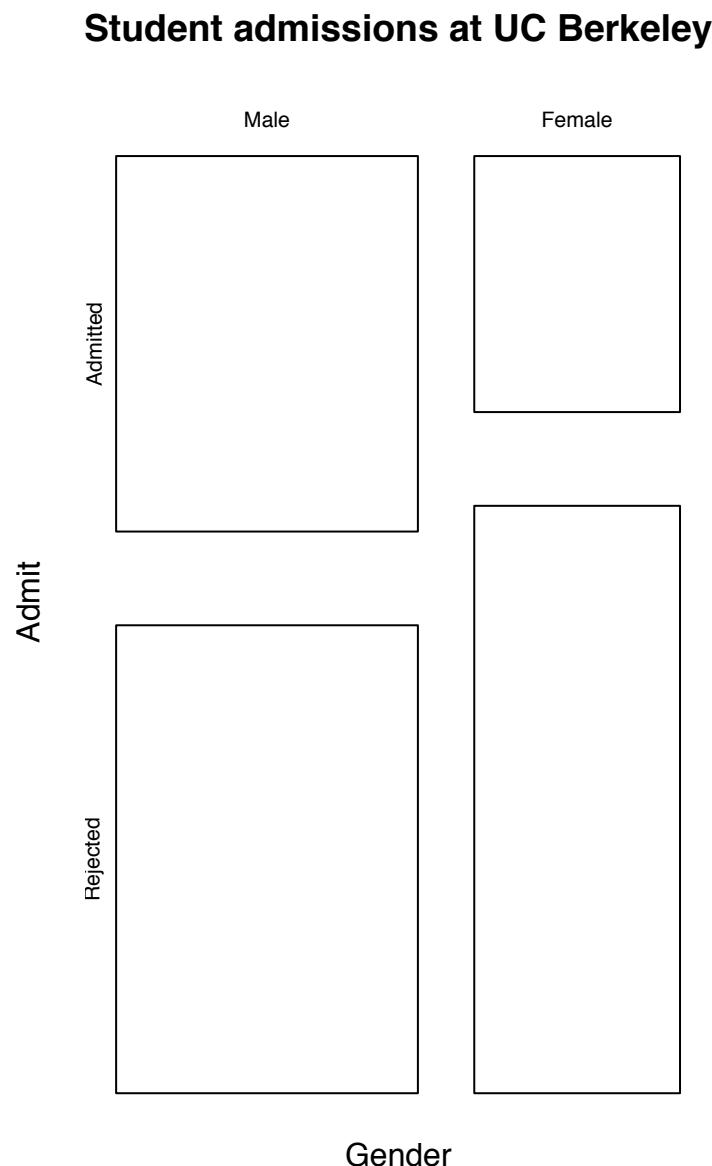


Figure 2.10: Mosaic plot of UCBAdmissions

rates had large numbers of female applicants. The generally accepted explanation for the discrepant marginal admission rates is that men tended to apply to departments that were easy to get into while women tended to apply to departments that were harder to get into. A more sinister explanation is that the university gave more resources to departments with many male applicants, allowing them to admit a greater proportion of their applicants. The data we've analyzed are consistent with both explanations; the choice between them must be made on other grounds.

One lesson here for statisticians is the power of simple data displays and summaries. Another is the need to consider the unique aspects of each data set. The explanation of different admissions rates for men and women could only be discovered by someone familiar with how universities and graduate schools work, not by following some general rules about how to do statistical analyses.

The next example is about the duration of eruptions and interval to the next eruption of the Old Faithful geyser. It explores two kinds of relationships — the relationship between duration and eruption and also the relationship of each variable with time.

### Example 2.5 (Old Faithful)

*Old Faithful* is a geyser in Yellowstone National Park and a great tourist attraction. As DENBY AND PREGIBON [1987] explain, “From August 1 to August 8, 1978, rangers and naturalists at Yellowstone National Park recorded the *duration* of eruption and *interval* to the next eruption (both in minutes) for eruptions of Old Faithful between 6 a.m. and midnight. The intent of the study was to predict the time of the next eruption, to be posted at the Visitor’s Center so that visitors to Yellowstone can usefully budget their time.” The R dataset `faithful` contains the data. In addition to the references listed there, the data and analyses can also be found in WEISBERG [1985] and DENBY AND PREGIBON [1987]. The latter analysis emphasizes graphics, and we shall follow some of their suggestions here.

We begin exploring the data with stripcharts and density estimates of durations and intervals. These are shown in Figure 2.11. The figure suggests bimodal distributions. For *duration* there seems to be one bunch of data around two minutes and another around four or five minutes. For *interval*, the modes are around 50 minutes and 80 minutes. A plot of *interval* versus *duration*, Figure 2.12, suggests that the bimodality is present in the joint distribution of the two variables. Because the data were collected over time, it might be useful to plot the data in the order of collection. That’s Figure 2.13. The horizontal scale in Figure 2.13 is so compressed that it’s hard to see what’s going on. Figure 2.14 repeats Figure 2.13 but divides the time interval into two subintervals to make the plots easier to read. The subintervals overlap slightly. The persistent up-and-down character of Figure 2.14 shows that, for the most part, long

and short durations are interwoven, as are long and short intervals. (Figure 2.14 is potentially misleading. The data were collected over an eight day period. There are eight separate sequences of eruptions with gaps in between. The faithful data set does not tell us where the gaps are. DENBY AND PREGIBON [1987] tell us where the gaps are and use the eight separate days to find errors in data transcription.) Just this simple analysis, a collection of four figures, has given us insight into the data that will be very useful in predicting the time of the next eruption.

Figures 2.11, 2.12, 2.13, and 2.14 were produced with the following R code.

```
data(faithful)
attach(faithful)

par (mfcoll=c(2,2))
stripchart (eruptions, method="jitter", pch=1, xlim=c(1,6),
 xlab="duration (min)", main="(a)")
plot (density (eruptions), type="l", xlim=c(1,6),
 xlab="duration (min)", main="(b)")
stripchart (waiting, method="jitter", pch=1, xlim=c(40,100),
 xlab="waiting (min)", main="(c)")
plot (density (waiting), type="l", xlim=c(40,100),
 xlab="waiting (min)", main="(d)")

par (mfrow=c(1,1))
plot (eruptions, waiting, xlab="duration of eruption",
 ylab="time to next eruption")

par (mfrow=c(2,1))
plot.ts (eruptions, xlab="data number", ylab="duration",
 main="a")
plot.ts (waiting, xlab="data number", ylab="waiting time",
 main="b")

par (mfrow=c(4,1))
plot.ts (eruptions[1:150], xlab="data number",
 ylab="duration", main="a1")
plot.ts (eruptions[130:272], xlab="data number",
 ylab="duration", main="a2")
plot.ts (waiting[1:150], xlab="data number",
```

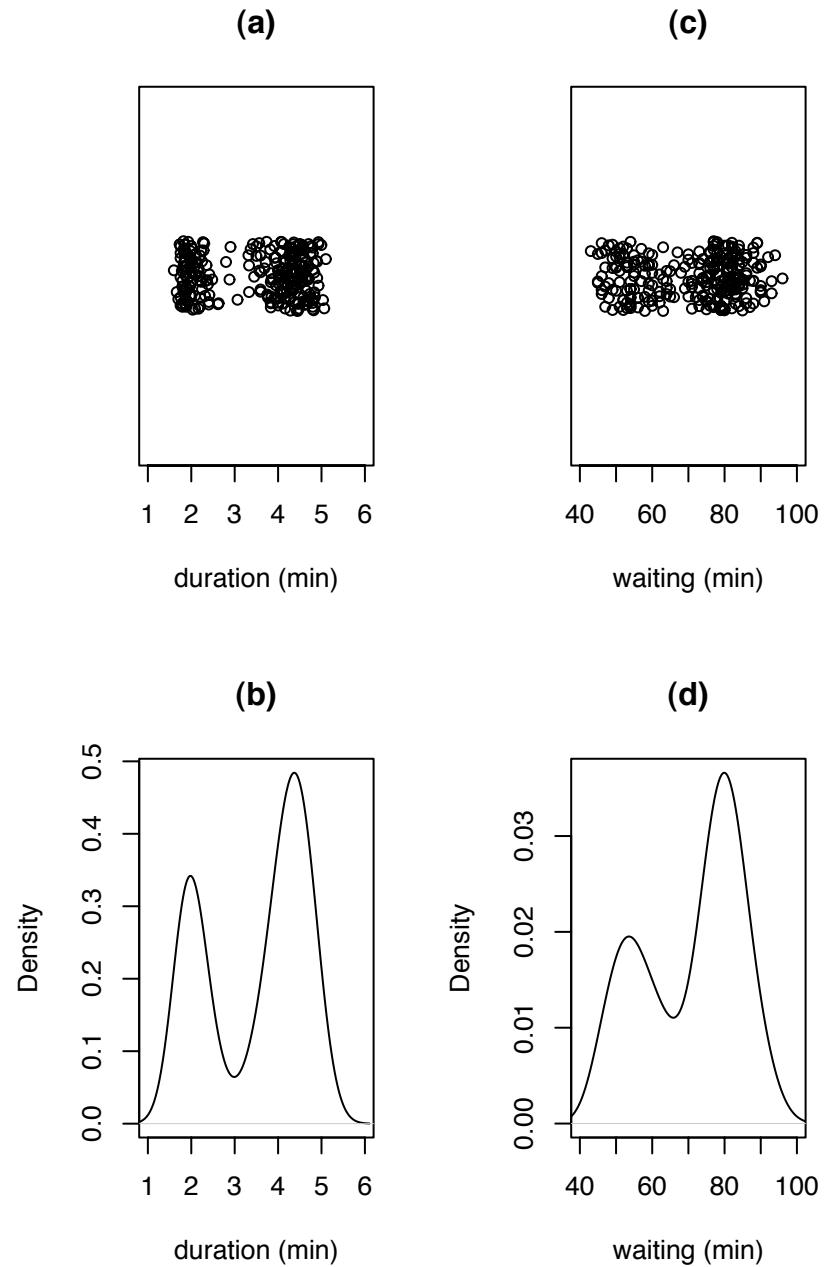


Figure 2.11: Old Faithful data: duration of eruptions and waiting time between eruptions. Stripcharts: (a) and (c). Density estimates: (b) and (d).

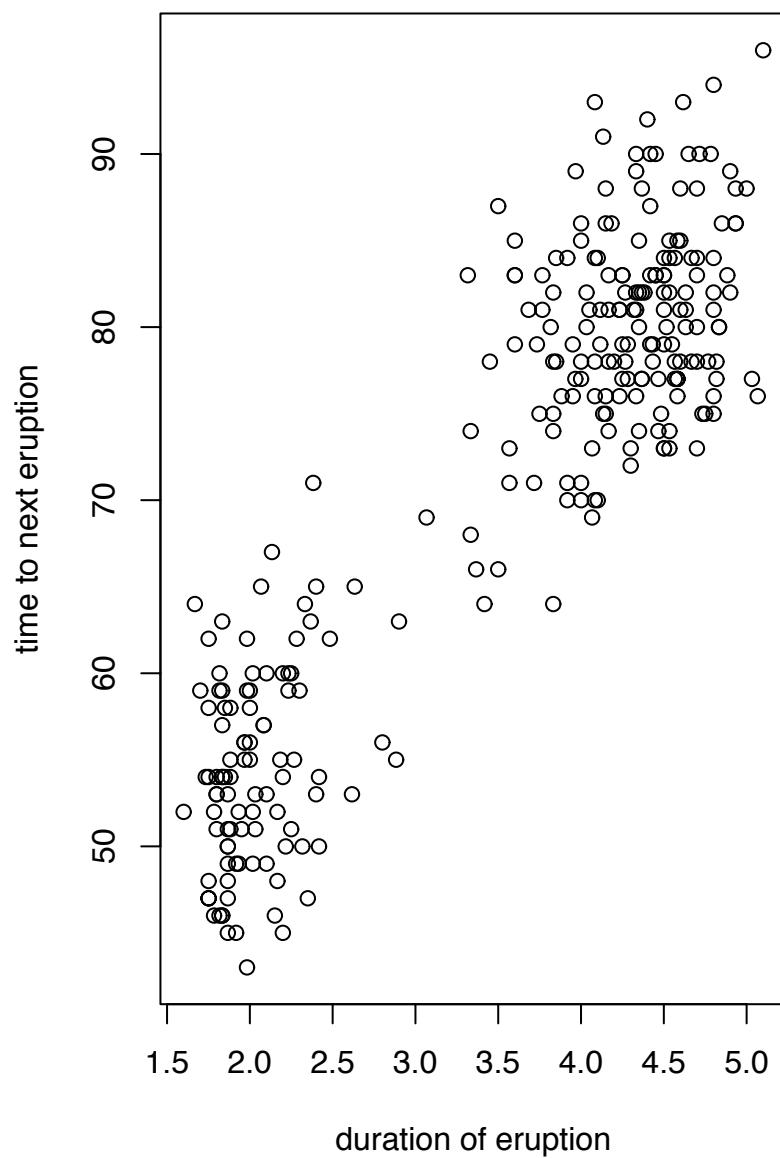


Figure 2.12: Waiting time versus duration in the Old Faithful dataset

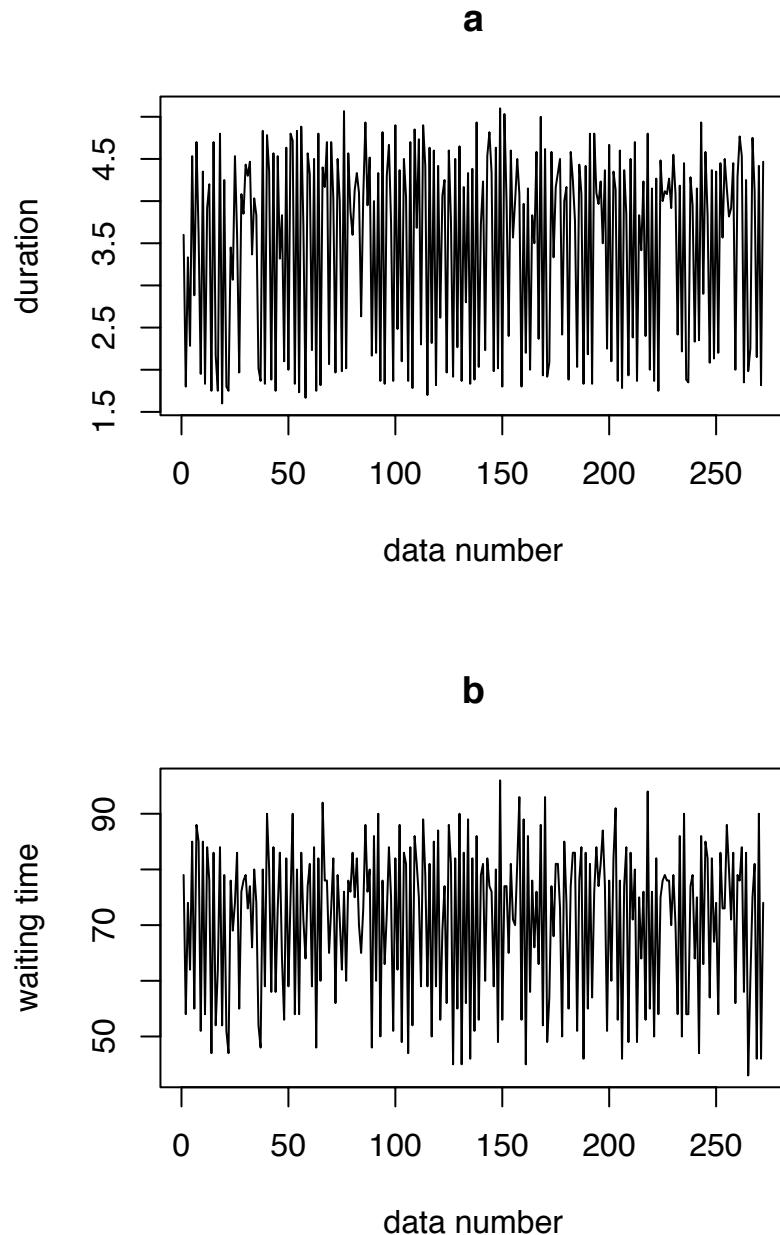


Figure 2.13: (a): duration and (b): waiting time plotted against data number in the Old Faithful dataset

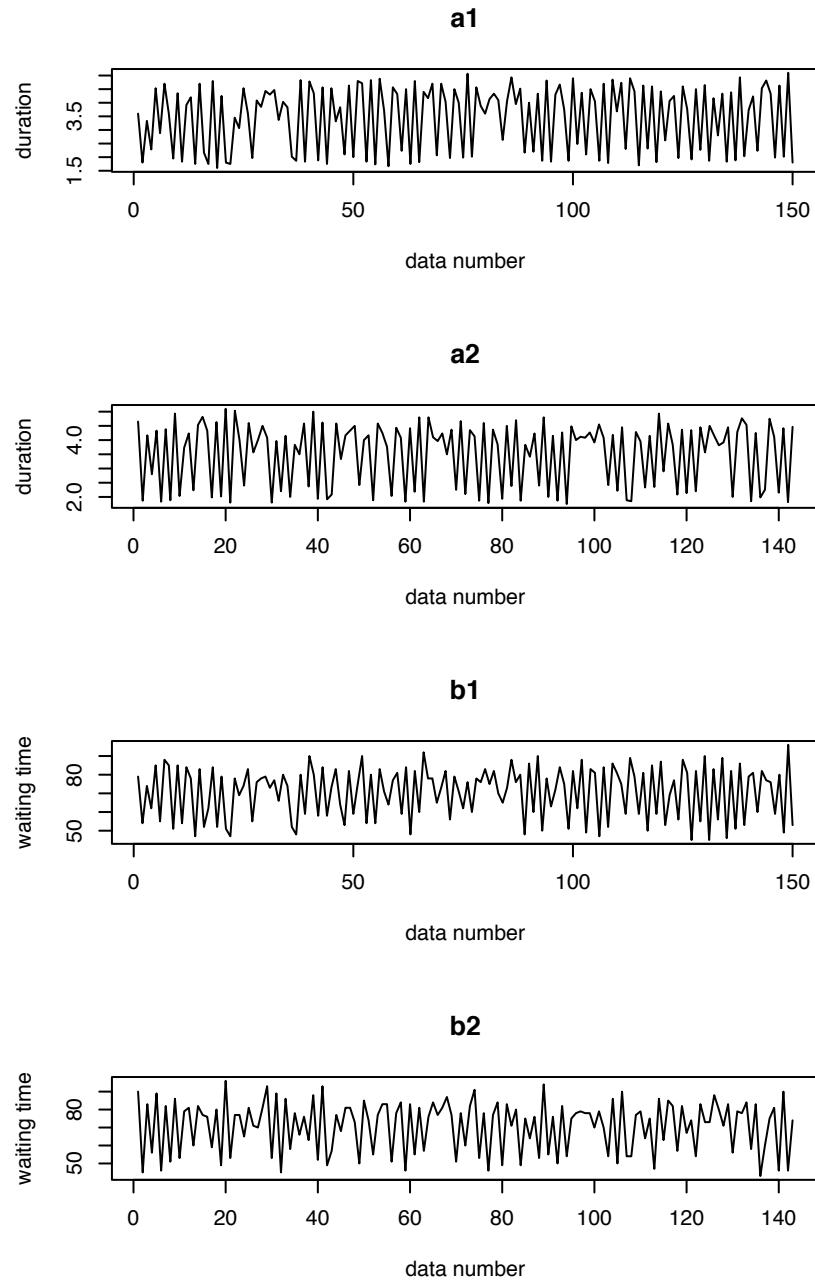


Figure 2.14: (a1), (a2): duration and (b1), (b2): waiting time plotted against data number in the Old Faithful dataset

```

 ylab="waiting time", main="b1")
plot.ts (waiting[130:272], xlab="data number",
 ylab="waiting time", main="b2")

```

Figures 2.15 and 2.16 introduce *coplots*, a tool for visualizing the relationship among three variables. They represent the ocean temperature data from Example 1.5. In Figure 2.15 there are six panels in which temperature is plotted against latitude. Each panel is made from the points in a restricted range of longitude. The upper panel, the one spanning the top of the Figure, shows the six different ranges of longitude. For example, the first longitude range runs from about -10 to about -17. Points whose longitude is in the interval  $(-17, -10)$  go into the upper right panel of scatterplots. These are the points very close to the mouth of the Mediterranean Sea. Looking at that panel we see that temperature increases very steeply from South to North, until about  $35^\circ$ , at which point they start to decrease steeply as we go further North. That's because we're crossing the Mediterranean tongue at a point very close to its source.

The other longitude ranges are about  $(-20, -13)$ ,  $(-25, -16)$ ,  $(-30, -20)$ ,  $(-34, -25)$  and  $(-40, -28)$ . They are used to create the scatterplot panels in the upper center, upper left, lower right, lower center, and lower left, respectively. The general impression is

- temperatures decrease slightly as we move East to West,
- the angle in the scatterplot becomes slightly shallower as we move East to West, and
- there are some points that don't fit the general pattern.

Notice that the longitude ranges are overlapping and not of equal width. The ranges are chosen by R to have a little bit of overlap and to put roughly equal numbers of points into each range.

Figure 2.16 reverses the roles of latitude and longitude. The impression is that temperature increases gradually from West to East. These two figures give a fairly clear picture of the Mediterranean tongue.

Figures 2.15 and 2.16 were produced by

```

coplot (temp ~ lat | lon)
coplot (temp ~ lon | lat)

```

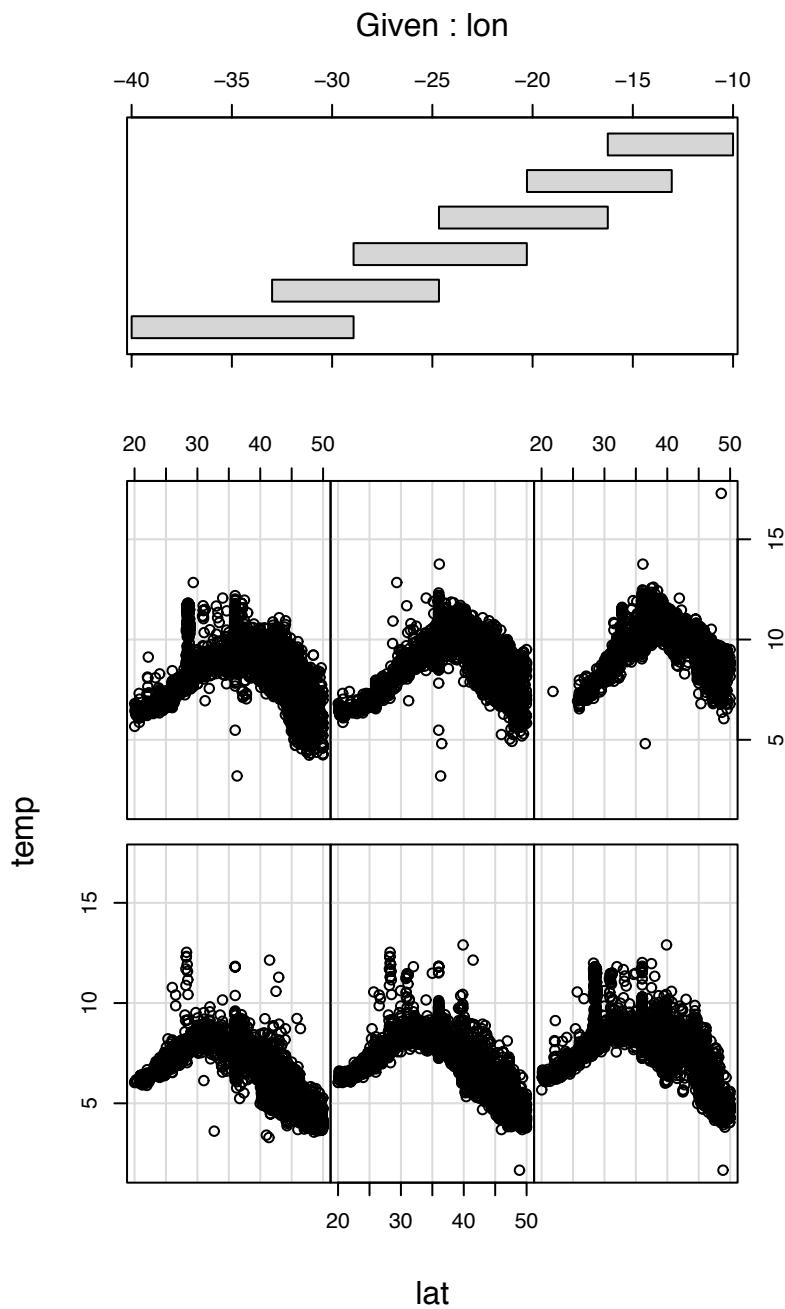


Figure 2.15: Temperature versus latitude for different values of longitude

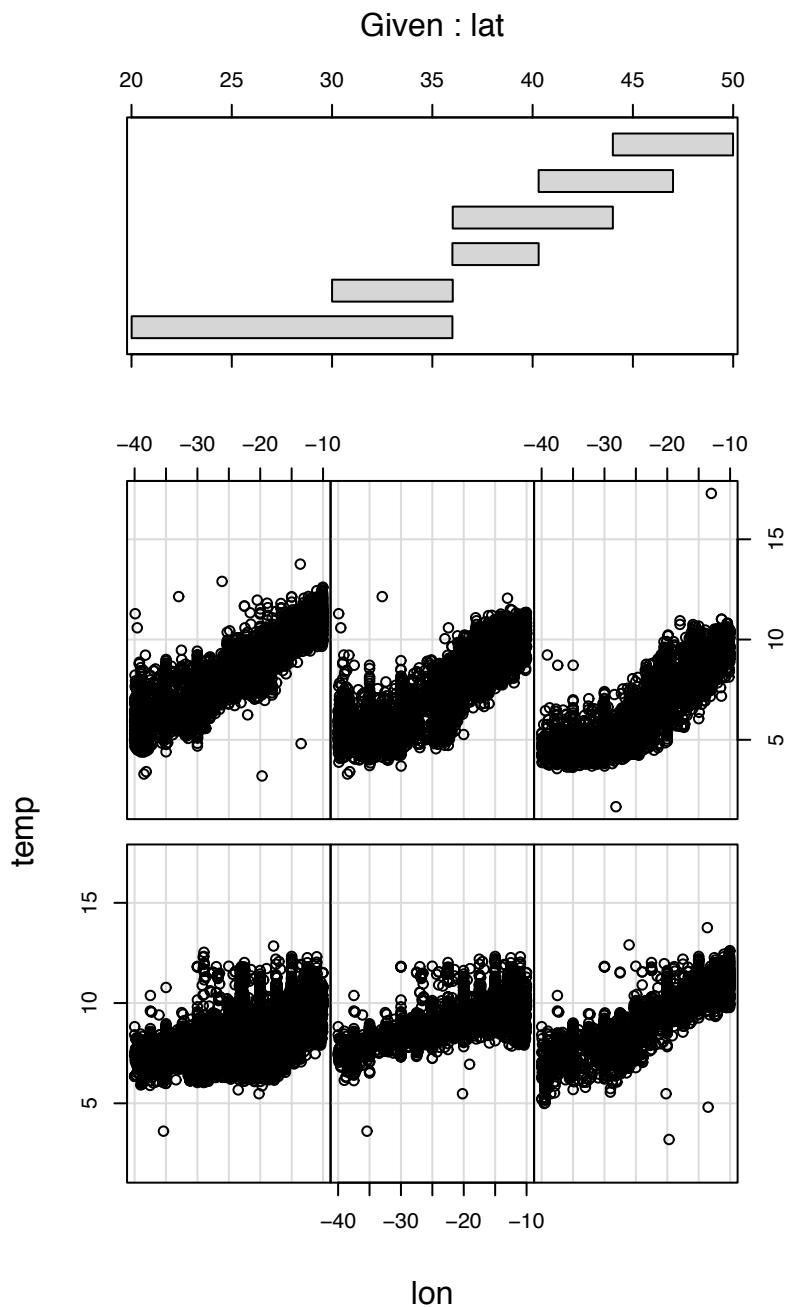


Figure 2.16: Temperature versus longitude for different values of latitude

Example 2.6 shows one way to display the relationship between two sequences of events.

**Example 2.6** (Neurobiology)

To learn how the brain works, neurobiologists implant electrodes into animal brains. These electrodes are fine enough to record the firing times of individual neurons. A sequence of firing times of a neuron is called a *spike train*. Figure 2.17 shows the spike train from one neuron in the gustatory cortex of a rat while the rat was in an experiment on taste. This particular rat was in the experiment for a little over 80 minutes. Those minutes are marked on the *y*-axis. The *x*-axis is marked in seconds. Each dot on the plot shows a time at which the neuron fired. We can see, for example, that this neuron fired about nine times in the first five seconds, then was silent for about the next ten seconds. We can also see, for example, that this neuron undergoes some episodes of very rapid firing lasting up to about 10 seconds.

Since this neuron is in the gustatory cortex — the part of the brain responsible for taste — it is of interest to see how the neuron responds to various tastes. During the experiment the rat was licking a tube that sometimes delivered a drop of water and sometimes delivered a drop of water in which a chemical, or *tastant*, was dissolved. The 55 short vertical lines on the plot show the times at which the rat received a drop of 300 millimolar (.3 M) solution of NaCl. We can examine the plot for relationships between deliveries of NaCl and activity of the neuron.

Figure 2.17 was produced by

```
datadir <- "~/research/neuro/data/stapleton/"
spikes <- list (
 sig002a = scan (paste (datadir, "sig002a.txt", sep="")),
 sig002b = scan (paste (datadir, "sig002b.txt", sep="")),
 sig002c = scan (paste (datadir, "sig002c.txt", sep="")),
 sig003a = scan (paste (datadir, "sig003a.txt", sep="")),
 sig003b = scan (paste (datadir, "sig003b.txt", sep="")),
 sig004a = scan (paste (datadir, "sig004a.txt", sep="")),
 sig008a = scan (paste (datadir, "sig008a.txt", sep="")),
 sig014a = scan (paste (datadir, "sig014a.txt", sep="")),
 sig014b = scan (paste (datadir, "sig014b.txt", sep="")),
 sig017a = scan (paste (datadir, "sig017a.txt", sep=""))
)
tastants <- list (
```

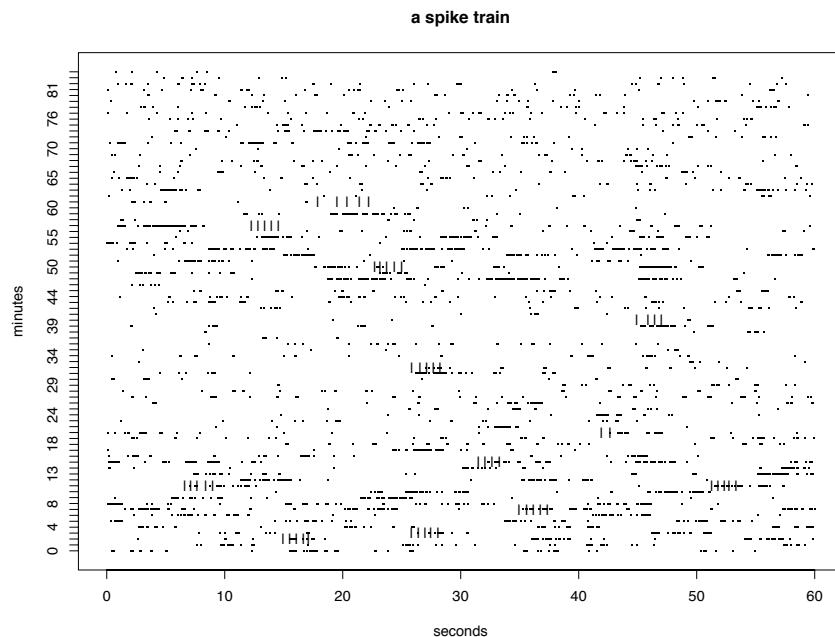


Figure 2.17: Spike train from a neuron during a taste experiment. The dots show the times at which the neuron fired. The solid lines show times at which the rat received a drop of a .3 M solution of NaCl.

```

MSG100 = scan (paste (datadir, "MSG100.txt", sep="")),
MSG300 = scan (paste (datadir, "MSG300.txt", sep="")),
NaCl100 = scan (paste (datadir, "NaCl100.txt", sep="")),
NaCl300 = scan (paste (datadir, "NaCl300.txt", sep="")),
water = scan (paste (datadir, "water.txt", sep=""))
)
stripchart (spikes[[8]] %~ 60 ~ spikes[[8]] %% 60, pch=".",
 main="a spike train", xlab="seconds", ylab="minutes")
points (tastants$NaCl300 %% 60, tastants$NaCl300 %% 60 + 1,
 pch="|")

```

- The line `datadir <- ...` stores the name of the directory in which I keep the neuro data. When used in `paste` it identifies individual files.
- The command `list()` creates a list. The elements of a list can be anything. In this case the list named `spikes` has ten elements whose names are `sig002a`, `sig002b`, ..., and `sig017a`. The list named `tastants` has five elements whose names are `MSG100`, `MSG300`, `NaCl100`, `NaCl300`, and `water`. Lists are useful for keeping related objects together, especially when those objects aren't all of the same type.
- Each element of the list is the result of a `scan()`. `scan()` reads a file and stores the result in a vector. So `spikes` is a list of ten vectors. Each vector contains the firing times, or a spike train, of one neuron. `tastants` is a list of five vectors. Each vector contains the times at which a particular tastant was delivered.
- There are two ways to refer an element of a list. For example, `spikes[[8]]` refers to the eighth element of `spikes` while `tastants$NaCl300` refers to the element named `NaCl300`.
- Lists are useful for keeping related objects together, especially when those objects are not the same type. In this example `spikes$sig002a` is a vector whose length is the number of times neuron 002a fired, while the length of `spikes$sig002b` is the number of times neuron 002b fired. Since those lengths are not the same, the data don't fit neatly into a matrix, so we use a list instead.

## 2.3 Likelihood

### 2.3.1 The Likelihood Function

It often happens that we observe data from a distribution that is not known precisely but whose general form is known. For example, we may know that the data come from a Poisson distribution,  $X \sim \text{Poi}(\lambda)$ , but we don't know the value of  $\lambda$ . We may know that  $X \sim \text{Bin}(n, \theta)$  but not know  $\theta$ . Or we may know that the values of  $X$  are densely clustered around some central value and sparser on both sides, so we decide to model  $X \sim N(\mu, \sigma)$ , but we don't know the values of  $\mu$  and  $\sigma$ . In these cases there is a whole family of probability distributions indexed by either  $\lambda$ ,  $\theta$ , or  $(\mu, \sigma)$ . We call  $\lambda$ ,  $\theta$ , or  $(\mu, \sigma)$  the unknown *parameter*; the family of distributions is called a *parametric family*. Often, the goal of the statistical analysis is to learn about the value of the unknown parameter. Of course, learning which value of the parameter is the true one, or which values of the parameter are plausible in light of the data, is the same as learning which member of the family is the true one, or which members of the family are plausible in light of the data.

The different values of the parameter, or the different members of the family, represent different theories or hypotheses about nature. A sensible way to discriminate among the theories is according to how well they explain the data. Recall the Seedlings data (Examples 1.4, 1.6, 1.7 and 1.9) in which  $X$  was the number of new seedlings in a forest quadrat,  $X \sim \text{Poi}(\lambda)$ , and different values of  $\lambda$  represent different theories or hypotheses about the arrival rate of new seedlings. When  $X$  turned out to be 3, how well a value of  $\lambda$  explains the data is measured by  $\Pr[X = 3 | \lambda]$ . This probability, as a function of  $\lambda$ , is called the *likelihood function* and denoted  $\ell(\lambda)$ . It says how well each value of  $\lambda$  explains the datum  $X = 3$ . Figure 1.6 (pg. 19) is a plot of the likelihood function.

In a typical problem we know the data come from a parametric family indexed by a parameter  $\theta$ , i.e.  $X_1, \dots, X_n \sim \text{i.i.d.} f(x | \theta)$ , but we don't know  $\theta$ . The joint density of all the data is

$$f(X_1, \dots, X_n | \theta) = \prod f(X_i | \theta). \quad (2.2)$$

Equation 2.2, as a function of  $\theta$ , is the likelihood function. We sometimes write  $f(\text{Data} | \theta)$  instead of indicating each individual datum. To emphasize that we are thinking of a function of  $\theta$  we may also write the likelihood function as  $\ell(\theta)$  or  $\ell(\theta | \text{Data})$ .

The interpretation of the likelihood function is always in terms of ratios. If, for example,  $\ell(\theta_1)/\ell(\theta_2) > 1$ , then  $\theta_1$  explains the data better than  $\theta_2$ . If  $\ell(\theta_1)/\ell(\theta_2) = k$ , then  $\theta_1$  explains the data  $k$  times better than  $\theta_2$ . To illustrate, suppose students in a statistics class conduct a study to estimate the fraction of cars on Campus Drive that are red. Student A decides to observe the first 10 cars and record  $X$ , the number that are red. Student A

observes

$$NR, R, NR, NR, NR, R, NR, NR, NR, R$$

and records  $X = 3$ . She did a Binomial experiment; her statistical model is  $X \sim \text{Bin}(10, \theta)$ ; her likelihood function is  $\ell_A(\theta) = \binom{10}{3}\theta^3(1-\theta)^7$ . It is plotted in Figure 2.18. Because only ratios matter, the likelihood function can be rescaled by any arbitrary positive constant. In Figure 2.18 it has been rescaled so the maximum is 1. The interpretation of Figure 2.18 is that values of  $\theta$  around  $\theta \approx 0.3$  explain the data best, but that any value of  $\theta$  in the interval from about 0.1 to about 0.6 explains the data not too much worse than the best. I.e.,  $\theta \approx 0.3$  explains the data only about 10 times better than  $\theta \approx 0.1$  or  $\theta \approx 0.6$ , and a factor of 10 is not really very much. On the other hand, values of  $\theta$  less than about 0.05 or greater than about 0.7 explain the data much worse than  $\theta \approx 0.3$ .

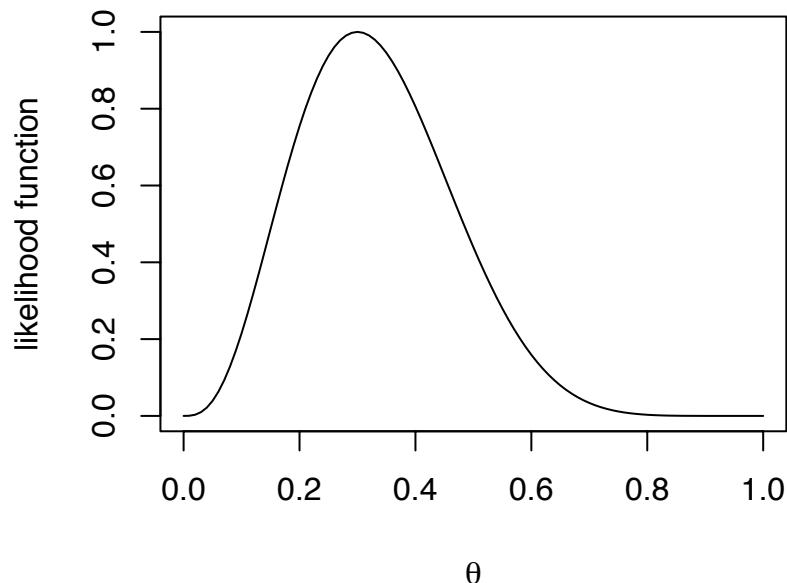


Figure 2.18: Likelihood function  $\ell(\theta)$  for the proportion  $\theta$  of red cars on Campus Drive

Figure 2.18 was produced by the following snippet.

```
theta <- seq(0, 1, by=.01) # some values of theta
y <- dbinom(3, 10, theta) # calculate l(theta)
y <- y / max(y) # rescale
plot(theta, y, type="l", xlab=expression(theta),
 ylab="likelihood function")
```

- `expression` is R's way of getting mathematical symbols and formulae into plot labels. For more information, type `help(plotmath)`.

To continue the example, Student B decides to observe cars until the third red one drives by and record  $Y$ , the total number of cars that drive by until the third red one. Students A and B went to Campus Drive at the same time and observed the same cars. B records  $Y = 10$ . For B the likelihood function is

$$\begin{aligned}\ell_B(\theta) &= P[Y = 10 | \theta] \\ &= P[2 \text{ reds among first 9 cars}] \times P[10^{\text{th}} \text{ car is red}] \\ &= \binom{9}{2} \theta^2 (1 - \theta)^7 \times \theta \\ &= \binom{9}{2} \theta^3 (1 - \theta)^7\end{aligned}$$

$\ell_B$  differs from  $\ell_A$  by the multiplicative constant  $\binom{9}{2}/\binom{10}{3}$ . But since multiplicative constants don't matter, A and B really have the same likelihood function and hence exactly the same information about  $\theta$ . Student B would also use Figure 2.18 as the plot of her likelihood function.

Student C decides to observe every car for a period of 10 minutes and record  $Z_1, \dots, Z_k$  where  $k$  is the number of cars that drive by in 10 minutes and each  $Z_i$  is either 1 or 0 according to whether the  $i^{\text{th}}$  car is red. When C went to Campus Drive with A and B, only 10 cars drove by in the first 10 minutes. Therefore C recorded exactly the same data as A and B. Her likelihood function is

$$\ell_C(\theta) = (1 - \theta)\theta(1 - \theta)(1 - \theta)(1 - \theta)\theta(1 - \theta)(1 - \theta)\theta = \theta^3(1 - \theta)^7$$

$\ell_C$  is proportional to  $\ell_A$  and  $\ell_B$  and hence contains exactly the same information and looks exactly like Figure 2.18. So even though the students planned different experiments they ended up with the same data, and hence the same information about  $\theta$ .

The next example follows the Seedling story and shows what happens to the likelihood function as data accumulates.

**Example 2.7** (Seedlings, cont.)

Examples 1.4, 1.6, 1.7, and 1.9 reported data from a single quadrat on the number of new seedlings to emerge in a given year. In fact, ecologists collected data from multiple quadrats over multiple years. In the first year there were 60 quadrats and a total of 40 seedlings so the likelihood function was

$$\begin{aligned}\ell(\lambda) &\equiv p(\text{Data} | \lambda) \\ &= p(y_1, \dots, y_{60} | \lambda) \\ &= \prod_1^{60} p(y_i | \lambda) \\ &= \prod_1^{60} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \\ &\propto e^{-60\lambda} \lambda^{40}\end{aligned}$$

Note that  $\prod y_i!$  is a multiplicative factor that does not depend on  $\lambda$  and so is irrelevant to  $\ell(\lambda)$ . Note also that  $\ell(\lambda)$  depends only on  $\sum y_i$ , not on the individual  $y_i$ 's. I.e., we only need to know  $\sum y_i = 40$ ; we don't need to know the individual  $y_i$ 's.  $\ell(\lambda)$  is plotted in Figure 2.19. Compare to Figure 1.6 (pg. 19). Figure 2.19 is much more peaked. That's because it reflects much more information, 60 quadrats instead of 1. The extra information pins down the value of  $\lambda$  much more accurately.

Figure 2.19 was created with

```
lam <- seq(0, 2, length=50)
lik <- dpois(40, 60*lam)
lik <- lik / max(lik)
plot(lam, lik, xlab=expression(lambda),
 ylab="likelihood", type="l")
```

The next example is about a possible cancer cluster in California.

**Example 2.8** (Slater School)

This example was reported in BRODEUR [1992]. See LAVINE [1999] for further analysis.

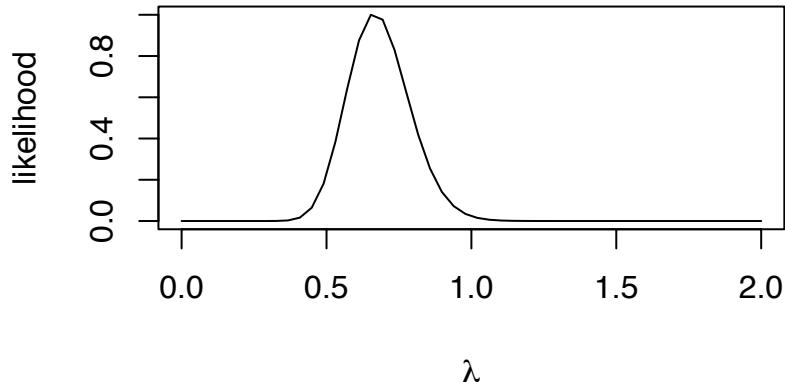


Figure 2.19:  $\ell(\theta)$  after  $\sum y_i = 40$  in 60 quadrats.

The Slater school is an elementary school in Fresno, California where teachers and staff were “concerned about the presence of two high-voltage transmission lines that ran past the school . . . .” Their concern centered on the “high incidence of cancer at Slater. . . .” To address their concern, Dr. Raymond Neutra of the California Department of Health Services’ Special Epidemiological Studies Program conducted a statistical analysis on the

“eight cases of invasive cancer, . . . , the total years of employment of the hundred and forty-five teachers, teachers’ aides, and staff members, . . . , [and] the number of person-years in terms of National Cancer Institute statistics showing the annual rate of invasive cancer in American women between the ages of forty and forty-four — the age group encompassing the average age of the teachers and staff at Slater — [which] enabled him to calculate that 4.2 cases of cancer could have been expected to occur among the Slater teachers and staff members . . . .”

For our purposes we can assume that  $X$ , the number of invasive cancer cases at the Slater School has the Binomial distribution  $X \sim \text{Bin}(145, \theta)$ . We observe  $x = 8$ . The likelihood function

$$\ell(\theta) \propto \theta^8(1 - \theta)^{137} \quad (2.3)$$

is pictured in Figure 2.20. From the Figure it appears that values of  $\theta$  around .05 or .06, explain the data better than values less than .05 or greater than .06, but that values of  $\theta$  anywhere from about .02 or .025 up to about .11 explain the data reasonably well.

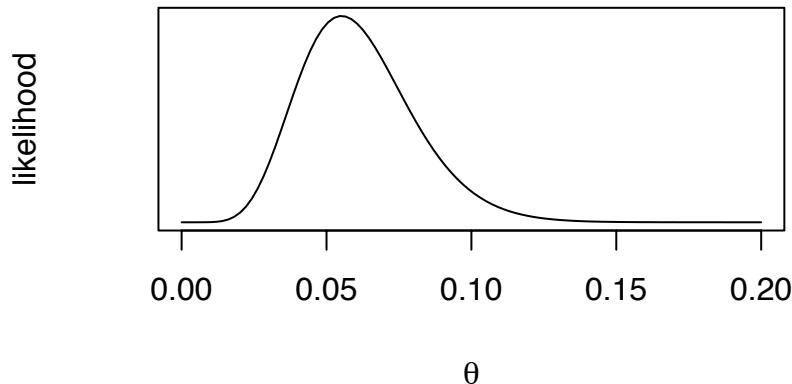


Figure 2.20: Likelihood for Slater School

Figure 2.20 was produced by the following R code.

```
theta <- seq(0, .2, length=100)
lik <- dbinom(8, 145, theta)
lik <- lik / max(lik)
plot(theta, lik, xlab=expression(theta),
 ylab="likelihood", type="l", yaxt="n")
```

The first line of code creates a sequence of 100 values of  $\theta$  at which to compute  $\ell(\theta)$ , the second line does the computation, the third line rescales so the maximum likelihood is 1, and the fourth line makes the plot.

Examples 2.7 and 2.8 show how likelihood functions are used. They reveal which values of a parameter the data support (equivalently, which values of a parameter explain the data well) and values they don't support (which values explain the data poorly). There is

no hard line between support and non-support. Rather, the plot of the likelihood functions shows the smoothly varying levels of support for different values of the parameter.

Because likelihood ratios measure the strength of evidence for or against one hypothesis as opposed to another, it is important to ask how large a likelihood ratio needs to be before it can be considered strong evidence. Or, to put it another way, how strong is the evidence in a likelihood ratio of 10, or 100, or 1000, or more? One way to answer the question is to construct a *reference experiment*, one in which we have an intuitive understanding of the strength of evidence and can calculate the likelihood; then we can compare the calculated likelihood to the known strength of evidence.

For our reference experiment imagine we have two coins. One is a fair coin, the other is two-headed. We randomly choose a coin. Then we conduct a sequence of coin tosses to learn which coin was selected. Suppose the tosses yield  $n$  consecutive Heads.  $P[n \text{ Heads} | \text{fair}] = 2^{-n}$ ;  $P[n \text{ Heads} | \text{two-headed}] = 1$ . So the likelihood ratio is  $2^n$ . That's our reference experiment. A likelihood ratio around 8 is like tossing three consecutive Heads; a likelihood ratio around 1000 is like tossing ten consecutive Heads.

In Example 2.8  $\operatorname{argmax} \ell(\theta) \approx .055$  and  $\ell(.025)/\ell(.055) \approx .13 \approx 1/8$ , so the evidence against  $\theta = .025$  as opposed to  $\theta = .055$  is about as strong as the evidence against the fair coin when three consecutive Heads are tossed. The same can be said for the evidence against  $\theta = .1$ . Similarly,  $\ell(.011)/\ell(.055) \approx \ell(.15)/\ell(.055) \approx .001$ , so the evidence against  $\theta = .011$  or  $\theta = .15$  is about as strong as 10 consecutive Heads. A fair statement of the evidence is that  $\theta$ 's in the interval from about  $\theta = .025$  to about  $\theta = .1$  explain the data not much worse than the maximum of  $\theta \approx .055$ . But  $\theta$ 's below about .01 or larger than about .15 explain the data not nearly as well as  $\theta$ 's around .055.

### 2.3.2 Likelihoods from the Central Limit Theorem

Sometimes it is not possible to compute the likelihood function exactly, either because it is too difficult or because we don't know what it is. But we can often compute an approximate likelihood function using the Central Limit Theorem. The following example is the simplest case, but typifies the more exotic cases we will see later on.

Suppose we sample  $X_1, X_2, \dots, X_n$  from a probability density  $f$ . We don't know what  $f$  is; we don't even know what parametric family it belongs to. Assume that  $f$  has a mean  $\mu$  and an SD  $\sigma$  (I.e, assume that the mean and variance are finite.) and that we would like to learn about  $\mu$ . If  $(\mu, \sigma)$  are the only unknown parameters then the likelihood function is  $\ell(\mu, \sigma) = f(\text{Data} | \mu, \sigma) = \prod f(X_i | \mu, \sigma)$ . But we don't know  $f$  and can't calculate  $\ell(\mu, \sigma)$ .

However, we can reason as follows.

1. Most of the information in the data for learning about  $\mu$  is contained in  $\bar{X}$ . That is,  $\bar{X}$  tells us a lot about  $\mu$  and the deviations  $\delta_i \equiv X_i - \bar{X}, i = 1, \dots, n$  tell us very little.

2. If  $n$  is large then the Central Limit Theorem tells us

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}), \quad \text{approximately}$$

3. We can estimate  $\sigma^2$  from the data by

$$\hat{\sigma}^2 = s^2 = \sum \delta_i^2/n$$

4. And therefore the function

$$\ell_M(\mu) \propto \exp\left(-\frac{1}{2}\left(\frac{\mu - \bar{X}}{\hat{\sigma}/\sqrt{n}}\right)^2\right) \quad (2.4)$$

is a good approximation to the likelihood function.

In the preceding reasoning we separated the data into two parts —  $\bar{X}$  and  $\{\delta_i\}$ ; used  $\{\delta_i\}$  to estimate  $\sigma$ ; and used  $\bar{X}$  to find a likelihood function for  $\mu$ . We cannot, in general, justify such a separation mathematically. We justified it if and when our main interest is in  $\mu$  and we believe  $\{\delta_i\}$  tell us little about  $\mu$ .

Function 2.4 is called a *marginal likelihood* function. TSOU AND ROYALL [1995] show that marginal likelihoods are good approximations to true likelihoods and can be used to make accurate inferences, at least in cases where the Central Limit Theorem applies. We shall use marginal likelihoods throughout this book.

### Example 2.9 (Slater School, continued)

We redo the Slater School example (Example 2.8) to illustrate the marginal likelihood and see how it compares to the exact likelihood. In that example the  $X_i$ 's were 1's and 0's indicating which teachers got cancer. There were 8 1's out of 145 teachers, so  $\bar{X} = 8/145 \approx .055$ . Also,  $\hat{\sigma}^2 = (8(137/145)^2 + 137(8/145)^2)/145 \approx .052$ , so  $\hat{\sigma} \approx .23$ . We get

$$\ell_M(\mu) \propto \exp\left(-\frac{1}{2}\left(\frac{\mu - .055}{.23/\sqrt{145}}\right)^2\right) \quad (2.5)$$

Figure 2.21 shows the marginal and exact likelihood functions. The marginal likelihood is a reasonably good approximation to the exact likelihood.

Figure 2.21 was produced by the following snippet.

```
theta <- seq (0, .2, length=100)
lik <- dbeta (theta, 9, 138)
```

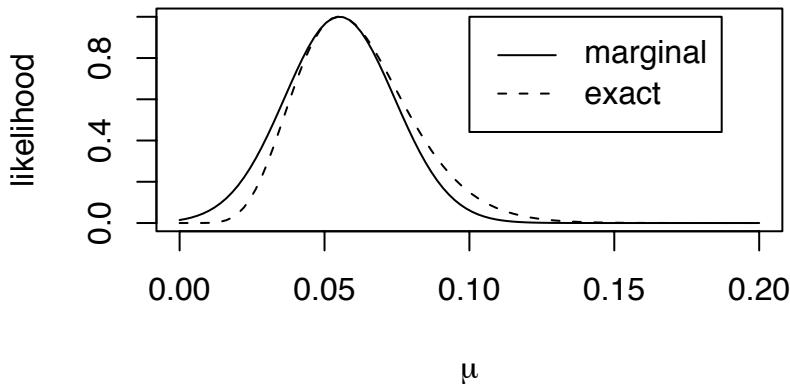


Figure 2.21: Marginal and exact likelihoods for Slater School

```

lik.mar <- dnorm (theta, 8/145,
 sqrt((8*(137/145)^2 + 137*(8/145)^2)/145)/sqrt(145))
lik <- lik/max(lik)
lik.mar <- lik.mar/max(lik.mar)
matplot (theta, cbind(lik,lik.mar), xlab=expression(mu),
 ylab="likelihood", type="l", lty=c(2,1), col=1)
legend (.1, 1, c("marginal", "exact"), lty=c(1,2))

```

### Example 2.10 (CEO salary)

How much are corporate CEO's paid? Forbes magazine collected data in 1993 that can begin to answer this question. The data are available on-line at DASL, the *Data and Story Library*, a collection of data sets for free use by statistics students. DASL says

"Forbes magazine published data on the best small firms in 1993. These were firms with annual sales of more than five and less than \$350 million. Firms were ranked by five-year average return on investment. The data extracted are the age and annual salary of the chief executive officer for

the first 60 ranked firms. In question are the distribution patterns for the ages and the salaries."

You can download the data from

<HTTP://LIB.STAT.CMU.EDU/DASL/DATAFILES/CEODAT.HTML>. The first few lines look like this:

AGE SAL

|    |     |
|----|-----|
| 53 | 145 |
| 43 | 621 |
| 33 | 262 |

In this example we treat the Forbes data as a random sample of size  $n = 60$  of CEO salaries for small firms. We're interested in the average salary  $\mu$ . Our approach is to calculate the marginal likelihood function  $\ell_M(\mu)$ .

Figure 2.22(a) shows a stripchart of the data. Evidently, most salaries are in the range of \$200 to \$400 thousand dollars, but with a long right-hand tail. Because the right-hand tail is so much larger than the left, the data are not even approximately Normally distributed. But the Central Limit Theorem tells us that  $\bar{X}$  is approximately Normally distributed, so the method of marginal likelihood applies. Figure 2.22(b) displays the marginal likelihood function  $\ell_M(\mu)$ .

Figure 2.22 was produced by the following snippet.

```
ceo <- read.table ("data/ceo_salaries/data",header=T)

par (mfrow=c(2,1))
stripchart (ceo$SAL, "jitter", pch=1, main="(a)",
 xlab="Salary (thousands of dollars)")

m <- mean (ceo$SAL, na.rm=T)
s <- sqrt (var(ceo$SAL,na.rm=T) / (length(ceo$SAL)-1))
x <- seq (340, 470, length=40)
y <- dnorm (x, m, s)
y <- y / max(y)
plot (x, y, type="l", xlab="mean salary",
 ylab="likelihood", main="(b)")
```

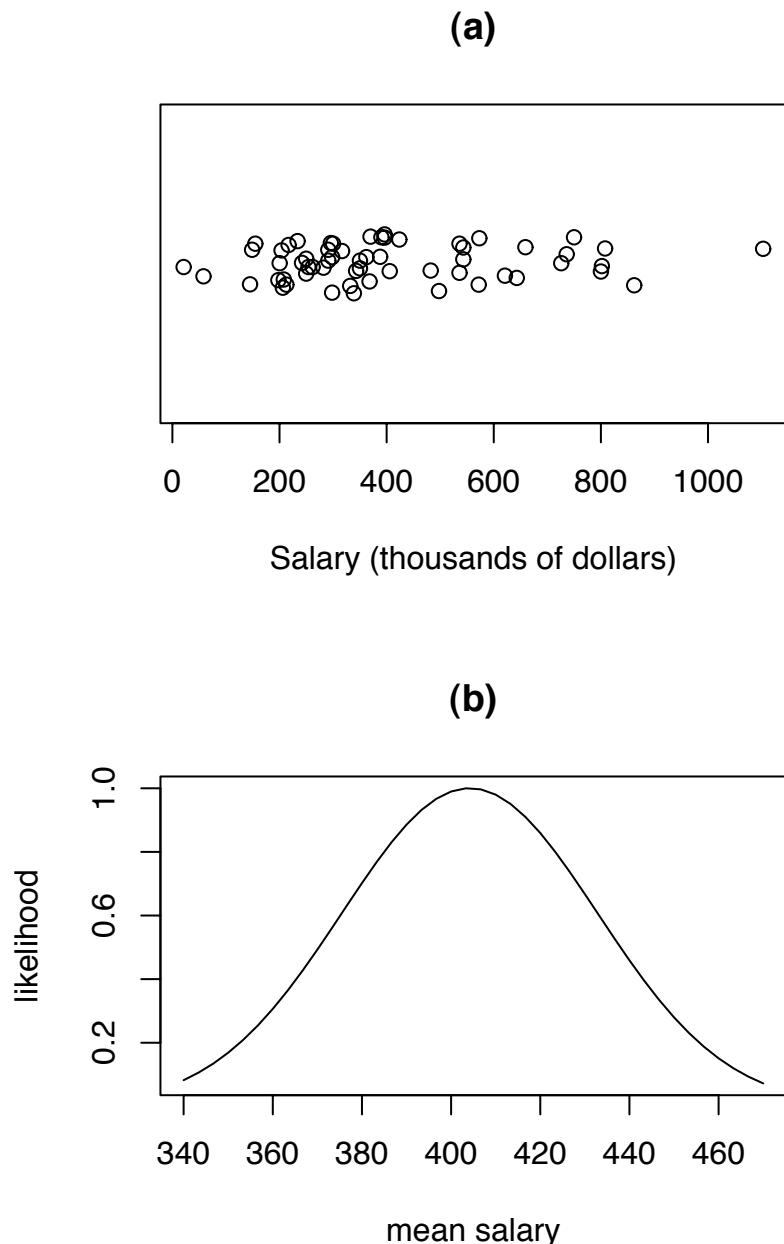


Figure 2.22: Marginal likelihood for mean CEO salary

- In `In s <- sqrt ... the (length(ceo$SAL)-1)` is there to account for one missing data point.
- `y <- y / max(y)` doesn't accomplish much and could be omitted.
- The data strongly support the conclusion that the mean salary is between about \$350 and \$450 thousand dollars. That's much smaller than the range of salaries on display in Figure 2.22(a). Why?
- Is inference about the mean salary useful in this data set? If not, what would be better?

### 2.3.3 Likelihoods for several parameters

What if there are two unknown parameters? Then the likelihood is a function of two variables. For example, if the  $X_i$ 's are a sample from  $N(\mu, \sigma)$  then the likelihood is a function of  $(\mu, \sigma)$ . The next example illustrates the point.

#### **Example 2.11** (FACE, continued)

This example continues Example 1.12 about a FACE experiment in Duke Forest. There were six rings; three were treated with excess CO<sub>2</sub>. The dominant canopy tree in the FACE experiment is *pinus taeda*, or loblolly pine. Figure 2.23a is a histogram of the final basal area of each loblolly pine in 1998 divided by its initial basal area in 1996. It shows that the trees in Ring 1 grew an average of about 30% but with variability that ranged from close to 0% on the low end to around 50% or 60% on the high end. Because the data are clustered around a central value and fall off roughly equally on both sides they can be well approximated by a Normal distribution. But with what mean and SD? What values of  $(\mu, \sigma)$  might reasonably produce the histogram in Figure 2.23a?

The likelihood function is

$$\begin{aligned}\ell(\mu, \sigma) &= \prod_1^n f(x_i | \mu, \sigma) \\ &= \prod_1^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} \\ &\propto \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum_1^n (x_i - \mu)^2}\end{aligned}$$

Figure 2.23b is a contour plot of the likelihood function. The dot in the center, where

$(\mu, \sigma) \approx (1.27, .098)$ , is where the likelihood function is highest. That is the value of  $(\mu, \sigma)$  that best explains the data. The next contour line is drawn where the likelihood is about 1/4 of its maximum; then the next is at 1/16 the maximum, the next at 1/64, and the last at 1/256 of the maximum. They show values of  $(\mu, \sigma)$  that explain the data less and less well.

Ecologists are primarily interested in  $\mu$  because they want to compare the  $\mu$ 's from different rings to see whether the excess CO<sub>2</sub> has affected the average growth rate. (They're also interested in the  $\sigma$ 's, but that's a secondary concern.) But  $\ell$  is a function of both  $\mu$  and  $\sigma$ , so it's not immediately obvious that the data tell us anything about  $\mu$  by itself. To investigate further, Figure 2.23c shows slices through the likelihood function at  $\sigma = .09, .10$ , and  $.11$ , the locations of the dashed lines in Figure 2.23b. The three curves are almost identical. Therefore, the relative support for different values of  $\mu$  does not depend very much on the value of  $\sigma$ , and therefore we are justified in interpreting any of the curves in Figure 2.23c as a "likelihood function" for  $\mu$  alone, showing how well different values of  $\mu$  explain the data. In this case, it looks as though values of  $\mu$  in the interval (1.25, 1.28) explain the data much better than values outside that interval.

Figure 2.23 was produced with

```
par (mfrow=c(2,2)) # a 2 by 2 array of plots
x <- ba98$BA.final / ba96$BA.init
x <- x[!is.na(x)]
hist (x, prob=T, xlab="basal area ratio",
 ylab="", main="(a)")
mu <- seq (1.2, 1.35, length=50)
sd <- seq (.08, .12, length=50)
lik <- matrix (NA, 50, 50)
for (i in 1:50)
 for (j in 1:50)
 lik[i,j] = prod (dnorm (x, mu[i], sd[j]))
lik <- lik / max(lik)
contour (mu, sd, lik, levels=4^(-4:0), drawlabels=F,
 xlab=expression(mu), ylab=expression(sigma),
 main="(b)")
abline (h = c (.09, .1, .11), lty=2)
lik.09 <- lik[,13] / max(lik[,13])
lik.10 <- lik[,26] / max(lik[,26])
lik.11 <- lik[,38] / max(lik[,38])
```

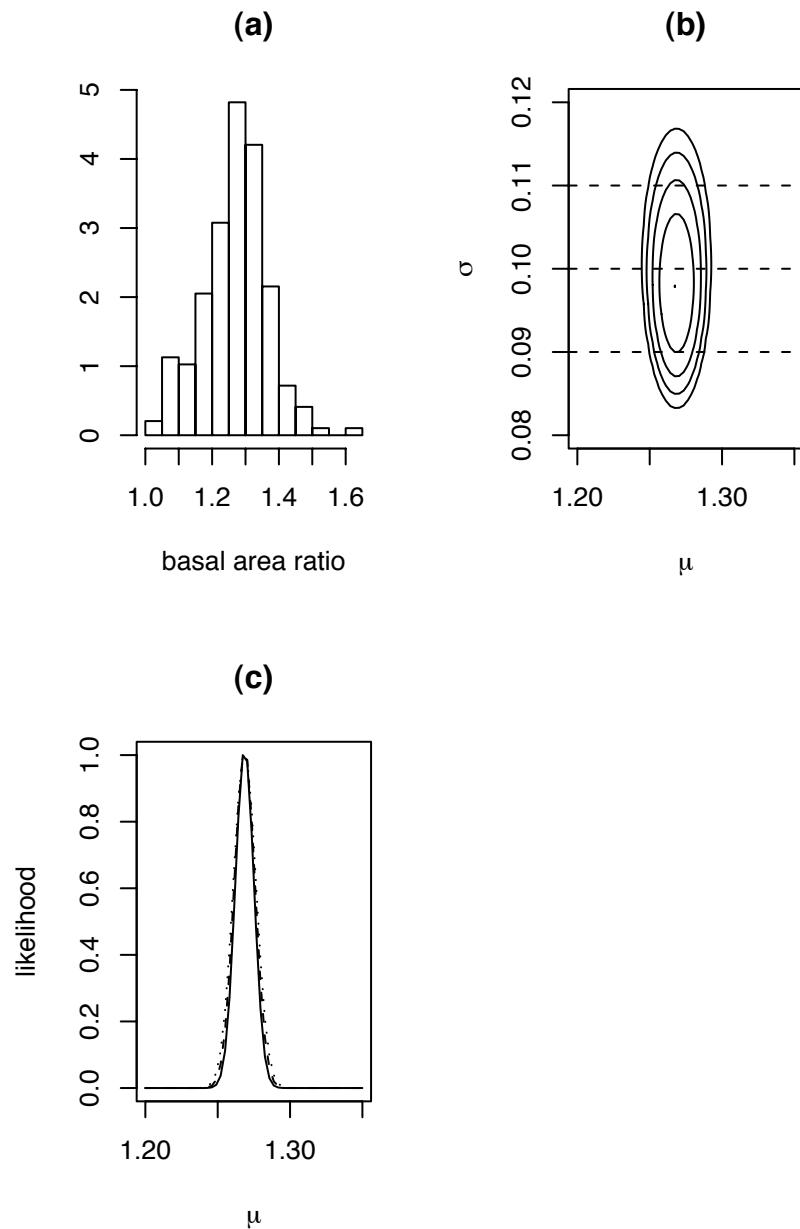


Figure 2.23: FACE Experiment, Ring 1. (a):  $(1998 \text{ final basal area}) \div (1996 \text{ initial basal area})$ ; (b): contours of the likelihood function. (c): slices of the likelihood function.

```
matplot (mu, cbind (lik.09, lik.10, lik.11), type="l",
 col=1, main="(c)",
 xlab=expression(mu), ylab="likelihood")
```

- The line `x <- x[!is.na(x)]` is there because some data is missing. This line selects only those data that are not missing and keeps them in `x`. When `x` is a vector, `is.na(x)` is another vector, the same length as `x`, with TRUE or FALSE, indicating where `x` is missing. The `!` is “not”, or negation, so `x[!is.na(x)]` selects only those values that are not missing.
- The lines `mu <- ...` and `sd <- ...` create a grid of  $\mu$  and  $\sigma$  values at which to evaluate the likelihood.
- The line `lik <- matrix ( NA, 50, 50 )` creates a matrix for storing the values of  $\ell(\mu, \sigma)$  on the grid. The next three lines are a loop to calculate the values and put them in the matrix.
- The line `lik <- lik / max(lik)` rescales all the values in the matrix so the maximum value is 1. Rescaling makes it easier to set the levels in the next line.
- `contour` produces a contour plot. `contour(mu, sd, lik, ...)` specifies the values on the x-axis, the values on the y-axis, and a matrix of values on the grid. The `levels` argument says at what levels to draw the contour lines, while `drawlabels=F` says not to print numbers on those lines. (Make your own contour plot without using `drawlabels` to see what happens.)
- `abline` is used for adding lines to plots. You can say either `abline(h=...)` or `abline(v=...)` to get horizontal and vertical lines, or `abline(intercept, slope)` to get arbitrary lines.
- `lik.09`, `lik.10`, and `lik.11` pick out three columns from the `lik` matrix. They are the three columns for the values of  $\sigma$  closest to  $\sigma = .09, .10, .11$ . Each column is rescaled so its maximum is 1.

### Example 2.12 (Quiz Scores, continued)

This example continues Example 2.3 about scores in Statistics 103. Figure 2.7 shows that most students scored between about 5 and 10, while 4 students were well below

the rest of the class. In fact, those students did not show up for every quiz so their averages were quite low. But the remaining students' scores were clustered together in a way that can be adequately described by a Normal distribution. What do the data say about  $(\mu, \sigma)$ ?

Figure 2.24 shows the likelihood function. The data support values of  $\mu$  from about 7.0 to about 7.6 and values of  $\sigma$  from about 0.8 to about 1.2. A good description of the data is that most of it follows a Normal distribution with  $(\mu, \sigma)$  in the indicated intervals, except for 4 students who had low scores not fitting the general pattern. Do you think the instructor should use this analysis to assign letter grades and, if so, how?

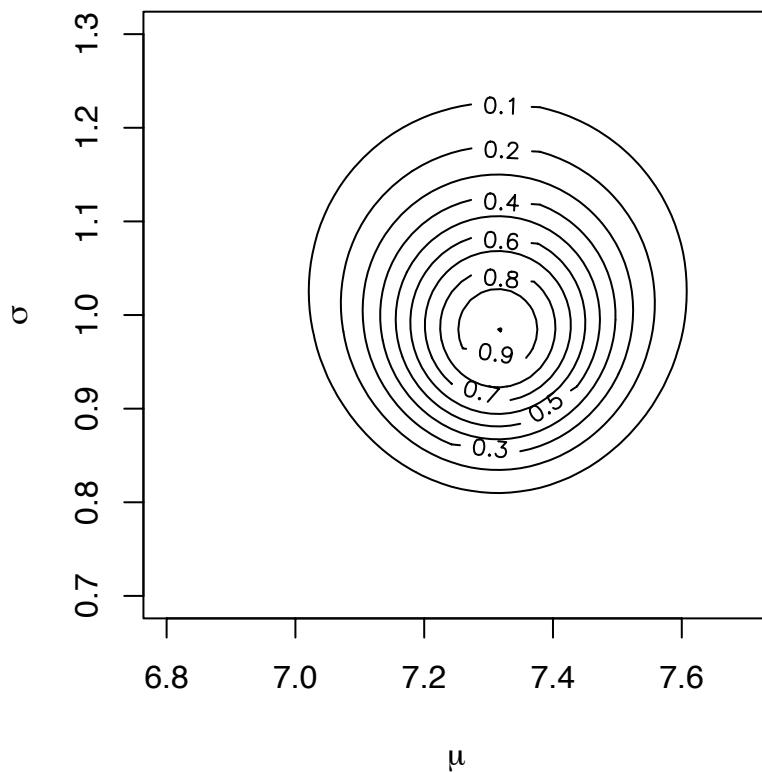


Figure 2.24: Likelihood function for Quiz Scores

Figure 2.24 was produced by

```
x <- sort(scores.ave)[5:58]
mu <- seq(6.8, 7.7, length=60)
sig <- seq(.7, 1.3, length=60)
lik <- matrix(NA, 60, 60)
for(i in 1:60)
 for(j in 1:60)
 lik[i,j] <- prod(dnorm(x, mu[i], sig[j]))
lik <- lik/max(lik)
contour(mu, sig, lik, xlab=expression(mu),
 ylab=expression(sigma))
```

Examples 2.11 and 2.12 have likelihood contours that are roughly circular, indicating that the likelihood function for one parameter does not depend very strongly on the value of the other parameter, and so we can get a fairly clear picture of what the data say about one parameter in isolation. But in other data sets two parameters may be inextricably entwined. Example 2.13 illustrates the problem.

### Example 2.13 (Seedlings, continued)

Examples 1.4, 1.6, and 1.7 introduced an observational study by ecologists to learn about tree seedling emergence and survival. Some species, Red Maple or *acer rubrum* for example, get a mark called a *bud scale scar* when they lose their leaves over winter. By looking for bud scale scars ecologists can usually tell whether an *acer rubrum* seedling is New (in its first summer), or Old (already survived through at least one winter). When they make their annual observations they record the numbers of New and Old *acer rubrum* seedlings in each quadrat. Every Old seedling in year  $t$  must have been either a New or an Old seedling in year  $t - 1$ .

Table 2.1 shows the 1992–1993 data for quadrat 6. Clearly the data are inconsistent; where did the Old seedling come from in 1993? When confronted with this paradox the ecologists explained that some New seedlings emerge from the ground after the date of the Fall census but before the winter. Thus they are not counted in the census their first year, but develop a bud scale scar and are counted as Old seedlings in their second year. One such seedling must have emerged in 1992, accounting for the Old seedling in 1993.

How shall we model the data? Let  $N_i^T$  be the true number of New seedlings in year  $i$ , i.e., including those that emerge after the census; and let  $N_i^O$  be the observed

| Year | No. of New seedlings | No. of Old seedlings |
|------|----------------------|----------------------|
| 1992 | 0                    | 0                    |
| 1993 | 0                    | 1                    |

Table 2.1: Numbers of New and Old seedlings in quadrat 6 in 1992 and 1993.

number of seedlings in year  $i$ , i.e., those that are counted in the census. As in Example 1.4 we model  $N_i^T \sim \text{Poi}(\lambda)$ . Furthermore, each seedling has some chance  $\theta_f$  of being found in the census. (Nominally  $\theta_f$  is the proportion of seedlings that emerge before the census, but in fact it may also include a component accounting for the failure of ecologists to find seedlings that have already emerged.) Treating the seedlings as independent and all having the same  $\theta_f$  leads to the model  $N_i^O \sim \text{Bin}(N_i^T, \theta_f)$ . The data are the  $N_i^O$ 's; the  $N_i^T$ 's are not observed. What do the data tell us about the two parameters  $(\lambda, \theta_f)$ ?

Ignore the Old seedlings for now and just look at 1992 data  $N_{1992}^O = 0$ . Dropping the subscript 1992, the likelihood function is

$$\begin{aligned}
 \ell(\lambda, \theta_f) &= P[N^O = 0 | \lambda, \theta_f] \\
 &= \sum_{n=0}^{\infty} P[N^O = 0, N^T = n | \lambda, \theta_f] \\
 &= \sum_{n=0}^{\infty} P[N^T = n | \lambda] P[N^O = 0 | N^T = n, \theta_f] \\
 &= \sum_{n=0}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!} (1 - \theta_f)^n \\
 &= \sum_{n=0}^{\infty} \frac{e^{-\lambda(1-\theta_f)} (\lambda(1 - \theta_f))^n}{e^{\lambda \theta_f} n!} \\
 &= e^{-\lambda \theta_f}
 \end{aligned} \tag{2.6}$$

Figure 2.25a plots  $\log_{10} \ell(\lambda, \theta_f)$ . (We plotted  $\log_{10} \ell$  instead of  $\ell$  for variety.) The contour lines are not circular. To see what that means, focus on the curve  $\log_{10} \ell(\lambda, \theta_f) = -1$  which runs from about  $(\lambda, \theta_f) = (2.5, 1)$  to about  $(\lambda, \theta_f) = (6, .4)$ . Points  $(\lambda, \theta_f)$  along that curve explain the datum  $N^O = 0$  about 1/10 as well as the m.l.e. (The m.l.e. is any pair where either  $\lambda = 0$  or  $\theta_f = 0$ .) Points below and to the left of that curve explain the datum better than 1/10 of the maximum.

The main parameter of ecological interest is  $\lambda$ , the rate at which New seedlings

tend to arrive. The figure shows that values of  $\lambda$  as large as 6 can have reasonably large likelihoods and hence explain the data reasonably well, at least if we believe that  $\theta_f$  might be as small as .4. To investigate further, Figure 2.25b is similar to 2.25a but includes values of  $\lambda$  as large as 1000. It shows that even values of  $\lambda$  as large as 1000 can have reasonably large likelihoods if they're accompanied by sufficiently small values of  $\theta_f$ . In fact, arbitrarily large values of  $\lambda$  coupled with sufficiently small values of  $\theta_f$  can have arbitrarily large likelihoods. So from the data alone, there is no way to rule out extremely large values of  $\lambda$ . Of course extremely large values of  $\lambda$  don't make ecological sense, both in their own right and because extremely small values of  $\theta_f$  are also not sensible. Scientific background information of this type is incorporated into statistical analysis often through Bayesian inference (Section 2.5). But the point here is that  $\lambda$  and  $\theta_f$  are linked, and the data alone does not tell us much about either parameter individually.

Figure 2.25(a) was produced with the following snippet.

```
lam <- seq (0, 6, by=.1)
th <- seq (0, 1, by=.02)
lik <- matrix (NA, length(lam), length(th))
for (i in seq(along=lam))
for (j in seq(along=th))
 lik[i,j] <- exp (-lam[i]*th[j])
contour (lam, th, log10(lik),
 levels=c(0,-.2,-.6,-1,-1.5,-2),
 xlab=expression(lambda),
 ylab=expression(theta[f]), main="(a)")
```

- `log10` computes the base 10 logarithm.

Figure 2.25(b) was produced with the following snippet.

```
lam2 <- seq (0, 1000, by=1)
lik2 <- matrix (NA, length(lam2), length(th))
for (i in seq(along=lam2))
for (j in seq(along=th))
 lik2[i,j] <- exp (-lam2[i]*th[j])
contour (lam2, th, log10(lik2), levels=c(0,-1,-2,-3),
 xlab=expression(lambda),
 ylab=expression(theta[f]), main="(b)")
```

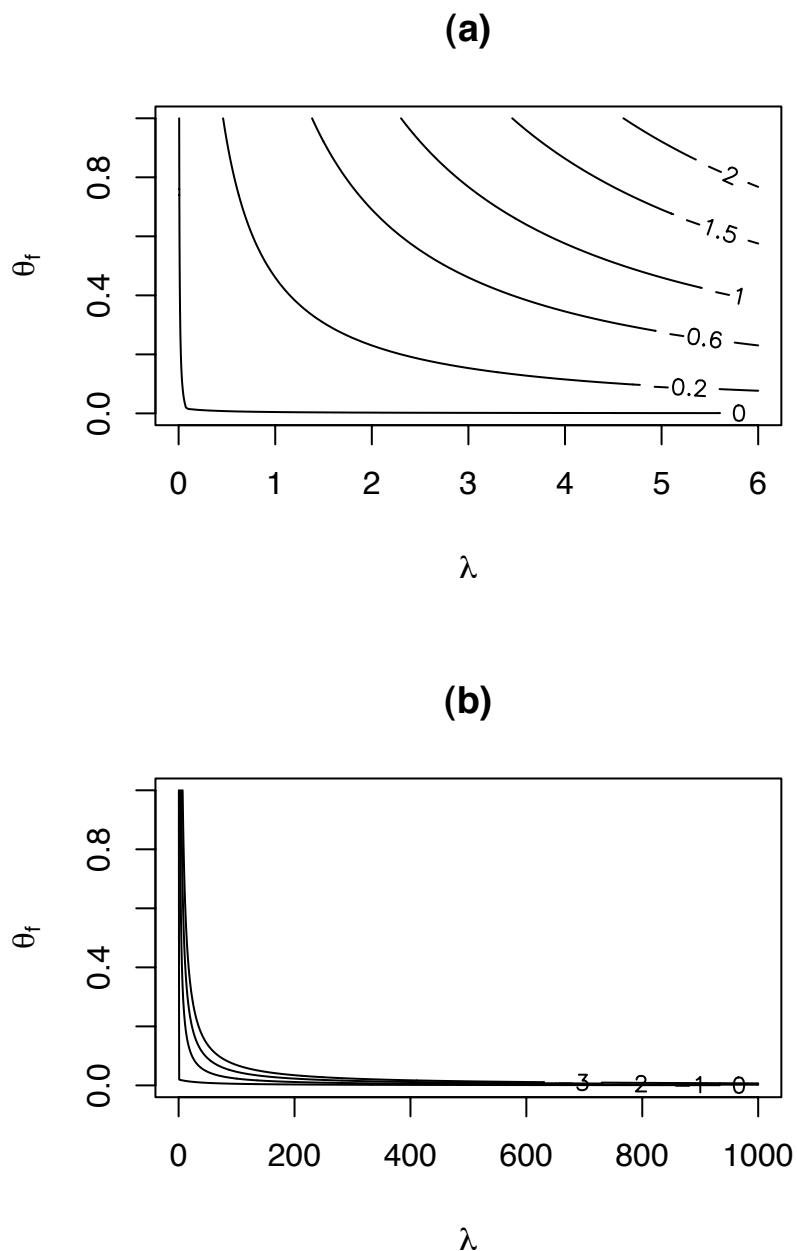


Figure 2.25: Log of the likelihood function for  $(\lambda, \theta_f)$  in Example 2.13

We have now seen two examples (2.11 and 2.12) in which likelihood contours are roughly circular and one (2.13) in which they're not. By far the most common and important case is similar to Example 2.11 because it applies when the Central Limit Theorem applies. That is, there are many instances in which we are trying to make an inference about a parameter  $\theta$  and can invoke the Central Limit Theorem saying that for some statistic  $t$ ,  $t \sim N(\theta, \sigma_t)$  approximately and where we can estimate  $\sigma_t$ . In these cases we can, if necessary, ignore any other parameters in the problem and make an inference about  $\theta$  based on  $\ell_M(\theta)$ .

## 2.4 Estimation

Sometimes the purpose of a statistical analysis is to compute a single best guess at a parameter  $\theta$ . An informed guess at the value of  $\theta$  is called an *estimate* and denoted  $\hat{\theta}$ . One way to estimate  $\theta$  is to find  $\hat{\theta} \equiv \operatorname{argmax}_\theta \ell(\theta)$ , the value of  $\theta$  for which  $\ell(\theta)$  is largest and hence the value of  $\theta$  that best explains the data. That's the subject of Section 2.4.1.

### 2.4.1 The Maximum Likelihood Estimate

In many statistics problems there is a unique value of  $\theta$  that maximizes  $\ell(\theta)$ . This value is called the *maximum likelihood estimate*, or m.l.e. of  $\theta$  and denoted  $\hat{\theta}$ .

$$\hat{\theta} \equiv \operatorname{argmax}_\theta p(y | \theta) = \operatorname{argmax}_\theta \ell(\theta).$$

For instance, in Example 2.8 and Figure 2.20  $\theta$  was the rate of cancer occurrence and we calculated  $\ell(\theta)$  based on  $y = 8$  cancers in 145 people. Figure 2.20 suggests that the m.l.e. is about  $\hat{\theta} \approx .05$ .

When  $\ell(\theta)$  is differentiable, the m.l.e. can be found by differentiating and equating to zero. In Example 2.8 the likelihood was  $\ell(\theta) \propto \theta^8(1 - \theta)^{137}$ . The derivative is

$$\begin{aligned} \frac{d\ell(\theta)}{d\theta} &\propto 8\theta^7(1 - \theta)^{137} - 137\theta^8(1 - \theta)^{136} \\ &= \theta^7(1 - \theta)^{136} [8(1 - \theta) - 137\theta] \end{aligned} \tag{2.7}$$

Equating to 0 yields

$$\begin{aligned} 0 &= 8(1 - \theta) - 137\theta \\ 145\theta &= 8 \\ \theta &= 8/145 \approx .055 \end{aligned}$$

So  $\hat{\theta} \approx .055$  is the m.l.e. Of course if the mode is flat, there are multiple modes, the maximum occurs at an endpoint, or  $\ell$  is not differentiable, then more care is needed.

Equation 2.7 shows more generally the m.l.e. for Binomial data. Simply replace 137 with  $n - y$  and 8 with  $y$  to get  $\hat{\theta} = y/n$ . In the Exercises you will be asked to find the m.l.e. for data from other types of distributions.

There is a trick that is often useful for finding m.l.e.'s. Because  $\log$  is a monotone function,  $\operatorname{argmax} \ell(\theta) = \operatorname{argmax} \log(\ell(\theta))$ , so the m.l.e. can be found by maximizing  $\log \ell$ . For i.i.d. data,  $\ell(\theta) = \prod p(y_i | \theta)$ ,  $\log \ell(\theta) = \sum \log p(y_i | \theta)$ , and it is often easier to differentiate the sum than the product. For the Slater example the math would look like this:

$$\begin{aligned}\log \ell(\theta) &= 8 \log \theta + 137 \log(1 - \theta) \\ \frac{d \log \ell(\theta)}{d\theta} &= \frac{8}{\theta} - \frac{137}{1 - \theta} \\ \frac{137}{1 - \theta} &= \frac{8}{\theta} \\ 137\theta &= 8 - 8\theta \\ \theta &= \frac{8}{145}.\end{aligned}$$

Equation 2.7 shows that if  $y_1, \dots, y_n \sim \text{Bern}(\theta)$  then the m.l.e. of  $\theta$  is

$$\hat{\theta} = n^{-1} \sum y_i = \text{sample mean}$$

The Exercises ask you to show the following.

1. If  $y_1, \dots, y_n \sim N(\mu, \sigma)$  then the m.l.e. of  $\mu$  is

$$\hat{\mu} = n^{-1} \sum y_i = \text{sample mean}$$

2. If  $y_1, \dots, y_n \sim \text{Poi}(\lambda)$  then the m.l.e. of  $\lambda$  is

$$\hat{\lambda} = n^{-1} \sum y_i = \text{sample mean}$$

3. If  $y_1, \dots, y_n \sim \text{Exp}(\lambda)$  then the m.l.e. of  $\lambda$  is

$$\hat{\lambda} = n^{-1} \sum y_i = \text{sample mean}$$

### 2.4.2 Accuracy of Estimation

Finding the m.l.e. is not enough. Statisticians want to quantify the accuracy of  $\hat{\theta}$  as an estimate of  $\theta$ . In other words, we want to know what other values of  $\theta$ , in addition to  $\hat{\theta}$ , have reasonably high likelihood (provide a reasonably good explanation of the data). And what does “reasonable” mean? Section 2.4.2 addresses this question.

As we saw from the reference experiment in section 2.3, the evidence is not very strong against any value of  $\theta$  such that  $\ell(\theta) > \ell(\hat{\theta})/10$ . So when considering estimation accuracy it is useful to think about sets such as

$$\text{LS}_{.1} \equiv \left\{ \theta : \frac{\ell(\theta)}{\ell(\hat{\theta})} \geq .1 \right\}$$

LS stands for *likelihood set*. More generally, for any  $\alpha \in (0, 1)$  we define the likelihood set of level  $\alpha$  to be

$$\text{LS}_\alpha \equiv \left\{ \theta : \frac{\ell(\theta)}{\ell(\hat{\theta})} \geq \alpha \right\}$$

$\text{LS}_\alpha$  is the set of  $\theta$ 's that explain the data reasonably well, and therefore the set of  $\theta$ 's best supported by the data, where the quantification of “reasonable” and “best” are determined by  $\alpha$ . The notion is only approximate and meant as a heuristic reference; in reality there is no strict cutoff between reasonable and unreasonable values of  $\theta$ . Also, there is no uniquely best value of  $\alpha$ . We frequently use  $\alpha \approx .1$  for convenience and custom.

In many problems the likelihood function  $\ell(\theta)$  is continuous and unimodal, i.e. strictly decreasing away from  $\hat{\theta}$ , and goes to 0 as  $\theta \rightarrow \pm\infty$ , as in Figures 2.19 and 2.20. In these cases,  $\theta \approx \hat{\theta} \Rightarrow \ell(\theta) \approx \ell(\hat{\theta})$ . So values of  $\theta$  close to  $\hat{\theta}$  explain the data almost as well as and are about as plausible as  $\hat{\theta}$  and  $\text{LS}_\alpha$  is an interval

$$\text{LS}_\alpha = [\theta_l, \theta_u]$$

where  $\theta_l$  and  $\theta_u$  are the lower and upper endpoints, respectively, of the interval.

In Example 2.9 (Slater School)  $\hat{\theta} = 8/145$ , so we can find  $\ell(\hat{\theta})$  on a calculator, or by using R's built-in function

```
dbinom(8, 145, 8/145)
```

which yields about .144. Then  $\theta_l$  and  $\theta_u$  can be found by trial and error. Since  $\text{dbinom}(8, 145, .023) \approx .013$  and  $\text{dbinom}(8, 145, .105) \approx .015$ , we conclude that  $\text{LS}_{.1} \approx [.023, .105]$  is a rough likelihood interval for  $\theta$ . Review Figure 2.20 to see whether this interval makes sense.

The data in Example 2.9 could pin down  $\theta$  to an interval of width about .08. In general, an experiment will pin down  $\theta$  to an extent determined by the amount of information in the

data. As data accumulates so does information and the ability to determine  $\theta$ . Typically the likelihood function becomes increasingly more peaked as  $n \rightarrow \infty$ , leading to increasingly accurate inference for  $\theta$ . We saw that in Figures 1.6 and 2.19. Example 2.14 illustrates the point further.

**Example 2.14** (Craps, continued)

Example 1.10 introduced a computer simulation to learn the probability  $\theta$  of winning the game of craps. In this example we use that simulation to illustrate the effect of gathering ever increasing amounts of data. We'll start by running the simulation just a few times, and examining the likelihood function  $\ell(\theta)$ . Then we'll add more and more simulations and see what happens to  $\ell(\theta)$ .

The result is in Figure 2.26. The flattest curve is for 3 simulations, and the curves become increasingly peaked for 9, 27, and 81 simulations. After only 3 simulations  $LS_{.1} \approx [.15, .95]$  is quite wide, reflecting the small amount of information. But after 9 simulations  $\ell(\theta)$  has sharpened so that  $LS_{.1} \approx [.05, .55]$  is much smaller. After 27 simulations  $LS_{.1}$  has shrunk further to about [.25, .7], and after 81 it has shrunk even further to about [.38, .61].

Figure 2.26 was produced with the following snippet.

```

n.sim <- c (3, 9, 27, 81)
th <- seq (0, 1, length=200)
lik <- matrix (NA, 200, length(n.sim))

for (i in seq(along=n.sim)) {
 wins <- 0
 for (j in 1:n.sim[i])
 wins <- wins + sim.craps()
 lik[,i] <- dbinom (wins, n.sim[i], th)
 lik[,i] <- lik[,i] / max(lik[,i])
}

matplot (th, lik, type="l", col=1, lty=1:4,
 xlab=expression(theta), ylab="likelihood")

```

In Figure 2.26 the likelihood function looks increasingly like a Normal density as the number of simulations increases. That is no accident; it is the typical behavior in many statistics problems. Section 2.4.3 explains the reason.

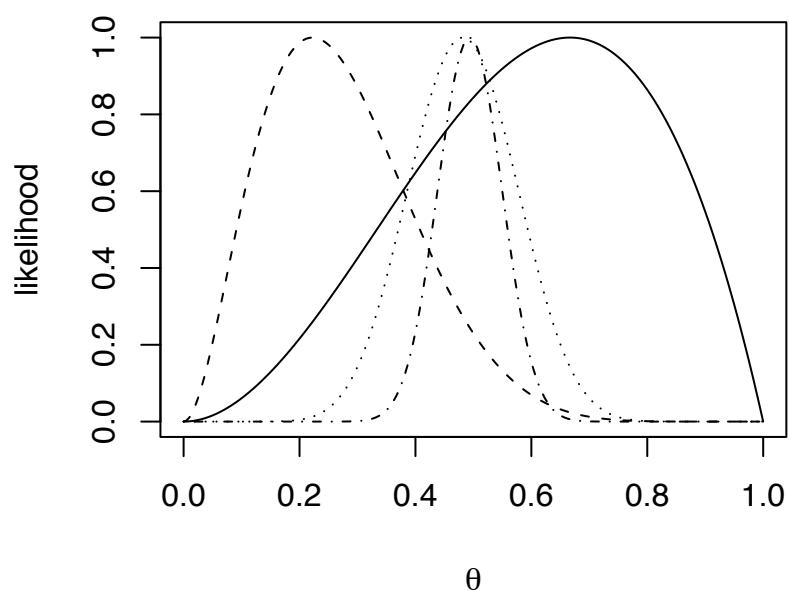


Figure 2.26: Likelihood function for the probability  $\theta$  of winning a game of craps. The four curves are for 3, 9, 27, and 81 simulations.

### 2.4.3 The sampling distribution of an estimator

The estimator  $\hat{\theta}$  is a function of the data  $y_1, \dots, y_n$ . If we repeat the experiment and get new data we also get a new  $\hat{\theta}$ . So  $\hat{\theta}$  is a random variable and has a distribution called the *sampling distribution* of  $\hat{\theta}$  and denoted  $F_{\hat{\theta}}$ . We studied  $F_{\hat{\theta}}$  in Example 1.11 where we used simulation to estimate the probability  $\theta$  of winning a game of craps. For each sample size of  $n = 50, 200, 1000$  we did 1000 simulations. Each simulation yielded a different  $\hat{\theta}$ . Those 1000  $\hat{\theta}$ 's are a random sample of size 1000 from  $F_{\hat{\theta}}$ . Figure 1.19 showed boxplots of the simulations.

Now we examine the sampling distribution of  $\hat{\theta}$  in more detail. There are at least two reasons for doing so. First,  $F_{\hat{\theta}}$  is another way, in addition to likelihood sets, of assessing the accuracy of  $\hat{\theta}$  as an estimator of  $\theta$ . If  $F_{\hat{\theta}}$  is tightly concentrated around  $\theta$  then  $\hat{\theta}$  is highly accurate. Conversely, if  $F_{\hat{\theta}}$  is highly dispersed, or not centered around  $\theta$ , then  $\hat{\theta}$  is an inaccurate estimator. Second, we may want to compare two possible estimators. I.e., if there are two potential estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$ , we can compare  $F_{\hat{\theta}_1}$  and  $F_{\hat{\theta}_2}$  and use the estimator whose sampling distribution is most tightly concentrated around  $\theta$ .

To illustrate, let's suppose we sample  $y_1, \dots, y_n$  from distribution  $F_Y$ , and want to estimate  $\theta \equiv \mathbb{E}[Y]$ . We consider two potential estimators, the sample mean  $\hat{\theta}_1 = (1/n) \sum y_i$  and the sample median  $\hat{\theta}_2$ . To see which estimator is better we do a simulation, as shown in the following snippet. The simulation is done at four different sample sizes,  $n = 4, 16, 64, 256$ , to see whether sample size matters. Here we'll let  $F_Y$  be  $N(0, 1)$ . But the choice between  $\hat{\theta}_1$  and  $\hat{\theta}_2$  might depend on what  $F_Y$  is, so a more thorough investigation would consider other choices of  $F_Y$ .

We do 1000 simulations at each sample size. Figure 2.27 shows the result. The figure suggests that the sampling distributions of both  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are centered at the true value of  $\theta$ . The distribution of  $\hat{\theta}_1$  is slightly less variable than that of  $\hat{\theta}_2$ , but not enough to make much practical difference.

Figure 2.27 was produced by the following snippet.

```
sampsize <- c(4, 16, 64, 256)
n.sim <- 1000

par(mfrow=c(2,2))
for (i in seq(along=sampsize)) {
 y <- matrix(rnorm(n.sim*sampsize[i], 0, 1),
 nrow=sampsize[i], ncol=n.sim)
 that.1 <- apply(y, 2, mean)
 that.2 <- apply(y, 2, median)}
```

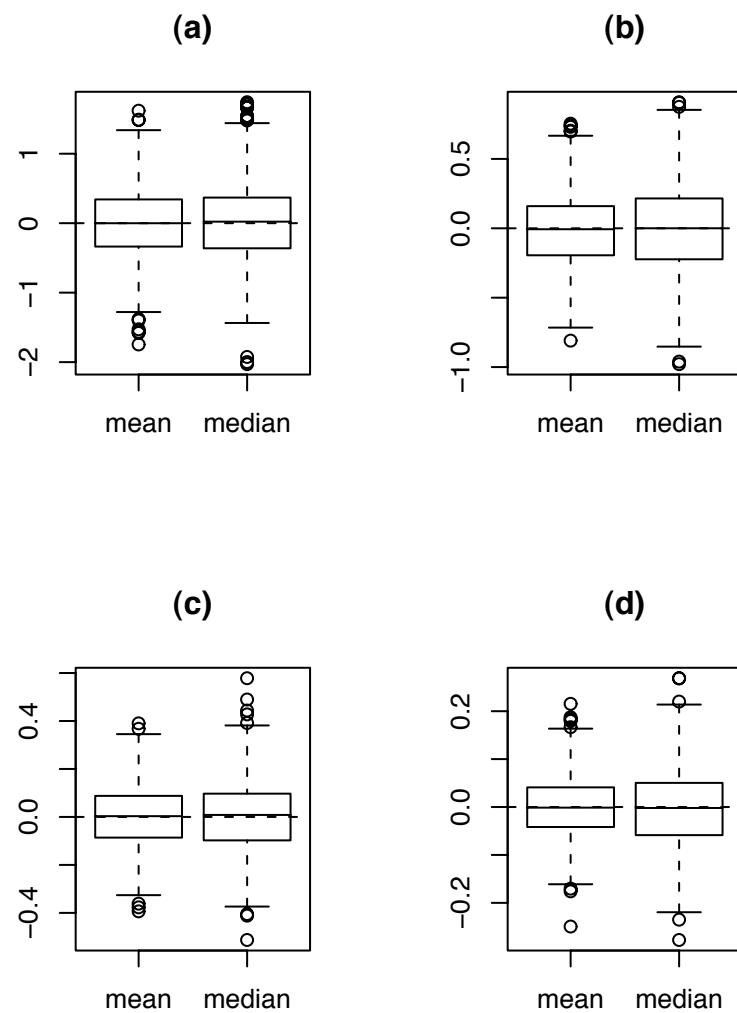


Figure 2.27: Sampling distribution of  $\hat{\theta}_1$ , the sample mean and  $\hat{\theta}_2$ , the sample median. Four different sample sizes. (a):  $n=4$ ; (b):  $n=16$ ; (c):  $n=64$ ; (d):  $n=256$

```

boxplot (that.1, that.2, names=c("mean","median"),
 main=paste("(",letters[i],")",sep=""))
abline (h=0, lty=2)
}

```

For us, comparing  $\hat{\theta}_1$  to  $\hat{\theta}_2$  is only a secondary point of the simulation. The main point is four-fold.

1. An estimator  $\hat{\theta}$  is a random variable and has a distribution.
2.  $F_{\hat{\theta}}$  is a guide to estimation accuracy.
3. Statisticians study conditions under which one estimator is better than another.
4. Simulation is useful.

When the m.l.e. is the sample mean, as it is when  $F_Y$  is a Bernoulli, Normal, Poisson or Exponential distribution, the Central Limit Theorem tells us that in large samples,  $\hat{\theta}$  is approximately Normally distributed. Therefore, in these cases, its distribution can be well described by its mean and SD. Approximately,

$$\hat{\theta} \sim N(\mu_{\hat{\theta}}, \sigma_{\hat{\theta}}).$$

where

$$\begin{aligned}\mu_{\hat{\theta}} &= \mu_Y \\ \sigma_{\hat{\theta}} &= \frac{\sigma_Y}{\sqrt{n}}\end{aligned}\tag{2.8}$$

both of which can be easily estimated from the sample. So we can use the sample to compute a good approximation to the sampling distribution of the m.l.e.

To see that more clearly, let's make 1000 simulations of the m.l.e. in  $n = 5, 10, 25, 100$  Bernoulli trials with  $p = .1$ . We'll make histograms of those simulations and overlay them with kernel density estimates and Normal densities. The parameters of the Normal densities will be estimated from the simulations. Results are shown in Figure 2.28.

Figure 2.28 was produced by the following snippet.

```

sampsize <- c (5, 10, 25, 100)
n.sim <- 1000
p.true <- .1

```

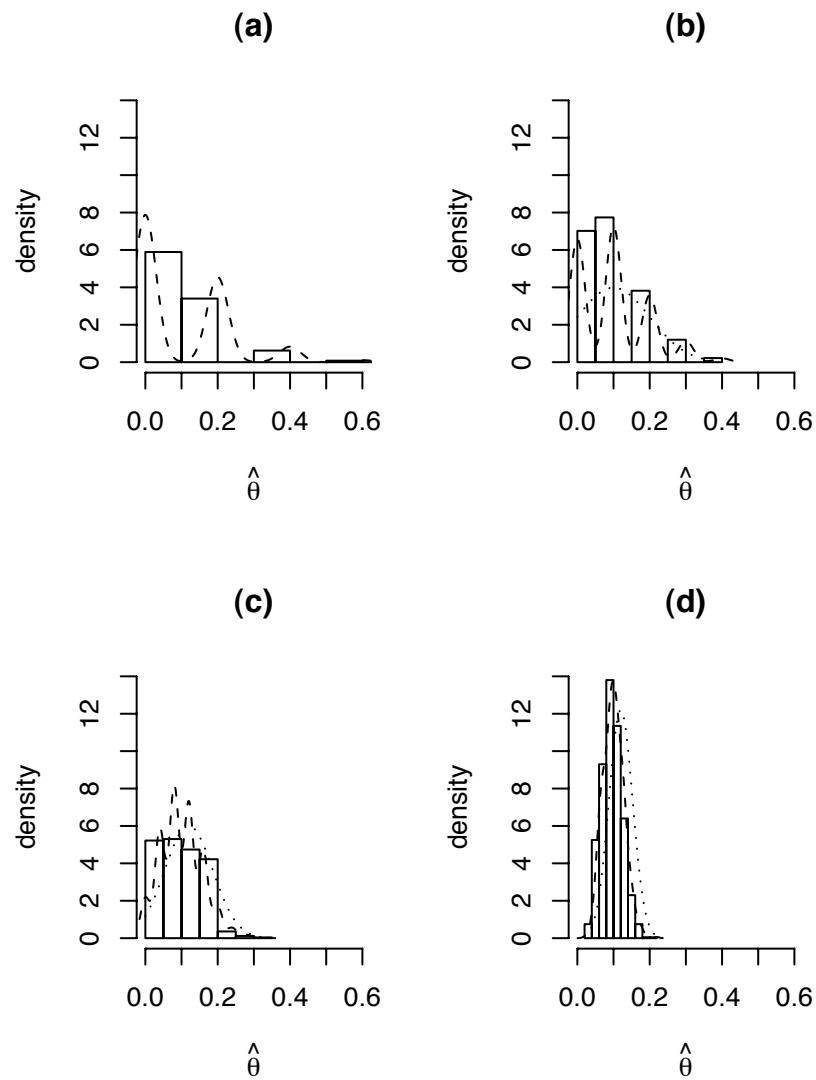


Figure 2.28: Histograms of  $\hat{\theta}$ , the sample mean, for samples from  $\text{Bin}(n, .1)$ . Dashed line: kernel density estimate. Dotted line: Normal approximation. **(a)**:  $n=4$ ; **(b)**:  $n=16$ ; **(c)**:  $n=64$ ; **(d)**:  $n=256$

```

par(mfrow=c(2,2))
for (i in seq(along=sampsizes)) {
 # n.sim Bernoulli samples of sampsizes[i]
 y <- matrix (rbinom (n.sim*sampsizes[i], 1, p.true),
 nrow=n.sim, ncol=sampsizes[i])

 # for each sample, compute the mean
 t.hat <- apply (y, 1, mean)

 # histogram of theta hat
 hist (t.hat, prob=T,
 xlim=c(0,.6), xlab=expression(hat(theta)),
 ylim=c(0,14), ylab="density",
 main=paste ("(", letters[i], ")", sep="")
)

 # kernel density estimate of theta hat
 lines (density (t.hat), lty=2)

 # Normal approximation to density of theta hat,
 # calculated from the first sample
 m <- mean(y[1,])
 sd <- sd(y[1,])/sqrt(sampsizes[i])
 t <- seq (min(t.hat), max(t.hat), length=40)
 lines (t, dnorm (t, m, sd), lty=3)
}

```

Notice that the Normal approximation is not very good for small  $n$ . That's because the underlying distribution  $F_Y$  is highly skewed, nothing at all like a Normal distribution. In fact, R was unable to compute the Normal approximation for  $n = 5$ . But for large  $n$ , the Normal approximation is quite good. That's the Central Limit Theorem kicking in. For any  $n$ , we can use the sample to estimate the parameters in Equation 2.8. For small  $n$ , those parameters don't help us much. But for  $n = 256$ , they tell us a lot about the accuracy of  $\hat{\theta}$ , and the Normal approximation computed from the first sample is a good match to the sampling distribution of  $\hat{\theta}$ .

The SD of an estimator is given a special name. It's called the *standard error* or SE of the estimator because it measures the typical size of estimation errors  $|\hat{\theta} - \theta|$ . When

$\hat{\theta} \sim N(\mu_{\hat{\theta}}, \sigma_{\hat{\theta}})$ , approximately, then  $\sigma_{\hat{\theta}}$  is the SE. For any Normal distribution, about 95% of the mass is within  $\pm 2$  standard deviations of the mean. Therefore,

$$\Pr[|\hat{\theta} - \theta| \leq 2\sigma_{\hat{\theta}}] \approx .95$$

In other words, estimates are accurate to within about two standard errors about 95% of the time, at least when Normal theory applies.

We have now seen two ways of assessing estimation accuracy — through  $\ell(\theta)$  and through  $F_{\hat{\theta}}$ . Often these two apparently different approaches almost coincide. That happens under the following conditions.

1. When  $\hat{\theta} \sim N(\theta, \sigma_{\hat{\theta}})$ , and  $\sigma_{\hat{\theta}} \approx \sigma / \sqrt{n}$ , an approximation often justified by the Central Limit Theorem, then we can estimate  $\theta$  to within about  $\pm 2\sigma_{\hat{\theta}}$ , around 95% of the time. So the interval  $(\hat{\theta} - 2\sigma_{\hat{\theta}}, \hat{\theta} + 2\sigma_{\hat{\theta}})$  is a reasonable estimation interval.
2. When most of the information in the data come from the sample mean, and in other cases when a marginal likelihood argument applies, then  $\ell(\theta) \approx \exp\left(-\frac{1}{2}\left(\frac{\theta - \bar{Y}}{\hat{\sigma}/\sqrt{n}}\right)^2\right)$  (Equation 2.4) and  $LS_1 \approx (\hat{\theta} - 2\sigma_{\hat{\theta}}, \hat{\theta} + 2\sigma_{\hat{\theta}})$ . So the two intervals are about the same.

## 2.5 Bayesian Inference

The essence of Bayesian inference is using probability distributions to describe our state of knowledge of some parameter of interest,  $\theta$ . We construct  $p(\theta)$ , either a pmf or pdf, to reflect our knowledge by making  $p(\theta)$  large for those values of  $\theta$  that seem most likely, and  $p(\theta)$  small for those values of  $\theta$  that seem least likely, according to our state of knowledge. Although  $p(\theta)$  is a probability distribution, it doesn't necessarily mean that  $\theta$  is a random variable. Rather,  $p(\theta)$  encodes our state of knowledge. And different people can have different states of knowledge, hence different probability distributions. For example, suppose you toss a fair coin, look at it, but don't show it to me. The outcome is not random; it has already occurred and you know what it is. But for me, each outcome is equally likely. I would encode my state of knowledge by assigning  $P(H) = P(T) = 1/2$ . You would encode your state of knowledge by assigning either  $P(H) = 1$  or  $P(T) = 1$  according to whether the coin was Heads or Tails. After I see the coin I would update my probabilities to be the same as yours.

For another common example, consider horse racing. When a bettor places a bet at 10 to 1, she is paying \$1 for a ticket that will pay \$10 if the horse wins. Her expected payoff for that bet is  $-\$1 + P[\text{horse wins}] \times \$10$ . For that to be a good deal she must think that  $P[\text{horse wins}] \geq .1$ . Of course other bettors may disagree.

Here are some other examples in which probability distributions must be assessed.

- In deciding whether to fund Head Start, legislators must assess whether the program is likely to be beneficial and, if so, the degree of benefit.
- When investing in the stock market, investors must assess the future probability distributions of stocks they may buy.
- When making business decisions, firms must assess the future probability distributions of outcomes.
- Weather forecasters assess the probability of rain.
- Public policy makers must assess whether the observed increase in average global temperature is anthropogenic and, if so, to what extent.
- Doctors and patients must assess and compare the distribution of outcomes under several alternative treatments.
- At the Slater School, Example 2.8, teachers and administrators must assess their probability distribution for  $\theta$ , the chance that a randomly selected teacher develops invasive cancer.

Information of many types goes into assessing probability distributions. But it is often useful to divide the information into two types: general background knowledge and information specific to the situation at hand. How do those two types of information combine to form an overall distribution for  $\theta$ ? Often we begin by summarizing just the background information as  $p(\theta)$ , the marginal distribution of  $\theta$ . The specific information at hand is data which we can model as  $p(y_1, \dots, y_n | \theta)$ , the conditional distribution of  $y_1, \dots, y_n$  given  $\theta$ . Next, the marginal and conditional densities are combined to give the joint distribution  $p(y_1, \dots, y_n, \theta)$ . Finally, the joint distribution yields  $p(\theta | y_1, \dots, y_n)$  the conditional distribution of  $\theta$  given  $y_1, \dots, y_n$ . And  $p(\theta | y_1, \dots, y_n)$  represents our state of knowledge accounting for both the background information and the data specific to the problem at hand.  $p(\theta)$  is called the *prior* distribution and  $p(\theta | y_1, \dots, y_n)$  is the *posterior* distribution.

A common application is in medical screening exams. Consider a patient being screened for a rare disease, one that affects 1 in 1000 people, say. The disease rate in the population is background information; the patient's response on the screening exam is data specific to this particular patient. Define an indicator variable  $D$  by  $D = 1$  if the patient has the disease and  $D = 0$  if not. Define a second random variable  $T$  by  $T = 1$  if the test result is positive and  $T = 0$  if the test result is negative. And suppose the test that is 95% accurate in the sense that  $P[T = 1 | D = 1] = P[T = 0 | D = 0] = .95$ . Finally, what is the

chance that the patient has the disease given that the test is positive? In other words, what is  $P[D = 1 | T = 1]$ ?

We have the marginal distribution of  $D$  and the conditional distribution of  $T$  given  $D$ . The procedure is to find the joint distribution of  $(D, T)$ , then the conditional distribution of  $D$  given  $T$ . The math is

$$\begin{aligned}
 P[D = 1 | T = 1] &= \frac{P[D = 1 \text{ and } T = 1]}{P[T = 1]} \\
 &= \frac{P[D = 1 \text{ and } T = 1]}{P[T = 1 \text{ and } D = 1] + P[T = 1 \text{ and } D = 0]} \\
 &= \frac{P[D = 1] P[T = 1 | D = 1]}{P[D = 1] P[T = 1 | D = 1] + P[D = 0] P[T = 1 | D = 0]} \quad (2.9) \\
 &= \frac{(.001)(.95)}{(.001)(.95) + (.999)(.05)} \\
 &= \frac{.00095}{.00095 + .04995} \approx .019.
 \end{aligned}$$

That is, a patient who tests positive has only about a 2% chance of having the disease, even though the test is 95% accurate.

Many people find this a surprising result and suspect a mathematical trick. But a quick heuristic check says that out of 1000 people we expect 1 to have the disease, and that person to test positive; we expect 999 people not to have the disease and 5% of those, or about 50, to test positive; so among the 51 people who test positive, only 1, or a little less than 2%, has the disease. The math is correct. This is an example where most people's intuition is at fault and careful attention to mathematics is required in order not to be led astray.

What is the likelihood function in this example? There are two possible values of the parameter, hence only two points in the domain of the likelihood function,  $D = 0$  and  $D = 1$ . So the likelihood function is

$$\ell(0) = .05; \quad \ell(1) = .95$$

Here's another way to look at the medical screening problem, one that highlights the mul-

tiplicative nature of likelihood.

$$\begin{aligned}
 \frac{P[D = 1 | T = 1]}{P[D = 0 | T = 1]} &= \frac{P[D = 1 \text{ and } T = 1]}{P[D = 0 \text{ and } T = 1]} \\
 &= \frac{P[D = 1] P[T = 1 | D = 1]}{P[D = 0] P[T = 1 | D = 0]} \\
 &= \left( \frac{P[D = 1]}{P[D = 0]} \right) \left( \frac{P[T = 1 | D = 1]}{P[T = 1 | D = 0]} \right) \\
 &= \left( \frac{1}{999} \right) \left( \frac{.95}{.05} \right) \\
 &\approx .019
 \end{aligned}$$

The LHS of this equation is the posterior odds of having the disease. The penultimate line shows that the posterior odds is the product of the prior odds and the likelihood ratio. Specifically, to calculate the posterior, we need only the likelihood ratio, not the absolute value of the likelihood function. And likelihood ratios are the means by which prior odds get transformed into posterior odds.

Let's look more carefully at the mathematics in the case where the distributions have densities. Let  $y$  denote the data, even though in practice it might be  $y_1, \dots, y_n$ .

$$\begin{aligned}
 p(\theta | y) &= \frac{p(\theta, y)}{p(y)} \\
 &= \frac{p(\theta, y)}{\int p(\theta, y) d\theta} \\
 &= \frac{p(\theta)p(y | \theta)}{\int p(\theta)p(y | \theta) d\theta}
 \end{aligned} \tag{2.10}$$

Equation 2.10 is the same as Equation 2.9, only in more general terms. Since we are treating the data as given and  $p(\theta | y)$  as a function of  $\theta$ , we are justified in writing

$$p(\theta | y) = \frac{p(\theta)\ell(\theta)}{\int p(\theta)\ell(\theta) d\theta}$$

or

$$p(\theta | y) = \frac{p(\theta)\ell(\theta)}{c}$$

where  $c = \int p(\theta)\ell(\theta) d\theta$  is a constant that does not depend on  $\theta$ . (An integral with respect to  $\theta$  does not depend on  $\theta$ ; after integration it does not contain  $\theta$ .) The effect of the constant  $c$  is to rescale the function in the numerator so that it integrates to 1. I.e.,  $\int p(\theta | y) d\theta = 1$ .

And since  $c$  plays this role, the likelihood function can absorb an arbitrary constant which will ultimately be compensated for by  $c$ . One often sees the expression

$$p(\theta | y) \propto p(\theta) \ell(\theta) \quad (2.11)$$

where the unmentioned constant of proportionality is  $c$ . We can find  $c$  either through Equation 2.10. But it is usually much easier to find  $c$  by using Equation 2.11, then setting  $c = [\int p(\theta) \ell(\theta) d\theta]^{-1}$  and that's the approach we will adopt throughout most of this book.

The next two examples show Bayesian statistics with real data.

**Example 2.15** (Seedlings, continued)

Recall the Seedlings examples (1.4, 1.6, 1.7, 1.9, 2.7, and 2.13) which modelled the number of New seedling arrivals as  $\text{Poi}(\lambda)$ . Prior to the experiment ecologists knew quite a bit about regeneration rates of *acer rubrum* in the vicinity of the experimental quadrats. They estimated that New seedlings would arise at a rate most likely around .5 to 2 seedlings per quadrat per year and less likely either more or less than that. Their knowledge could be encoded in the prior density displayed in Figure 2.29 which is  $p(\lambda) = 4\lambda^2 e^{-2\lambda}$ . (This is the  $\text{Gam}(3, 1/2)$  density; see Section 5.5.) Figure 2.29 also displays the likelihood function  $p(y | \lambda) \propto \lambda^y e^{-\lambda}$  found in Example 1.4 and Figure 1.6. Therefore, according to Equation 2.11, the posterior density is  $p(\lambda | y) \propto \lambda^y e^{-3\lambda}$ . In Section 5.5 we will see that this is the  $\text{Gam}(6, 1/3)$  density, up to a constant of proportionality. Therefore  $c$  in this example must be the constant that appears in the Gamma density:  $c = 1/[5! \times (1/3)^6]$ .

In Figure 2.29 the posterior density is more similar to the prior density than to the likelihood function. But the analysis deals with only a single data point. Let's see what happens as data accumulates. If we have observations  $y_1, \dots, y_n$ , the likelihood function becomes

$$\ell(\lambda) = \prod p(y_i | \lambda) = \prod \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \propto e^{-n\lambda} \lambda^{\sum y_i}$$

To see what this means in practical terms, Figure 2.30 shows (a): the same prior we used in Example 2.15, (b):  $\ell(\lambda)$  for  $n = 1, 4, 16$ , and (c): the posterior for  $n = 1, 4, 16$ , always with  $\bar{y} = 3$ .

1. As  $n$  increases the likelihood function becomes increasingly peaked. That's because as  $n$  increases, the amount of information about  $\lambda$  increases, and we know  $\lambda$  with increasing accuracy. The likelihood function becomes increasingly peaked around the true value of  $\lambda$  and interval estimates become increasingly narrow.
2. As  $n$  increases the posterior density becomes increasingly peaked and becomes increasingly like  $\ell(\lambda)$ . That's because as  $n$  increases, the amount of information in

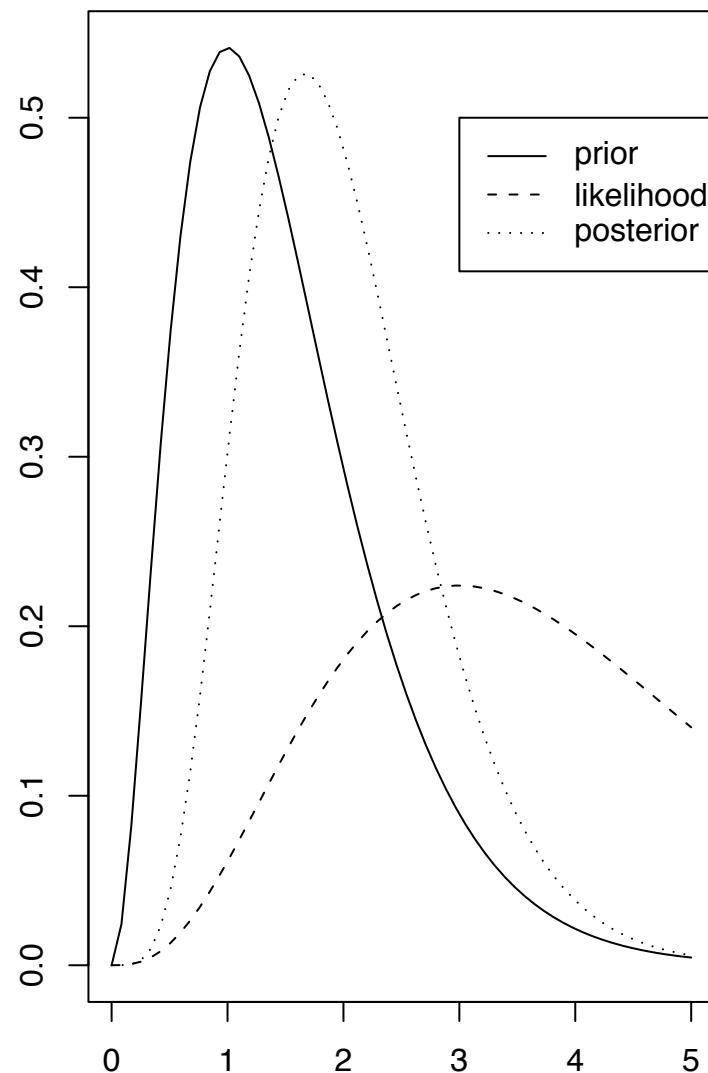


Figure 2.29: Prior, likelihood and posterior densities for  $\lambda$  in the seedlings example after the single observation  $y = 3$

the data increases and the likelihood function becomes increasingly peaked. Meanwhile, the prior density remains as it was. Eventually the data contains much more information than the prior, so the likelihood function becomes much more peaked than the prior and the likelihood dominates. So the posterior, the product of prior and likelihood, looks increasingly like the likelihood.

Another way to look at it is through the loglikelihood  $\ell(\lambda) = c + \log p(\lambda) + \sum_1^n \log p(y_i | \lambda)$ . As  $n \rightarrow \infty$  there is an increasing number of terms in the sum, so the sum eventually becomes much larger and much more important than  $\log p(\lambda)$ .

In practice, of course,  $\bar{y}$  usually doesn't remain constant as  $n$  increases. We saw in Example 1.6 that there were 40 new seedlings in 60 quadrats. With this data the posterior density is

$$p(\theta | y_1, \dots, y_{60}) \propto \lambda^{42} e^{-62\lambda} \quad (2.12)$$

which is the  $\text{Gam}(43, 1/62)$  density. It is pictured in Figure 2.31. Compare to Figure 2.29.

Example 2.16 shows Bayesian statistics at work for the Slater School. See LAVINE [1999] for further analysis.

### **Example 2.16** (Slater School, cont.)

At the time of the analysis reported in BRODEUR [1992] there were two other lines of evidence regarding the effect of power lines on cancer. First, there were some epidemiological studies showing that people who live near power lines or who work as power line repairmen develop cancer at higher rates than the population at large, though only slightly higher. And second, chemists and physicists who calculate the size of magnetic fields induced by power lines (the supposed mechanism for inducing cancer) said that the small amount of energy in the magnetic fields is insufficient to have any appreciable affect on the large biological molecules that are involved in cancer genesis. These two lines of evidence are contradictory. How shall we assess a distribution for  $\theta$ , the probability that a teacher hired at Slater School develops cancer?

Recall from page 135 that Neutra, the state epidemiologist, calculated "4.2 cases of cancer could have been expected to occur" if the cancer rate at Slater were equal to the national average. Therefore, the national average cancer rate for women of the age typical of Slater teachers is  $4.2/145 \approx .03$ . Considering the view of the physicists, our prior distribution should have a fair bit of mass on values of  $\theta \approx .03$ . And considering the epidemiological studies and the likelihood that effects would have been detected before 1992 if they were strong, our prior distribution should put most of its mass below  $\theta \approx .06$ . For the sake of argument let's adopt the prior depicted in Figure 2.32.

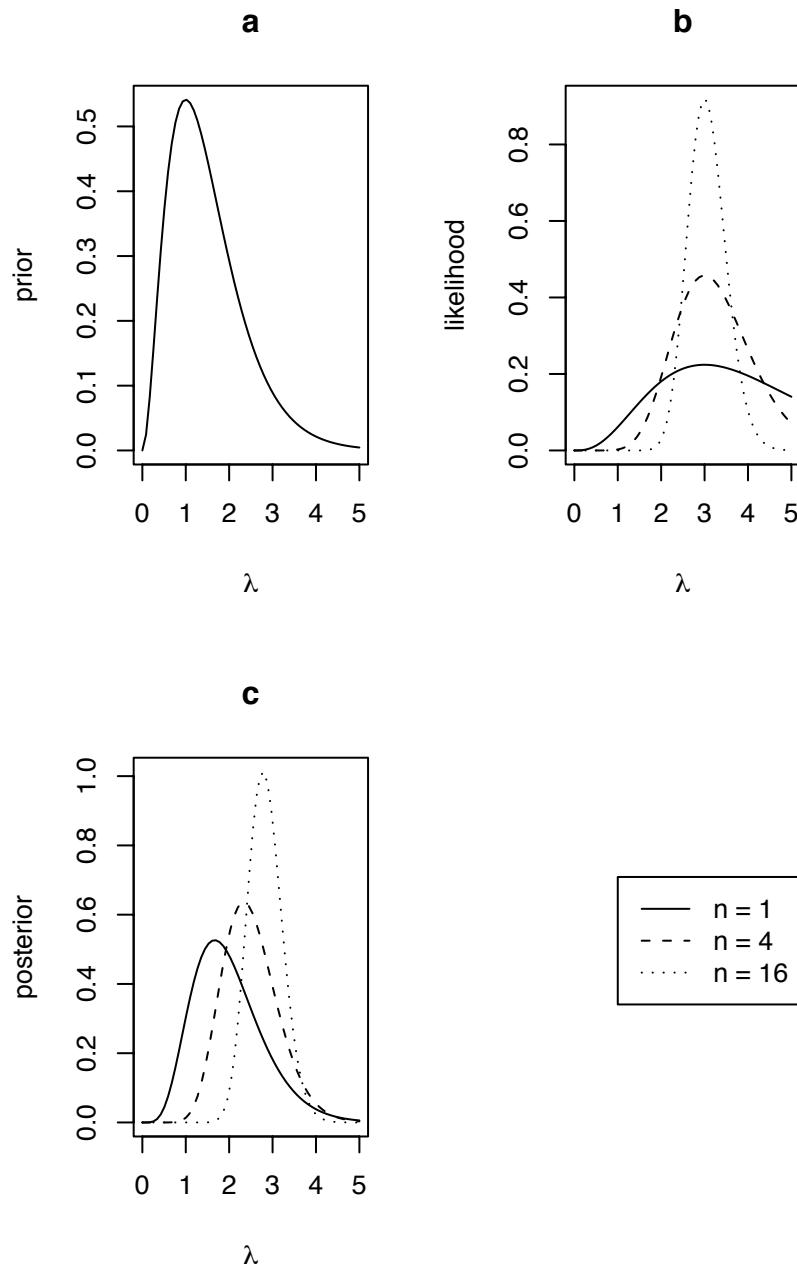


Figure 2.30: **a**:Prior, **b**:likelihood and **c**:posterior densities for  $\lambda$  with  $n = 1, 4, 16$

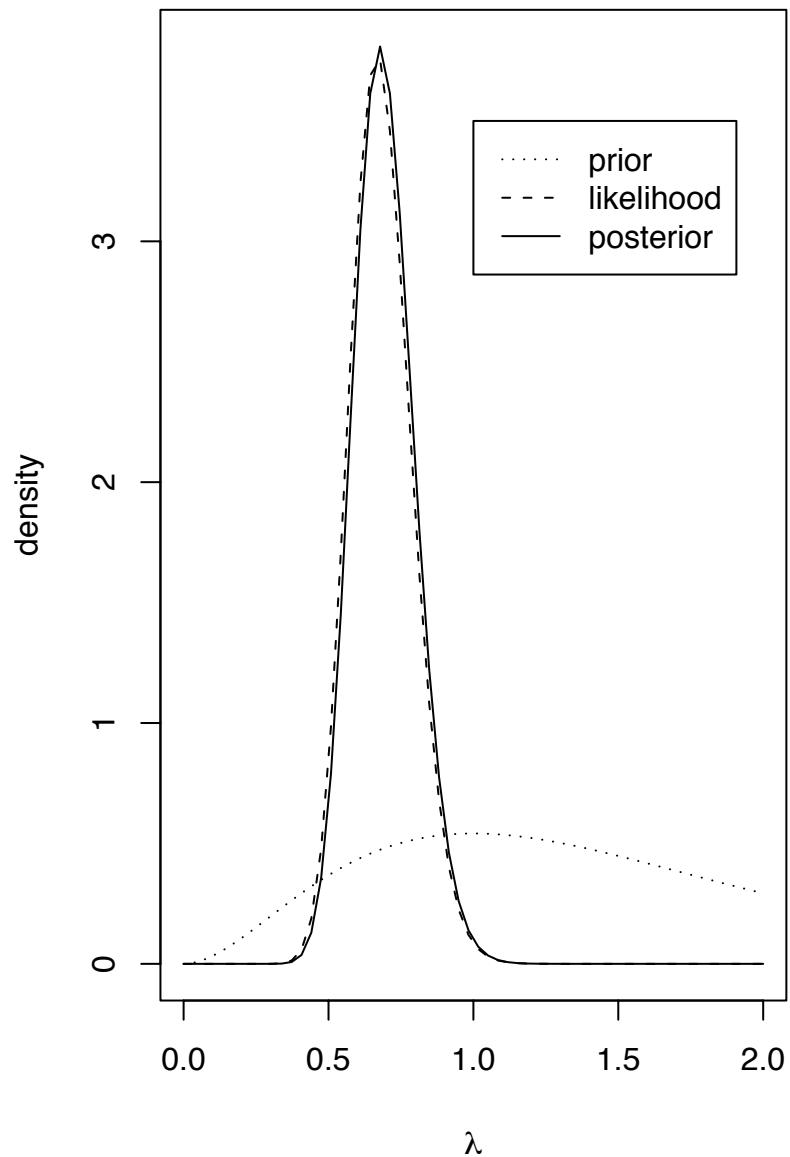


Figure 2.31: Prior, likelihood and posterior densities for  $\lambda$  with  $n = 60$ ,  $\sum y_i = 40$ .

Its formula is

$$p(\theta) = \frac{\Gamma(20)\Gamma(400)}{\Gamma(420)} \theta^{19} (1-\theta)^{399} \quad (2.13)$$

which we will see in Section 5.6 is the  $\text{Be}(20, 400)$  density. The likelihood function is  $\ell(\theta) \propto \theta^8(1-\theta)^{137}$  (Equation 2.3, Figure 2.20). Therefore the posterior density  $p(\theta|y) \propto \theta^{27}(1-\theta)^{536}$  which we will see in Section 5.6 is the  $\text{Be}(28, 537)$  density. Therefore we can easily write down the constant and get the posterior density

$$p(\theta|y) = \frac{\Gamma(28)\Gamma(537)}{\Gamma(565)} \theta^{27}(1-\theta)^{536}$$

which is also pictured in Figure 2.32.

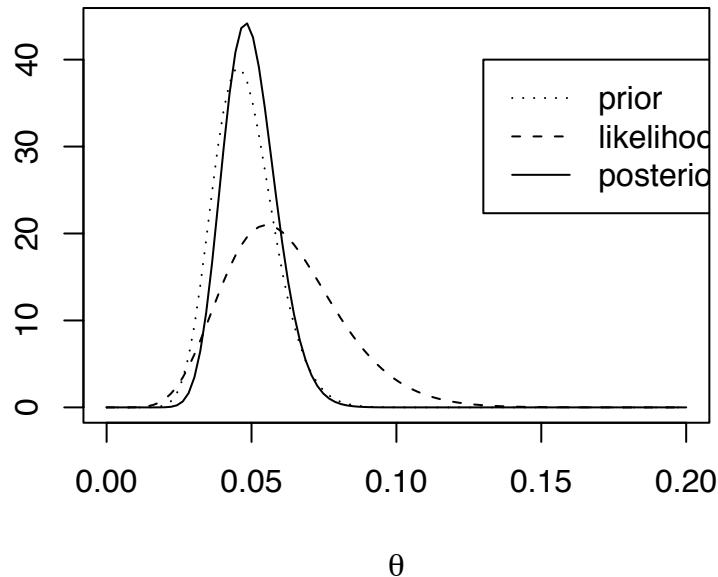


Figure 2.32: Prior, likelihood and posterior density for Slater School

Examples 2.15 and 2.16 have the convenient feature that the prior density had the same form —  $\lambda^a e^{-b\lambda}$  in one case and  $\theta^a(1-\theta)^b$  in the other — as the likelihood function, which made the posterior density and the constant  $c$  particularly easy to calculate. This was not a coincidence. The investigators knew the form of the likelihood function and looked for

a convenient prior of the same form that approximately represented their prior beliefs. This convenience, and whether choosing a prior density for this property is legitimate, are topics which deserve serious thought but which we shall not take up at this point.

## 2.6 Prediction

Sometimes the goal of statistical analysis is to make predictions for future observations. Let  $y_1, \dots, y_n, y_f$  be a sample from  $p(\cdot | \theta)$ . We observe  $y_1, \dots, y_n$  but not  $y_f$ , and want a prediction for  $y_f$ . There are three common forms that predictions take.

**point predictions** A point prediction is a single guess for  $y_f$ . It might be a *predictive mean*, *predictive median*, *predictive mode*, or any other type of point prediction that seems sensible.

**interval predictions** An interval prediction or *predictive interval*, is an interval of plausible values for  $y_f$ . A predictive interval is accompanied by a probability. For example, we might say that “The interval  $(0, 5)$  is a 90% predictive interval for  $y_f$ ” which would mean  $\Pr[y_f \in (0, 5)] = .90$ . In a given problem there are, for two reasons, many predictive intervals. First, there are 90% intervals, 95% intervals, 50% intervals, and so on. And second, there are many predictive intervals with the same probability. For instance, if  $(0, 5)$  is a 90% predictive interval, then it’s possible that  $(-1, 4.5)$  is also a 90% predictive interval.

**predictive distributions** A predictive distribution is a probability distribution for  $y_f$ . From a predictive distribution, different people could compute point predictions or interval predictions, each according to their needs.

In the real world, we don’t know  $\theta$ . After all, that’s why we collected data  $y_1, \dots, y_n$ . But for now, to clarify the types of predictions listed above, let’s pretend that we do know  $\theta$ . Specifically, let’s pretend that we know  $y_1, \dots, y_n, y_f \sim \text{i.i.d. } N(-2, 1)$ .

The main thing to note, since we know  $\theta$  (in this case, the mean and SD of the Normal distribution), is that  $y_1, \dots, y_n$  don’t help us at all. That is, they contain no information about  $y_f$  that is not already contained in the knowledge of  $\theta$ . In other words,  $y_1, \dots, y_n$  and  $y_f$  are conditionally independent given  $\theta$ . In symbols:

$$p(y_f | \theta, y_1, \dots, y_n) = p(y_f | \theta).$$

Therefore, our prediction should be based on the knowledge of  $\theta$  alone, not on any aspect of  $y_1, \dots, y_n$ .

A sensible point prediction for  $y_f$  is  $\hat{y}_f = -2$ , because -2 is the mean, median, and mode of the  $N(-2, 1)$  distribution. Some sensible 90% prediction intervals are  $(-\infty, -0.72)$ ,  $(-3.65, -0.36)$  and  $(-3.28, \infty)$ . We would choose one or the other depending on whether we wanted to describe the lowest values that  $y_f$  might take, a middle set of values, or the highest values. And, of course, the predictive distribution of  $y_f$  is  $N(-2, 1)$ . It completely describes the extent of our knowledge and ability to predict  $y_f$ .

In real problems, though, we don't know  $\theta$ . The simplest way to make a prediction consists of two steps. First use  $y_1, \dots, y_n$  to estimate  $\theta$ , then make predictions based on  $p(y_f | \hat{\theta})$ . Predictions made by this method are called *plug-in* predictions. In the example of the previous paragraph, if  $y_1, \dots, y_n$  yielded  $\hat{\mu} = -2$  and  $\hat{\sigma} = 1$ , then predictions would be exactly as described above.

For an example with discrete data, refer to Examples 1.4 and 1.6 in which  $\lambda$  is the arrival rate of new seedlings. We found  $\hat{\lambda} = 2/3$ . The entire plug-in predictive distribution is displayed in Figure 2.33.  $\hat{y}_f = 0$  is a sensible point prediction. The set  $\{0, 1, 2\}$  is a 97% plug-in prediction interval or prediction set (because  $\text{ppois}(2, 2/3) \approx .97$ ); the set  $\{0, 1, 2, 3\}$  is a 99.5% interval.

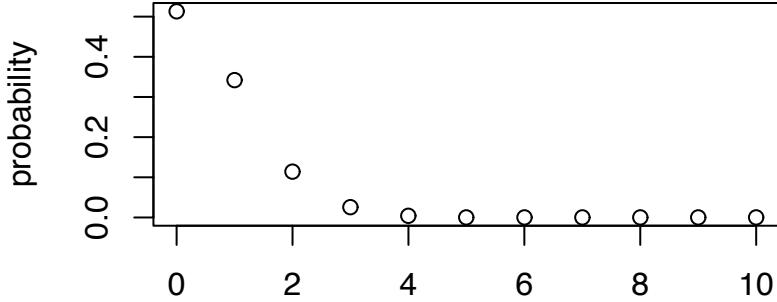


Figure 2.33: Plug-in predictive distribution  $y_f \sim \text{Poi}(\lambda = 2/3)$  for the seedlings example

There are two sources of uncertainty in making predictions. First, because  $y_f$  is random, we couldn't predict it perfectly even if we knew  $\theta$ . And second, we don't know  $\theta$ . In any given problem, either one of the two might be the more important source of uncer-

tainty. The first type of uncertainty can't be eliminated. But in theory, the second type can be reduced by collecting an increasingly large sample  $y_1, \dots, y_n$  so that we know  $\theta$  with ever more accuracy. Eventually, when we know  $\theta$  accurately enough, the second type of uncertainty becomes negligible compared to the first. In that situation, plug-in predictions do capture almost the full extent of predictive uncertainty.

But in many practical problems the second type of uncertainty is too large to be ignored. Plug-in predictive intervals and predictive distributions are too optimistic because they don't account for the uncertainty involved in estimating  $\theta$ . A Bayesian approach to prediction can account for this uncertainty. The prior distribution of  $\theta$  and the conditional distribution of  $y_1, \dots, y_n, y_f$  given  $\theta$  provide the full joint distribution of  $y_1, \dots, y_n, y_f, \theta$ , which in turn provides the conditional distribution of  $y_f$  given  $y_1, \dots, y_n$ . Specifically,

$$\begin{aligned} p(y_f | y_1, \dots, y_n) &= \int p(y_f, \theta | y_1, \dots, y_n) d\theta \\ &= \int p(\theta | y_1, \dots, y_n) p(y_f | \theta, y_1, \dots, y_n) d\theta \\ &= \int p(\theta | y_1, \dots, y_n) p(y_f | \theta) d\theta \end{aligned} \quad (2.14)$$

Equation 2.14 is just the  $y_f$  marginal density derived from the joint density of  $(\theta, y_f)$ , all densities being conditional on the data observed so far. To say it another way, the predictive density  $p(y_f)$  is  $\int p(\theta, y_f) d\theta = \int p(\theta)p(y_f | \theta) d\theta$ , but where  $p(\theta)$  is really the posterior  $p(\theta | y_1, \dots, y_n)$ . The role of  $y_1, \dots, y_n$  is to give us the posterior density of  $\theta$  instead of the prior.

The predictive distribution in Equation 2.14 will be somewhat more dispersed than the plug-in predictive distribution. If we don't know much about  $\theta$  then the posterior will be widely dispersed and Equation 2.14 will be much more dispersed than the plug-in predictive distribution. On the other hand, if we know a lot about  $\theta$  then the posterior distribution will be tight and Equation 2.14 will be only slightly more dispersed than the plug-in predictive distribution.

### **Example 2.17** (Seedlings, cont.)

Refer to Examples 1.4 and 2.15 about  $Y$ , the number of new seedlings emerging each year in a forest quadrat. Our model is  $Y \sim \text{Poi}(\lambda)$ . The prior (page 165) was  $p(\lambda) = 4\lambda^2 e^{-2\lambda}$ . Before collecting any data our predictive distribution would be based

on that prior. For any number  $y$  we could calculate

$$\begin{aligned}
 p_{Y_f}(y) &\equiv \Pr[Y_f = y] = \int p_{Y_f|\Lambda}(y|\lambda)p_{\Lambda}(\lambda)d\lambda \\
 &= \int \frac{\lambda^y e^{-\lambda}}{y!} \frac{2^3}{\Gamma(3)} \lambda^2 e^{-2\lambda} d\lambda \\
 &= \frac{2^3}{y!\Gamma(3)} \int \lambda^{y+2} e^{-3\lambda} d\lambda \\
 &= \frac{2^3 \Gamma(y+3)}{y!\Gamma(3)3^{y+3}} \int \frac{3^{y+3}}{\Gamma(y+3)} \lambda^{y+2} e^{-3\lambda} d\lambda \\
 &= \binom{y+2}{y} \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^y,
 \end{aligned} \tag{2.15}$$

(We will see in Chapter 5 that this is a Negative Binomial distribution.) Thus for example, according to our prior,

$$\begin{aligned}
 \Pr[Y_f = 0] &= \left(\frac{2}{3}\right)^3 = \frac{8}{27} \\
 \Pr[Y_f = 1] &= 3 \left(\frac{2}{3}\right)^3 \frac{1}{3} = \frac{8}{27} \\
 &\text{etc.}
 \end{aligned}$$

Figure 2.34 displays these probabilities.

In the first quadrat we found  $y_1 = 3$  and the posterior distribution (Example 2.15, pg. 165)

$$p(\lambda|y_1 = 3) = \frac{3^6}{5!} \lambda^5 e^{-\lambda/3}.$$

So, by calculations similar to Equation 2.15, the predictive distribution after observing  $y_1 = 3$  is

$$\begin{aligned}
 p_{Y_f|Y_1}(y|y_1 = 3) &= \int p_{Y_f|\Lambda}(y|\lambda)p_{\Lambda|Y_1}(\lambda|y_1 = 3)d\lambda \\
 &= \binom{y+5}{y} \left(\frac{3}{4}\right)^6 \left(\frac{1}{4}\right)^y
 \end{aligned} \tag{2.16}$$

So, for example,

$$\Pr[Y_f = 0 | y_1 = 3] = \left(\frac{3}{4}\right)^6$$

$$\Pr[Y_f = 1 | y_1 = 3] = 6 \left(\frac{3}{4}\right)^6 \frac{1}{4}$$

etc.

Figure 2.34 displays these probabilities.

Finally, when we collected data from 60 quadrats, we found

$$p(\lambda | y_1, \dots, y_{60}) = \frac{62^{43}}{42!} \lambda^{42} e^{-62\lambda} \quad (2.17)$$

Therefore , by calculations similar to Equation 2.15, the predictive distribution is

$$\Pr[Y_f = y | y_1, \dots, y_{60}] = \binom{y+42}{y} \left(\frac{62}{63}\right)^6 \left(\frac{1}{63}\right)^y \quad (2.18)$$

Figure 2.34 displays these probabilities.

A priori, and after only  $n = 1$  observation,  $\lambda$  is not known very precisely; both types of uncertainty are important; and the Bayesian predictive distribution is noticeably different from the plug-in predictive distribution. But after  $n = 60$  observations  $\lambda$  is known fairly well; the second type of uncertainty is negligible; and the Bayesian predictive distribution is very similar to the plug-in predictive distribution.

## 2.7 Hypothesis Testing

Scientific inquiry often takes the form of hypothesis testing. In each instance there are two hypotheses — the *null* hypothesis  $H_0$  and the *alternative* hypothesis  $H_a$ .

### **medicine**

- $H_0$ : the new drug and the old drug are equally effective.
- $H_a$ : the new drug is better than the old.

### **public health**

- $H_0$  exposure to high voltage electric lines is benign.

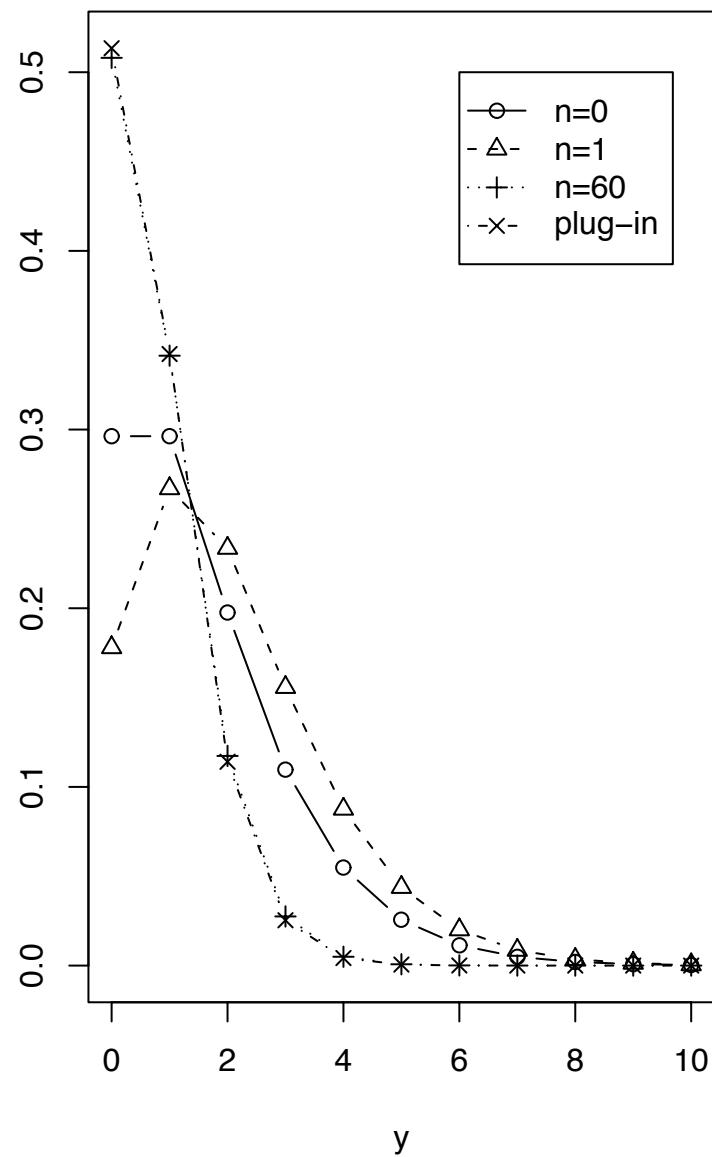


Figure 2.34: Predictive distributions of  $y_f$  in the seedlings example after samples of size  $n = 0, 1, 60$ , and the plug-in predictive

- $H_a$ : exposure to high voltage electric lines promotes cancer.

### public policy

- $H_0$ : Head Start has no effect.
- $H_a$ : Head Start is beneficial.

### astronomy

- $H_0$ : The sun revolves around the Earth.
- $H_a$ : The Earth revolves around the sun.

### physics

- $H_0$ : Newtonian mechanics holds.
- $H_a$ : Relativity holds.

### public trust

- $H_0$ : Winning lottery numbers are random.
- $H_a$ : Winning lottery numbers have patterns.

### ESP

- $H_0$ : There is no ESP.
- $H_a$ : There is ESP.

### ecology

- $H_0$ : Forest fires are irrelevant to forest diversity.
- $H_a$ : Forest fires enhance forest diversity.

By tradition  $H_0$  is the hypothesis that says nothing interesting is going on or the current theory is correct, while  $H_a$  says that something unexpected is happening or our current theories need updating. Often the investigator is hoping to disprove the null hypothesis and to suggest the alternative hypothesis in its place.

It is worth noting that while the two hypotheses are logically exclusive, they are not logically exhaustive. For instance, it's logically possible that forest fires decrease diversity even though that possibility is not included in either hypothesis. So one could write  $H_a$ : *Forest fires decrease forest diversity*, or even  $H_a$ : *Forest fires change forest diversity*.

Which alternative hypothesis is chosen makes little difference for the theory of hypothesis testing, though it might make a large difference to ecologists.

Statisticians have developed several methods called hypothesis tests. We focus on just one for the moment, useful when  $H_0$  is specific. The fundamental idea is to see whether the data are “compatible” with the specific  $H_0$ . If so, then there is no reason to doubt  $H_0$ ; if not, then there is reason to doubt  $H_0$  and possibly to consider  $H_a$  in its stead. The meaning of “compatible” can change from problem to problem but typically there is a four step process.

1. Formulate a scientific null hypothesis and translate it into statistical terms.
2. Choose a low dimensional statistic, say  $w = w(y_1, \dots, y_n)$  such that the distribution of  $w$  is specified under  $H_0$  and likely to be different under  $H_a$ .
3. Calculate, or at least approximate, the distribution of  $w$  under  $H_0$ .
4. Check whether the observed value of  $w$ , calculated from  $y_1, \dots, y_n$ , is compatible with its distribution under  $H_0$ .

How would this work in the examples listed at the beginning of the chapter? What follows is a very brief description of how hypothesis tests might be carried out in some of those examples. To focus on the key elements of hypothesis testing, the descriptions have been kept overly simplistic. In practice, we would have to worry about confounding factors, the difficulties of random sampling, and many other issues.

**public health** Sample a large number of people with high exposure to power lines. For each person, record  $X_i$ , a Bernoulli random variable indicating whether that person has cancer. Model  $X_1, \dots, X_n \sim$  i.i.d.  $\text{Bern}(\theta_1)$ . Repeat for a sample of people with low exposure; getting  $Y_1, \dots, Y_n \sim$  i.i.d.  $\text{Bern}(\theta_2)$ . Estimate  $\theta_1$  and  $\theta_2$ . Let  $w = \hat{\theta}_1 - \hat{\theta}_2$ .  $H_0$  says  $\mathbb{E}[w] = 0$ . Either the Binomial distribution or the Central Limit Theorem tells us the SD’s of  $\hat{\theta}_1$  and  $\hat{\theta}_2$ , and hence the SD of  $w$ . Ask *How many SD’s is w away from its expected value of 0*. If it’s off by many SD’s, more than about 2 or 3, that’s evidence against  $H_0$ .

**public policy** Test a sample children who have been through Head Start. Model their test scores as  $X_1, \dots, X_n \sim$  i.i.d.  $N(\mu_1, \sigma_1)$ . Do the same for children who have not been through Head Start, getting  $Y_1, \dots, Y_n \sim$  i.i.d.  $N(\mu_2, \sigma_2)$ .  $H_0$  says  $\mu_1 = \mu_2$ . Let  $w = \hat{\mu}_1 - \hat{\mu}_2$ . The parameters  $\mu_1, \mu_2, \sigma_1, \sigma_2$  can all be estimated from the data; therefore  $w$  can be calculated and its SD estimated. Ask *How many SD’s is w away from its expected value of 0*. If it’s off by many SD’s, more than about 2 or 3, that’s evidence against  $H_0$ .

**ecology** We could either do an observational study, beginning with one sample of plots that had had frequent forest fires in the past and another sample that had had few fires. Or we could do an experimental study, beginning with a large collection of plots and subjecting half to a regime of regular burning and the other half to a regime of no burning. In either case we would measure and compare species diversity in both sets of plots. If diversity is similar in both groups, there is no reason to doubt  $H_0$ . But if diversity is sufficiently different (Sufficient means *large compared to what is expected by chance under  $H_0$* ) that would be evidence against  $H_0$ .

To illustrate in more detail, let's consider testing a new blood pressure medication. The scientific null hypothesis is that the new medication is not any more effective than the old. We'll consider two ways a study might be conducted and see how to test the hypothesis both ways.

METHOD 1 A large number of patients are enrolled in a study and their blood pressures are measured. Half are randomly chosen to receive the new medication (treatment); half receive the old (control). After a prespecified amount of time, their blood pressure is remeasured. Let  $Y_{C,i}$  be the change in blood pressure from the beginning to the end of the experiment for the  $i$ 'th control patient and  $Y_{T,i}$  be the change in blood pressure from the beginning to the end of the experiment for the  $i$ 'th treatment patient. The model is

$$\begin{aligned} Y_{C,1}, \dots, Y_{C,n} &\sim \text{i.i.d. } f_C; & \mathbb{E}[Y_{C,i}] &= \mu_C; & \text{Var}(Y_{C,i}) &= \sigma_C^2 \\ Y_{T,1}, \dots, Y_{T,n} &\sim \text{i.i.d. } f_T; & \mathbb{E}[Y_{T,i}] &= \mu_T; & \text{Var}(Y_{T,i}) &= \sigma_T^2 \end{aligned}$$

for some unknown means  $\mu_C$  and  $\mu_T$  and variances  $\sigma_C^2$  and  $\sigma_T^2$ . The translation of the hypotheses into statistical terms is

$$\begin{aligned} H_0 : \mu_T &= \mu_C \\ H_a : \mu_T &\neq \mu_C \end{aligned}$$

Because we're testing a difference in means, let  $w = \bar{Y}_T - \bar{Y}_C$ . If the sample size  $n$  is reasonably large, then the Central Limit Theorem says approximately  $w \sim N(0, \sigma_w^2)$  under  $H_0$  with  $\sigma_w^2 = (\sigma_T^2 + \sigma_C^2)/n$ . The mean of 0 comes from  $H_0$ . The variance  $\sigma_w^2$  comes from adding variances of independent random variables.  $\sigma_T^2$  and  $\sigma_C^2$  and therefore  $\sigma_w^2$  can be estimated from the data. So we can calculate  $w$  from the data and see whether it is within about 2 or 3 SD's of where  $H_0$  says it should be. If it isn't, that's evidence against  $H_0$ .

METHOD 2 A large number of patients are enrolled in a study and their blood pressure is measured. They are matched together in pairs according to relevant medical characteristics. The two patients in a pair are chosen to be as similar to each other as possible. In each pair, one patient is randomly chosen to receive the new medication (treatment); the

other receives the old (control). After a prespecified amount of time their blood pressures are measured again. Let  $Y_{T,i}$  and  $Y_{C,i}$  be the change in blood pressure for the  $i$ 'th treatment and  $i$ 'th control patients. The researcher records

$$X_i = \begin{cases} 1 & \text{if } Y_{T,i} > Y_{C,i} \\ 0 & \text{otherwise} \end{cases}$$

The model is

$$X_1, \dots, X_n \sim \text{i.i.d. Bern}(p)$$

for some unknown probability  $p$ . The translation of the hypotheses into statistical terms is

$$\begin{aligned} H_0 &: p = .5 \\ H_a &: p \neq .5 \end{aligned}$$

Let  $w = \sum X_i$ . Under  $H_0$ ,  $w \sim \text{Bin}(n, .5)$ . To test  $H_0$  we plot the  $\text{Bin}(n, .5)$  distribution and see where  $w$  falls on the plot. Figure 2.35 shows the plot for  $n = 100$ . If  $w$  turned out to be between about 40 and 60, then there would be little reason to doubt  $H_0$ . But on the other hand, if  $w$  turned out to be less than 40 or greater than 60, then we would begin to doubt. The larger  $|w - 50|$ , the greater the cause for doubt.

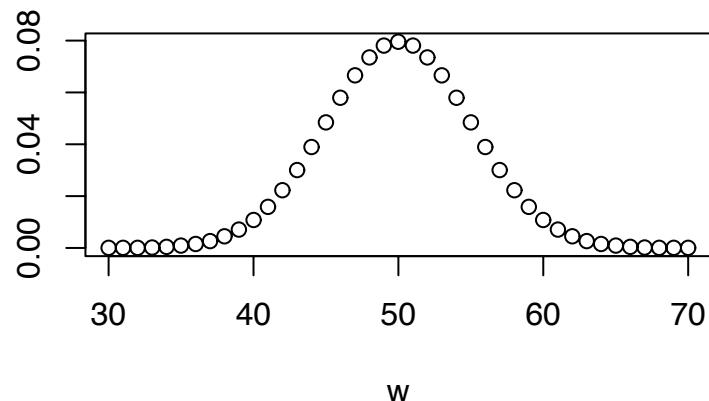


Figure 2.35: pdf of the  $\text{Bin}(100, .5)$  distribution

This blood pressure example exhibits a feature common to many hypothesis tests. First, we're testing a difference in means. I.e.,  $H_0$  and  $H_a$  disagree about a mean, in this case the mean change in blood pressure from the beginning to the end of the experiment. So we take  $w$  to be the difference in sample means. Second, since the experiment is run on a large number of people, the Central Limit Theorem says that  $w$  will be approximately Normally distributed. Third, we can calculate or estimate the mean  $\mu_0$  and SD  $\sigma_0$  under  $H_0$ . So fourth, we can compare the value of  $w$  from the data to what  $H_0$  says its distribution should be.

In Method 1 above, that's just what we did. In Method 2 above, we didn't use the Normal approximation; we used the Binomial distribution. But we could have used the approximation. From facts about the Binomial distribution we know  $\mu_0 = n/2$  and  $\sigma_0 = \sqrt{n}/2$  under  $H_0$ . For  $n = 100$ , Figure 2.36 compares the exact Binomial distribution to the Normal approximation.

In general, when the Normal approximation is valid, we compare  $w$  to the  $N(\mu_0, \sigma_0)$  density, where  $\mu_0$  is calculated according to  $H_0$  and  $\sigma_0$  is either calculated according to  $H_0$  or estimated from the data. If  $t \equiv |w - \mu_0|/\sigma_0$  is bigger than about 2 or 3, that's evidence against  $H_0$ .

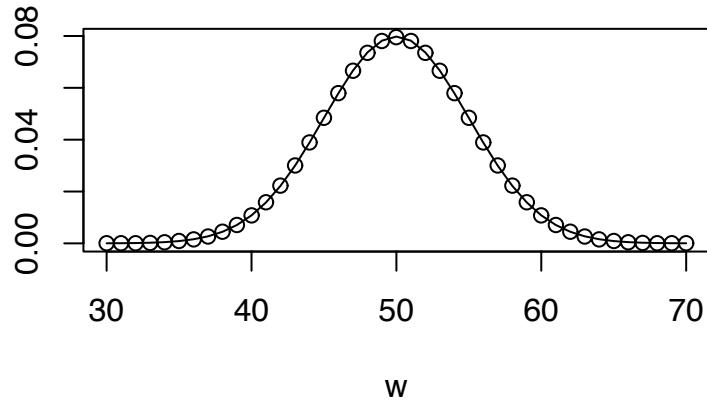


Figure 2.36: pdfs of the  $\text{Bin}(100, .5)$  (dots) and  $N(50, 5)$  (line) distributions

The following example shows hypothesis testing at work.

**Example 2.18** (Tooth Growth, continued)

This continues Example 2.1 (pg. 99). Let's concentrate on a particular dosage, say dose = 0.5, and test the null hypothesis that, on average, the delivery method (supp) makes no difference to tooth growth, as opposed to the alternative that it does make a difference. Those are the scientific hypotheses. The data for testing the hypothesis are  $x_1, \dots, x_{10}$ , the 10 recordings of growth when supp = VC and  $y_1, \dots, y_{10}$ , the 10 recordings of growth when supp = OJ. The  $x_i$ 's are 10 independent draws from one distribution; the  $y_i$ 's are 10 independent draws from another:

$$\begin{aligned}x_1, \dots, x_{10} &\sim \text{i.i.d. } f_{\text{VC}} \\y_1, \dots, y_{10} &\sim \text{i.i.d. } f_{\text{OJ}}\end{aligned}$$

Define the two means to be  $\mu_{\text{VC}} \equiv \mathbb{E}[x_i]$  and  $\mu_{\text{OJ}} \equiv \mathbb{E}[y_i]$ . The scientific hypothesis and its alternative, translated into statistical terms become

$$\begin{aligned}H_0 : \mu_{\text{VC}} &= \mu_{\text{OJ}} \\H_a : \mu_{\text{VC}} &\neq \mu_{\text{OJ}}\end{aligned}$$

Those are the hypotheses in statistical terms.

Because we're testing a difference in means, we choose our one dimensional summary statistic to be  $w = |\bar{x} - \bar{y}|$ . Small values of  $w$  support  $H_0$ ; large values support  $H_a$ . But how small is small; how large is large? The Central Limit Theorem says

$$\begin{aligned}\bar{x} &\sim N\left(\mu_{\text{VC}}, \frac{\sigma_{\text{VC}}}{\sqrt{n}}\right) \\\bar{y} &\sim N\left(\mu_{\text{OJ}}, \frac{\sigma_{\text{OJ}}}{\sqrt{n}}\right)\end{aligned}$$

approximately, so that under  $H_0$ ,

$$w \sim N\left(0, \sqrt{\frac{\sigma_{\text{VC}}^2 + \sigma_{\text{OJ}}^2}{n}}\right),$$

approximately. The statistic  $w$  can be calculated, its SD estimated, and its approximate density plotted as in Figure 2.37. We can see from the Figure, or from the fact that  $t/\sigma_t \approx 3.2$  that the observed value of  $t$  is moderately far from its expected value under  $H_0$ . The data provide moderately strong evidence against  $H_0$ .

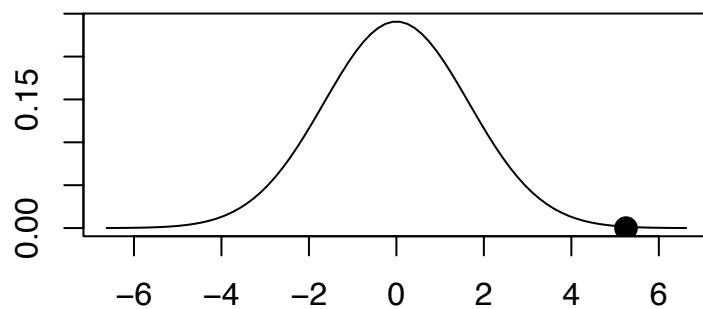


Figure 2.37: Approximate density of summary statistic  $t$ . The black dot is the value of  $t$  observed in the data.

Figure 2.37 was produced with the following R code.

```

x <- ToothGrowth$len[ToothGrowth$supp=="VC"
 & ToothGrowth$dose==0.5]
y <- ToothGrowth$len[ToothGrowth$supp=="OJ"
 & ToothGrowth$dose==0.5]
t <- abs (mean(x) - mean(y))
sd <- sqrt ((var(x) + var(y))/length(x))
tvals <- seq (-4*sd, 4*sd, len=80)
plot (tvals, dnorm(tvals,0,sd), type="l",
 xlab="", ylab="", main="")
points (t, 0, pch=16, cex=1.5)

```

The `points(...)` adds the observed value of  $t$  to the plot.

In the next example it is difficult to estimate the distribution of  $w$  under  $H_0$ ; so we use simulation to work it out.

**Example 2.19** (Baboons)

Because baboons are promiscuous, when a baby is born it is not obvious, at least to humans, who the father is. But do the baboons themselves know who the father is? BUCHAN ET AL. [2003] report a study of baboon behavior that attempts to answer that question. For more information see [HTTP://WWW.PRINCETON.EDU/~BABOON](http://WWW.PRINCETON.EDU/~BABOON). Baboons live in social groups comprised of several adult males, several adult females, and juveniles. Researchers followed several groups of baboons periodically over a period of several years to learn about baboon behavior. The particular aspect of behavior that concerns us here is that adult males sometimes come to the aid of juveniles. If adult males know which juveniles are their own children, then it's at least possible that they tend to aid their own children more than other juveniles. The data set baboons (available on the web site)<sup>1</sup> contains data on all the recorded instances of adult males helping juveniles. The first four lines of the file look like this.

| Recip | Father | Maleally | Dadpresent | Group |
|-------|--------|----------|------------|-------|
| ABB   | EDW    | EDW      | Y          | OMO   |
| ABB   | EDW    | EDW      | Y          | OMO   |
| ABB   | EDW    | EDW      | Y          | OMO   |
| ABB   | EDW    | POW      | Y          | OMO   |

1. Recip identifies the juvenile who received help. In the four lines shown here, it is always ABB.
2. Father identifies the father of the juvenile. Researchers know the father through DNA testing of fecal samples. In the four lines shown here, it is always EDW.
3. Maleally identifies the adult male who helped the juvenile. In the fourth line we see that POW aided ABB who is not his own child.
4. Dadpresent tells whether the father was present in the group when the juvenile was aided. In this data set it is always Y.
5. Group identifies the social group in which the incident occurred. In the four lines shown here, it is always OMO.

Let  $w$  be the number of cases in which a father helps his own child. The snippet

---

<sup>1</sup>We have slightly modified the data to avoid some irrelevant complications.

```
dim (baboons)
sum (baboons$Father == baboons$Maleally)
```

reveals that there are  $n = 147$  cases in the data set, and that  $w = 87$  are cases in which a father helps his own child. The next step is to work out the distribution of  $w$  under  $H_0$ : *adult male baboons do not know which juveniles are their children*.

Let's examine one group more closely, say the OMO group. Typing

```
baboons[baboons$Group == "OMO",]
```

displays the relevant records. There are 13 of them. EDW was the father in 9, POW was the father in 4. EDW provided the help in 9, POW in 4. The father was the ally in 9 cases; in 4 he was not.  $H_0$  implies that EDW and POW would distribute their help randomly among the 13 cases. If  $H_0$  is true, i.e., if EDW distributes his 9 helps and POW distributes his 4 helps randomly among the 13 cases, what would be the distribution of  $W$ , the number of times a father helps his own child? We can answer that question by a simulation in R. (We could also answer it by doing some math or by knowing the hypergeometric distribution, but that's not covered in this text.)

```
dads <- baboons$Father [baboons$Group == "OMO"]
ally <- baboons$Maleally [baboons$Group == "OMO"]
N.sim <- 1000
w <- rep (NA, N.sim)
for (i in 1:N.sim) {
 perm <- sample (dads)
 w[i] <- sum (perm == ally)
}
hist(w)
table(w)
```

Try out the simulation for yourself. It shows that the observed number in the data,  $w = 9$ , is not so unusual under  $H_0$ .

What about the other social groups? If we find out how many there are, we can do a similar simulation for each. Let's write an R function to help.

```
g.sim <- function (group, N.sim) {
 dads <- baboons$Father [baboons$Group == group]
```

```

ally <- baboons$Maleally [baboons$Group == group]
w <- rep (NA, N.sim)
for (i in 1:N.sim) {
 perm <- sample (dads)
 w[i] <- sum (perm == ally)
}
return(w)
}

```

Figure 2.38 shows histograms of `g.sim` for each group, along with a dot showing the observed value of `w` in the data set. For some of the groups the observed value of `w`, though a bit on the high side, might be considered consistent with  $H_0$ . For others, the observed value of `w` falls outside the range of what might be reasonably expected by chance. In a case like this, where some of the evidence is strongly against  $H_0$  and some is only weakly against  $H_0$ , an inexperienced statistician might believe the overall case against  $H_0$  is not very strong. But that's not true. In fact, every one of the groups contributes a little evidence against  $H_0$ , and the total evidence against  $H_0$  is very strong. To see this, we can combine the separate simulations into one. The following snippet of code does this. Each male's help is randomly reassigned to a juvenile within his group. The number of times when a father helps his own child is summed over the different groups. Simulated numbers are shown in the histogram in Figure 2.39. The dot in the figure is at 84, the actual number of instances in the full data set. Figure 2.39 suggests that it is almost impossible that the 84 instances arose by chance, as  $H_0$  would suggest. We should reject  $H_0$  and reach the conclusion that (a) adult male baboons do know who their own children are, and (b) they give help preferentially to their own children.

Figure 2.38 was produced with the following snippet.

```

groups <- unique (baboons$Group)
n.groups <- length(groups)
par (mfrow=c(3,2))
for (i in 1:n.groups) {
 good <- baboons$Group == groups[i]
 w.obs <- sum (baboons$Father[good]
 == baboons$Maleally[good])
 w.sim <- g.sim (groups[i], N.sim)
 hist (w.sim, xlab="w", ylab="", main=groups[i],

```

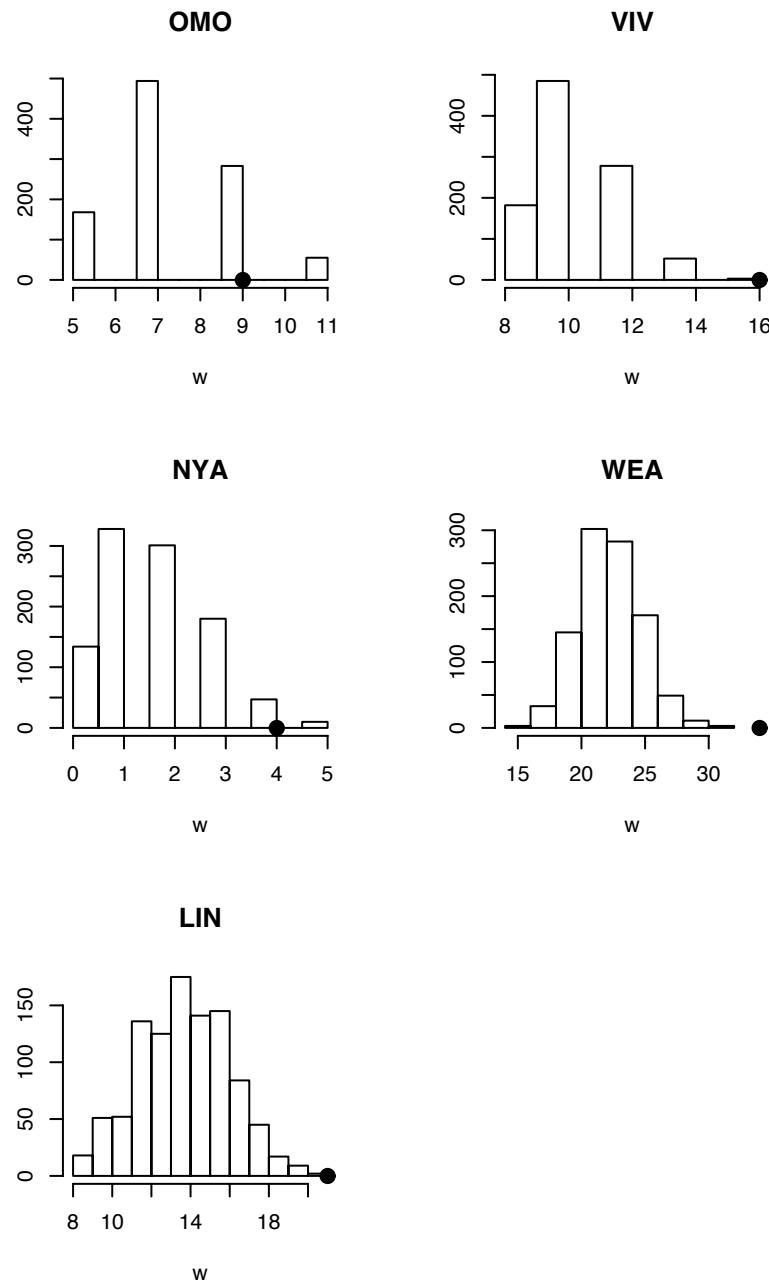


Figure 2.38: Number of times baboon father helps own child in Example 2.19. Histograms are simulated according to  $H_0$ . Dots are observed data.

```

 xlim=range(c(w.obs,w.sim)))
points (w.obs, 0, pch=16, cex=1.5)
print (w.obs)
}

```

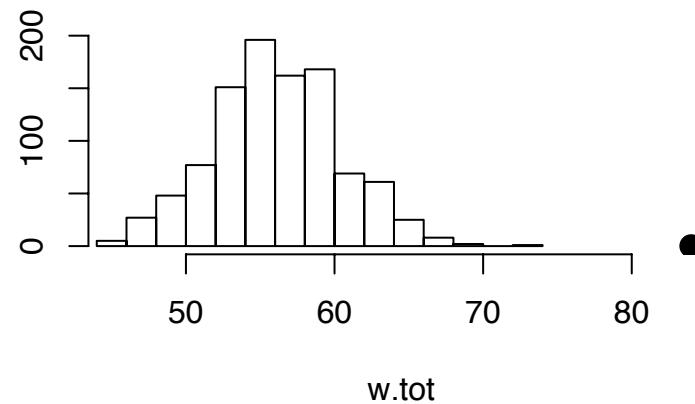


Figure 2.39: Histogram of simulated values of `w.tot`. The dot is the value observed in the baboon data set.

Figure 2.39 was produced with the following snippet.

```

w.obs <- rep (NA, n.groups)
w.sim <- matrix (NA, n.groups, N.sim)
for (i in 1:n.groups) {
 good <- baboons$Group == groups[i]
 w.obs[i] <- sum (baboons$Father[good]
 == baboons$Maleally[good])
 w.sim[i,] <- g.sim (groups[i], N.sim)
}

```

```
w.obs.tot <- sum (w.obs)
w.sim.tot <- apply (w.sim, 2, sum)
hist (w.sim.tot, xlab="w.tot", ylab="",
 xlim=range(c(w.obs.tot,w.sim.tot)))
points (w.obs.tot, 0, pch=16, cex=1.5)
print (w.obs.tot)
```

## 2.8 Exercises

1. (a) Justify Equation 2.1 on page 105.  
(b) Show that the function  $g(x)$  defined just after Equation 2.1 is a probability density. I.e., show that it integrates to 1.
2. This exercise uses the `ToothGrowth` data from Examples 2.1 and 2.18.
  - (a) Estimate the effect of delivery mode for doses 1.0 and 2.0. Does it seem that delivery mode has a different effect at different doses?
  - (b) Does it seem as though delivery mode changes the effect of dose?
  - (c) For each delivery mode, make a set of three boxplots to compare the three doses.
3. This exercise uses data from 272 eruptions of the Old Faithful geyser in Yellowstone National Park. The data are in the R dataset `faithful`. One column contains the duration of each eruption; the other contains the waiting time to the next eruption.
  - (a) Plot `eruption` versus `waiting`. Is there a pattern? What is going on?
  - (b) Try `ts.plot(faithful$eruptions[1:50])`. Try other sets of eruptions, say `ts.plot(faithful$eruptions[51:100])`. (There is nothing magic about 50, but if you plot all 272 eruptions then the pattern might be harder to see. Choose any convenient number that lets you see what's going on.) What is going on?
4. This exercise relies on data from the neurobiology experiment described in Example 2.6.
  - (a) Download the data from the book's website.

- (b) Reproduce Figure 2.17.
  - (c) Make a plot similar to Figure 2.17 but for a different neuron and different tastant.
  - (d) Write an R function that accepts a neuron and tastant as input and produces a plot like Figure 2.17.
  - (e) Use the function from the previous part to look for neurons that respond to particular tastants. Describe your results.
5. This exercise relies on Example 2.8 about the Slater school. There were 8 cancers among 145 teachers. Figure 2.20 shows the likelihood function. Suppose the same incidence rate had been found among more teachers. How would that affect  $\ell(\theta)$ ? Make a plot similar to Figure 2.20, but pretending that there had been 80 cancers among 1450 teachers. Compare to Figure 2.20. What is the result? Does it make sense? Try other numbers if it helps you see what is going on.
6. This exercise continues Exercise 36 in Chapter 1. Let  $p$  be the fraction of the population that uses illegal drugs.
- (a) Suppose researchers know that  $p \approx .1$ . Jane and John are given the randomized response question. Jane answers “yes”; John answers “no”. Find the posterior probability that Jane uses cocaine; find the posterior probability that John uses cocaine.
  - (b) Now suppose that  $p$  is not known and the researchers give the randomized response question to 100 people. Let  $X$  be the number who answer “yes”. What is the likelihood function?
  - (c) What is the mle of  $p$  if  $X=50$ , if  $X=60$ , if  $X=70$ , if  $X=80$ , if  $X=90$ ?
7. This exercise deals with the likelihood function for Poisson distributions.
- (a) Let  $x_1, \dots, x_n \sim \text{i.i.d. Poi}(\lambda)$ . Find  $\ell(\lambda)$  in terms of  $x_1, \dots, x_n$ .
  - (b) Show that  $\ell(\lambda)$  depends only on  $\sum x_i$  and not on the specific values of the individual  $x_i$ 's.
  - (c) Let  $y_1, \dots, y_n$  be a sample from  $\text{Poi}(\lambda)$ . Show that  $\hat{\lambda} = \bar{y}$  is the m.l.e.
  - (d) Find the m.l.e. in Example 1.4.
8. The book *Data* ANDREWS AND HERZBERG [1985] contains lots of data sets that have been used for various purposes in statistics. One famous data set records the annual number of deaths by horsekicks in the Prussian Army from 1875-1894 for each

of 14 corps. Download the data from **STATLIB** at [HTTP://LIB.STAT.CMU.EDU/DATASETS/ANDREWS/T04.1](http://lib.stat.cmu.edu/datasets/ANDREWS/T04.1). (It is Table 4.1 in the book.) Let  $Y_{ij}$  be the number of deaths in year  $i$ , corps  $j$ , for  $i = 1875, \dots, 1894$  and  $j = 1, \dots, 14$ . The  $Y_{ij}$ s are in columns 5–18 of the table.

- (a) What are the first four columns of the table?
  - (b) What is the last column of the table?
  - (c) What is a good model for the data?
  - (d) Suppose you model the data as i.i.d.  $\text{Poi}(\lambda)$ . (Yes, that's a good answer to the previous question.)
    - i. Plot the likelihood function for  $\lambda$ .
    - ii. Find  $\hat{\lambda}$ .
    - iii. What can you say about the rate of death by horsekick in the Prussian cavalry at the end of the 19th century?
  - (e) Is there any evidence that different corps had different death rates? How would you investigate that possibility?
9. Use the data from Example 2.8. Find the m.l.e. for  $\theta$ .
10. John is a runner and frequently runs from his home to his office. He wants to measure the distance, so once a week he drives to work and measures the distance on his car's odometer. Unfortunately, the odometer records distance only to the nearest mile. (John's odometer changes abruptly from one digit to the next. When the odometer displays a digit  $d$ , it is not possible to infer how close it is to becoming  $d + 1$ .) John drives the route ten times and records the following data from the odometer: 3, 3, 3, 4, 3, 4, 3, 3, 4, 3. Find the m.l.e. of the exact distance. You may make reasonable assumptions as needed.
11.  $X_1, \dots, X_n \sim \text{Normal}(\mu, 1)$ . Multiple choice: The m.l.e.  $\hat{\mu}$  is found from the equation
- (a)  $\frac{d}{d\mu} \frac{d}{dx} f(x_1, \dots, x_n | \mu) = 0$
  - (b)  $\frac{d}{d\mu} f(x_1, \dots, x_n | \mu) = 0$
  - (c)  $\frac{d}{dx} f(x_1, \dots, x_n | \mu) = 0$
12. This exercise deals with the likelihood function for Normal distributions.
- (a) Let  $y_1, \dots, y_n \sim \text{i.i.d. N}(\mu, 1)$ . Find  $\ell(\mu)$  in terms of  $y_1, \dots, y_n$ .

- (b) Show that  $\ell(\mu)$  depends only on  $\sum y_i$  and not on the specific values of the individual  $y_i$ 's.
- (c) Let  $n = 10$  and choose a value for  $\mu$ . Use R to generate a sample of size 10 from  $N(\mu, 1)$ . Plot the likelihood function. How accurately can you estimate  $\mu$  from a sample of size 10?
- (d) Let  $y_1, \dots, y_{10} \sim$  i.i.d.  $N(\mu, \sigma)$  where  $\sigma$  is known but not necessarily equal to 1. Find  $\ell(\mu)$  in terms of  $y_1, \dots, y_{10}$  and  $\sigma$ .
- (e) Let  $y_1, \dots, y_{10} \sim$  i.i.d.  $N(\mu, \sigma)$  where  $\mu$  is known but  $\sigma$  is unknown. Find  $\ell(\sigma)$  in terms of  $y_1, \dots, y_{10}$  and  $\mu$ .
13. Let  $y_1, \dots, y_n$  be a sample from  $N(\mu, 1)$ . Show that  $\hat{\mu} = \bar{y}$  is the m.l.e.
14. Let  $y_1, \dots, y_n$  be a sample from  $N(\mu, \sigma)$  where  $\mu$  is known. Show that  $\hat{\sigma}^2 = n^{-1} \sum (y_i - \mu)^2$  is the m.l.e.
15. Recall the **discoveries** data from page 10 on the number of great discoveries each year. Let  $Y_i$  be the number of great discoveries in year  $i$  and suppose  $Y_i \sim \text{Poi}(\lambda)$ . Plot the likelihood function  $\ell(\lambda)$ . Figure 1.3 suggested that  $\lambda \approx 3.1$  explained the data reasonably well. How sure can we be about the 3.1?
16. Justify each step of Equation 2.6.
17. Page 156 discusses a simulation experiment comparing the sample mean and sample median as estimators of a population mean. Figure 2.27 shows the results of the simulation experiment. Notice that the vertical scale decreases from panel (a) to (b), to (c), to (d). Why? Give a precise mathematical formula for the amount by which the vertical scale should decrease. Does the actual decrease agree with your formula?
18. In the medical screening example on page 162, find the probability that the patient has the disease given that the test is negative.
19. Country A suspects country B of having hidden chemical weapons. Based on secret information from their intelligence agency they calculate  $P[B \text{ has weapons}] = .8$ . But then country B agrees to inspections, so A sends inspectors. If there are no weapons then of course the inspectors won't find any. But if there are weapons then they will be well hidden, with only a 20% chance of being found. I.e.,
- $$P[\text{finding weapons} | \text{weapons exist}] = .2. \tag{2.19}$$

No weapons are found. Find the probability that B has weapons. I.e., find

$$\Pr[B \text{ has weapons} | \text{no weapons are found}].$$

20. Let  $T$  be the amount of time a customer spends on Hold when calling the computer help line. Assume that  $T \sim \exp(\lambda)$  where  $\lambda$  is unknown. A sample of  $n$  calls is randomly selected. Let  $t_1, \dots, t_n$  be the times spent on Hold.
  - (a) Choose a value of  $\lambda$  for doing simulations.
  - (b) Use R to simulate a sample of size  $n = 10$ .
  - (c) Plot  $\ell(\lambda)$  and find  $\hat{\lambda}$ .
  - (d) About how accurately can you determine  $\lambda$ ?
  - (e) Show that  $\ell(\lambda)$  depends only on  $\sum t_i$  and not on the values of the individual  $t_i$ 's.
21. There are two coins. One is fair; the other is two-headed. You randomly choose a coin and toss it.
  - (a) What is the probability the coin lands Heads?
  - (b) What is the probability the coin is two-headed given that it landed Heads?
  - (c) What is the probability the coin is two-headed given that it landed Tails? Give a formal proof, not intuition.
  - (d) You are about to toss the coin a second time. What is the probability that the second toss lands Heads given that the first toss landed Heads?
22. There are two coins. For coin A,  $P[H] = 1/4$ ; for coin B,  $P[H] = 2/3$ . You randomly choose a coin and toss it.
  - (a) What is the probability the coin lands Heads?
  - (b) What is the probability the coin is A given that it landed Heads? What is the probability the coin is A given that it landed Tails?
  - (c) You are about to toss the coin a second time. What is the probability the second toss lands Heads given that the first toss landed Heads?
23. At Dupont College (apologies to Tom Wolfe) Math SAT scores among math majors are distributed  $N(700, 50)$  while Math SAT scores among non-math majors are distributed  $N(600, 50)$ . 5% of the students are math majors. A randomly chosen student has a math SAT score of 720. Find the probability that the student is a math major.

24. The Great Randi is a professed psychic and claims to know the outcome of coin flips. This problem concerns a sequence of 20 coin flips that Randi will try to guess (or not guess, if his claim is correct).
- Take the prior  $P[\text{Randi is psychic}] = .01$ .
    - Before any guesses have been observed, find  $P[\text{first guess is correct}]$  and  $P[\text{first guess is incorrect}]$ .
    - After observing 10 consecutive correct guesses, find the updated  $P[\text{Randi is psychic}]$ .
    - After observing 10 consecutive correct guesses, find  $P[\text{next guess is correct}]$  and  $P[\text{next guess is incorrect}]$ .
    - After observing 20 consecutive correct guesses, find  $P[\text{next guess is correct}]$  and  $P[\text{next guess is incorrect}]$ .
  - Two statistics students, a skeptic and a believer discuss Randi after class.  
 Believer: *I believe her, I think she's psychic.*  
 Skeptic: *I doubt it. I think she's a hoax.*  
 Believer: *How could you be convinced? What if Randi guessed 10 in a row? What would you say then?*  
 Skeptic: *I would put that down to luck. But if she guessed 20 in a row then I would say  $P[\text{Randi can guess coin flips}] \approx .5$ .*  
 Find the skeptic's prior probability that Randi can guess coin flips.
  - Suppose that Randi doesn't claim to guess coin tosses perfectly, only that she can guess them at better than 50%. 100 trials are conducted. Randi gets 60 correct. Write down  $H_0$  and  $H_a$  appropriate for testing Randi's claim. Do the data support the claim? What if 70 were correct? Would that support the claim?
  - The Great Sandi, a statistician, writes the following R code to calculate a probability for Randi.

```
y <- rbinom(500, 100, .5)
sum(y == 60) / 500
```

What is Sandi trying to calculate? Write a formula (Don't evaluate it.) for the quantity Sandi is trying to calculate.

25. Let  $w$  be the fraction of free throws that Shaquille O'Neal (or any other player of your choosing) makes during the next NBA season. Find a density that approximately represents your prior opinion for  $w$ .

26. Let  $t$  be the amount of time between the moment when the sun first touches the horizon in the afternoon and the moment when it sinks completely below the horizon. Without making any observations, assess your distribution for  $t$ .
27. Assess your prior distribution for  $b$ , the proportion of M&M's that are brown. Buy as many M&M's as you like and count the number of browns. Calculate your posterior distribution.
28.
  - (a) Let  $y \sim N(\theta, 1)$  and let the prior distribution for  $\theta$  be  $\theta \sim N(0, 1)$ .
    - i. When  $y$  has been observed, what is the posterior density of  $\theta$ ?
    - ii. Show that the density in part i. is a Normal density.
    - iii. Find its mean and SD.
  - (b) Let  $y \sim N(\theta, \sigma_y)$  and let the prior distribution for  $\theta$  be  $\theta \sim N(m, \sigma)$ . Suppose that  $\sigma_y$ ,  $m$ , and  $\sigma$  are known constants.
    - i. When  $y$  has been observed, what is the posterior density of  $\theta$ ?
    - ii. Show that the density in part i. is a Normal density.
    - iii. Find its mean and SD.
  - (c) Let  $y_1, \dots, y_n$  be a sample of size  $n$  from  $N(\theta, \sigma_y)$  and let the prior distribution for  $\theta$  be  $\theta \sim N(m, \sigma)$ . Suppose that  $\sigma_y$ ,  $m$ , and  $\sigma$  are known constants.
    - i. When  $y_1, \dots, y_n$  have been observed, what is the posterior density of  $\theta$ ?
    - ii. Show that the density in part i. is a Normal density.
    - iii. Find its mean and SD.
29. Verify Equations 2.16, 2.17, and 2.18.
30. Refer to the discussion of predictive intervals on page 172. Justify the claim that  $(-\infty, -0.72)$ ,  $(-3.65, -0.36)$ , and  $(-3.28, \infty)$  are 90% prediction intervals. Find the corresponding 80% prediction intervals.
31.
  - (a) Following Example 2.17 (pg. 173), find  $\Pr[y_f = k | y_1, \dots, y_n]$  for  $k = 1, 2, 3, 4$ .
  - (b) Using the results from part (a), make a plot analogous to Figure 2.33 (pg. 172).
32. Suppose you want to test whether the random number generator in R generates each of the digits  $0, 1, \dots, 9$  with probability 0.1. How could you do it? You may consider first testing whether R generates 0 with the right frequency, then repeating the analysis for each digit.
33.
  - (a) Repeat the analysis of Example 2.18 (pg. 181), but for `dose = 1` and `dose = 2`.

- (b) Test the hypothesis that increasing the dose from 1 to 2 makes no difference in tooth growth.
- (c) Test the hypothesis that the effect of increasing the dose from 1 to 2 is the same for  $\text{supp} = \text{VC}$  as it is for  $\text{supp} = \text{OJ}$ .
- (d) Do the answers to parts (a), (b) and (c) agree with your subjective assessment of Figures 2.2, 2.3, and 2.6?
34. Continue Exercise 37 from Chapter 1. The autoganzfeld trials resulted in  $X = 122$ .
- What is the parameter in this problem?
  - Plot the likelihood function.
  - Test the “no ESP, no cheating” hypothesis.
  - Adopt and plot a reasonable and mathematically tractable prior distribution for the parameter. Compute and plot the posterior distribution.
  - Find the probability of a match on the next trial given  $X = 122$ .
  - What do you conclude?
35. Three biologists named Asiago, Brie, and Cheshire are studying a mutation in morning glories, a species of flowering plant. The mutation causes the flowers to be white rather than colored. But it is not known whether the mutation has any effect on the plants’ fitness. To study the question, each biologist takes a random sample of morning glories having the mutation, counts the seeds that each plant produces, and calculates a likelihood set for the average number of seeds produced by mutated morning glories.
- Asiago takes a sample of size  $n_A = 100$  and calculates a  $\text{LS}_{.1}$  set. Brie takes a sample of size  $n_B = 400$  and calculates a  $\text{LS}_{.1}$  set. Cheshire takes a sample of size  $n_C = 100$  and calculates a  $\text{LS}_{.2}$  set.
- Who will get the longer interval, Asiago or Brie? About how much longer will it be? Explain.
  - Who will get the longer interval, Asiago or Cheshire? About how much longer will it be? Explain.
36. In the 1990’s, a committee at MIT wrote **A Study on the Status of Women Faculty in Science at MIT**. In 1994 there were 15 women among the 209 tenured women in the six departments of the School of Science. They found, among other things, that the amount of resources (money, lab space, etc.) given to women was, on average,

less than the amount given to men. The report goes on to pose the question: *Given the tiny number of women faculty in any department one might ask if it is possible to obtain significant data to support a claim of gender differences . . .*

What does statistics say about it? Focus on a single resource, say laboratory space. The distribution of lab space is likely to be skewed. I.e., there will be a few people with lots more space than most others. So let's model the distribution of lab space with an Exponential distribution. Let  $x_1, \dots, x_{15}$  be the amounts of space given to tenured women, so  $x_i \sim \text{Exp}(\lambda_w)$  for some unknown parameter  $\lambda_w$ . Let  $M$  be the average lab space given to tenured men. Assume that  $M$  is known to be 100, from the large number of tenured men. If there is no discrimination, then  $\lambda_w = 100$ . ( $\lambda_w$  is  $E(x_i)$ .)

Chris Stats writes the following R code.

```
y <- rexp(15,.01)
m <- mean(y)
s <- sqrt(var(y) / 15)
lo <- m - 2*s
hi <- m + 2*s
```

What is  $y$  supposed to represent? What is  $(\text{lo}, \text{hi})$  supposed to represent?

Now Chris puts the code in a loop.

```
n <- 0
for (i in 1:1000) {
 y <- rexp(15,.01)
 m <- mean(y)
 s <- sqrt(var(y))
 lo <- m - 2*s
 hi <- m + 2*s
 if (lo < 100 & hi > 100) n <- n+1
}
print (n/1000)
```

What is  $n/1000$  supposed to represent? If a sample size of 15 is sufficiently large for the Central Limit Theorem to apply, then what, approximately, is the value of  $n/1000$ ?

37. Refer to the R code in Example 2.1 (pg. 99). Why was it necessary to have a brace (“{”) after the line

```
for (j in 1:3)
```

but not after the line

```
for (i in 1:2)?
```

## CHAPTER 3

# REGRESSION

### 3.1 Introduction

Regression is the study of how the distribution of one variable,  $Y$ , changes according to the value of another variable,  $X$ . R comes with many data sets that offer regression examples. Four are shown in Figure 3.1.

1. The data set `attenu` contains data on several variables from 182 earthquakes, including hypocenter-to-station distance and peak acceleration. Figure 3.1 (a) shows acceleration plotted against distance. There is a clear relationship between  $X =$  distance and the distribution of  $Y =$  acceleration. When  $X$  is small, the distribution of  $Y$  has a long right-hand tail. But when  $X$  is large,  $Y$  is always small.
2. The data set `airquality` contains data about air quality in New York City. Ozone levels  $Y$  are plotted against temperature  $X$  in Figure 3.1 (b). When  $X$  is small then the distribution of  $Y$  is concentrated on values below about 50 or so. But when  $X$  is large,  $Y$  can range up to about 150 or so.
3. Figure 3.1 (c) shows data from `mtcars`. Weight is on the abscissa and the type of transmission (`manual=1`, `automatic=0`) is on the ordinate. The distribution of weight is clearly different for cars with automatic transmissions than for cars with manual transmissions.
4. The data set `faithful` contains data about eruptions of the Old Faithful geyser in Yellowstone National Park. Figure 3.1 (d) shows  $Y =$  time to next eruption plotted against  $X =$  duration of current eruption. Small values of  $X$  tend to indicate small values of  $Y$ .

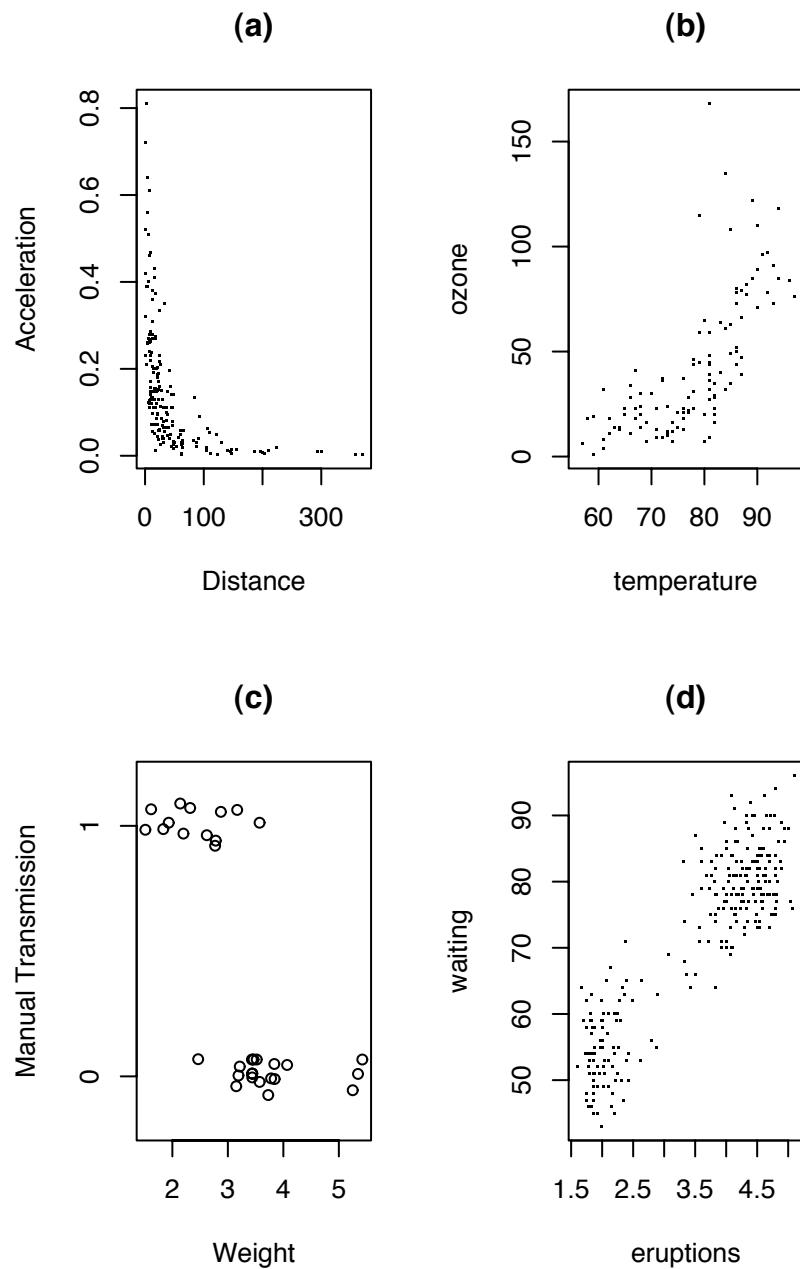


Figure 3.1: Four regression examples

Figure 3.1 was produced by the following R snippet.

```
par (mfrow=c(2,2))

data (attenu)
plot (attenu$dist, attenu$accel, xlab="Distance",
 ylab="Acceleration", main="(a)", pch=".")

data (airquality)
plot (airquality$Temp, airquality$Ozone, xlab="temperature",
 ylab="ozone", main="(b)", pch=".")

data (mtcars)
stripchart (mtcars$wt ~ mtcars$am, pch=1, xlab="Weight",
 method="jitter", ylab="Manual Transmission",
 main="(c)")

data (faithful)
plot (faithful, pch=".", main="(d)")
```

Both continuous and discrete variables can turn up in regression problems. In the `attenu`, `airquality` and `faithful` datasets, both  $X$  and  $Y$  are continuous. In `mtcars`, it seems natural to think of how the distribution of  $Y = \text{weight}$  varies with  $X = \text{transmission}$ , in which case  $X$  is discrete and  $Y$  is continuous. But we could also consider how the fraction of cars  $Y$  with automatic transmissions varies as a function of  $X = \text{weight}$ , in which case  $Y$  is discrete and  $X$  is continuous.

In many regression problems we just want to display the relationship between  $X$  and  $Y$ . Often a scatterplot or stripchart will suffice, as in Figure 3.1. Other times, we will use a statistical model to describe the relationship. The statistical model may have unknown parameters which we may wish to estimate or otherwise make inference for. Examples of parametric models will come later. Our study of regression begins with data display.

In many instances a simple plot is enough to show the relationship between  $X$  and  $Y$ . But sometimes the relationship is obscured by the scatter of points. Then it helps to draw a smooth curve through the data. Examples 3.1 and 3.2 illustrate.

### **Example 3.1** (1970 Draft Lottery)

The result of the 1970 draft lottery is available at DASL . The website explains:

"In 1970, Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The first date drawn from the drum received draft number one and eligible men born on that date were drafted first. In a truly random lottery there should be no relationship between the date and the draft number."

Figure 3.2 shows the data, with  $X$  = day of year and  $Y$  = draft number. There is no apparent relationship between  $X$  and  $Y$ .

Figure 3.2 was produced with the following snippet.

```
plot (draft$Day.of.year, draft$Draft.No,
 xlab="Day of year", ylab="Draft number")
```

More formally, a relationship between  $X$  and  $Y$  usually means that the expected value of  $Y$  is different for different values of  $X$ . (We don't consider changes in SD or other aspects of the distribution here.) Typically, when  $X$  is a continuous variable, changes in  $Y$  are smooth, so we would adopt the model

$$\mathbb{E}[Y|X] = g(X) \quad (3.1)$$

for some unknown smooth function  $g$ .

R has a variety of built-in functions to estimate  $g$ . These functions are called *scatterplot smoothers*, for obvious reasons. Figure 3.3 shows the draft lottery data with two scatterplot smoother estimates of  $g$ . Both estimates show a clear trend, that birthdays later in the year were more likely to have low draft numbers. Following discovery of this trend, the procedure for drawing draft numbers was changed in subsequent years.

Figure 3.3 was produced with the following snippet.

```
x <- draft$Day.of.year
y <- draft$Draft.No
plot (x, y, xlab="Day of year", ylab="Draft number")
lines (lowess (x, y))
lines (supsmu (x, y), lty=2)
```

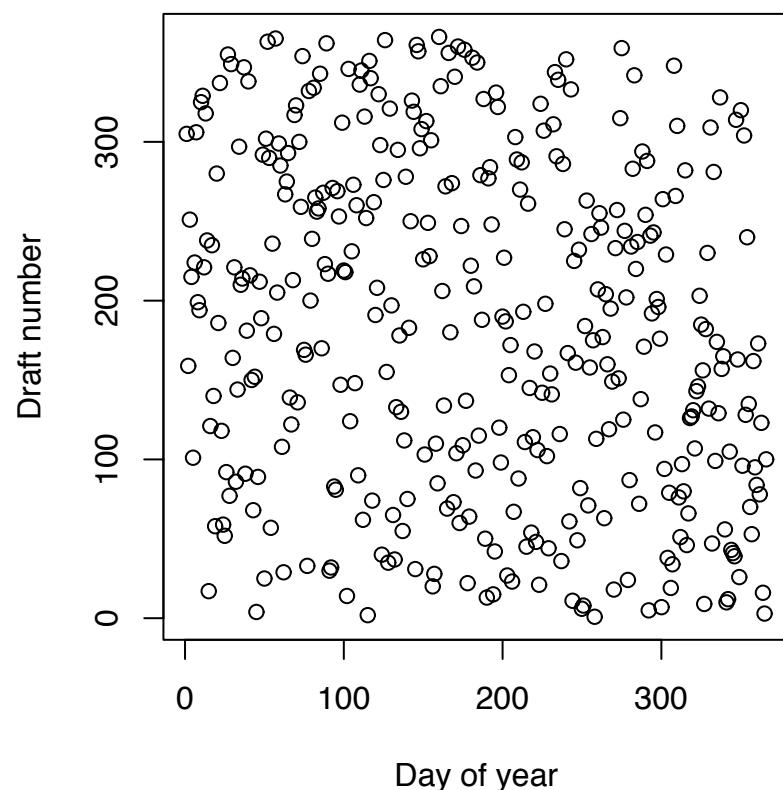


Figure 3.2: 1970 draft lottery. Draft number vs. day of year

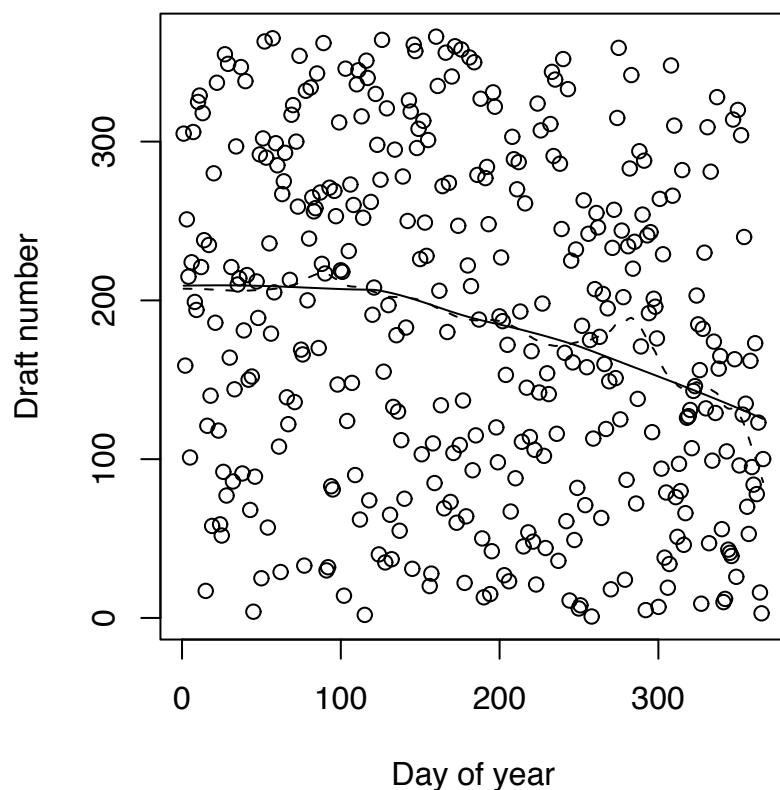


Figure 3.3: 1970 draft lottery. Draft number vs. day of year. Solid curve fit by lowess; dashed curve fit by supsmu.

- `lowess` (locally weighted scatterplot smoother) and `supsmu` (super smoother) are two of R's scatterplot smoothers. In the figure, the `lowess` curve is less wiggly than the `supsmu` curve. Each smoother has a tuning parameter that can make the curve more or less wiggly. Figure 3.3 was made with the default values for both smoothers.

**Example 3.2** (Seedlings, continued)

As mentioned in Example 1.6, the seedlings study was carried out at the Ceweeta Long Term Ecological Research station in western North Carolina. There were five plots at different elevations on a hillside. Within each plot there was a  $60\text{m} \times 1\text{m}$  strip running along the hillside divided into 60  $1\text{m} \times 1\text{m}$  quadrats. It is possible that the arrival rate of New seedlings and the survival rates of both Old and New seedlings are different in different plots and different quadrats. Figure 3.4 shows the total number of New seedlings in each of the quadrats in one of the five plots. The `lowess` curve brings out the spatial trend: low numbers to the left, a peak around quadrat 40, and a slight falling off by quadrat 60.

Figure 3.4 was produced by

```
plot (total.new, xlab="quadrat index",
 ylab="total new seedlings")
lines (lowess (total.new))
```

In a regression problem the data are pairs  $(x_i, y_i)$  for  $i = 1, \dots, n$ . For each  $i$ ,  $y_i$  is a random variable whose distribution depends on  $x_i$ . We write

$$y_i = g(x_i) + \epsilon_i. \quad (3.2)$$

Equation 3.2 expresses  $y_i$  as a systematic or explainable part  $g(x_i)$  and an unexplained part  $\epsilon_i$ .  $g$  is called the *regression function*. Often the statistician's goal is to estimate  $g$ . As usual, the most important tool is a simple plot, similar to those in Figures 3.1 through 3.4.

Once we have an estimate,  $\hat{g}$ , for the regression function  $g$  (either by a scatterplot smoother or by some other technique) we can calculate  $r_i \equiv y_i - \hat{g}(x_i)$ . The  $r_i$ 's are estimates of the  $\epsilon_i$ 's and are called *residuals*. The  $\epsilon_i$ 's themselves are called *errors*. Because the  $r_i$ 's are estimates they are sometimes written with the "hat" notation:

$$\hat{\epsilon}_i = r_i = \text{estimate of } \epsilon_i$$

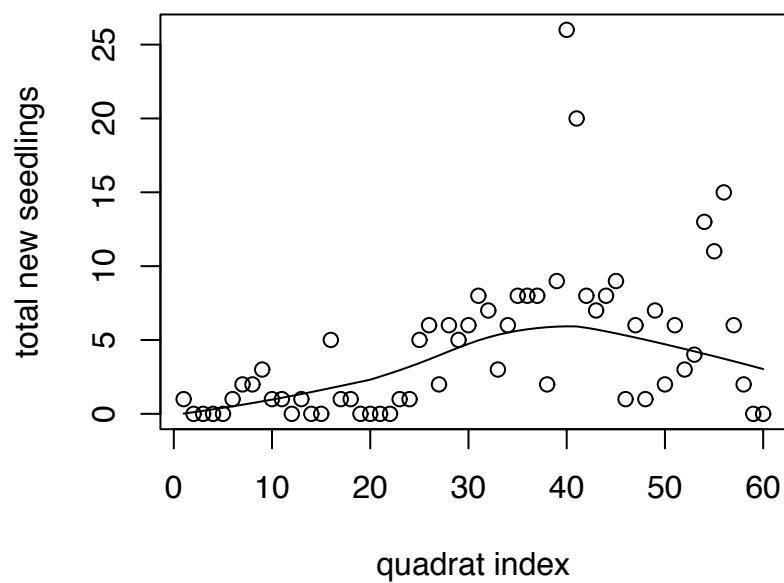


Figure 3.4: Total number of New seedlings 1993 – 1997, by quadrat.

Residuals are used to evaluate and assess the fit of models for  $g$ , a topic which is beyond the scope of this book.

In regression we use one variable to explain or predict the other. It is customary in statistics to plot the predictor variable on the  $x$ -axis and the predicted variable on the  $y$ -axis. The predictor is also called the *independent* variable, the *explanatory* variable, the *covariate*, or simply  $x$ . The predicted variable is called the *dependent* variable, or simply  $y$ . (In Economics  $x$  and  $y$  are sometimes called the *exogenous* and *endogenous* variables, respectively.) Predicting or explaining  $y$  from  $x$  is not perfect; knowing  $x$  does not tell us  $y$  exactly. But knowing  $x$  does tell us something about  $y$  and allows us to make more accurate predictions than if we didn't know  $x$ .

Regression models are agnostic about causality. In fact, instead of using  $x$  to predict  $y$ , we could use  $y$  to predict  $x$ . So for each pair of variables there are two possible regressions: using  $x$  to predict  $y$  and using  $y$  to predict  $x$ . Sometimes neither variable causes the other. For example, consider a sample of cities and let  $x$  be the number of churches and  $y$  be the number of bars. A scatterplot of  $x$  and  $y$  will show a strong relationship between them. But the relationship is caused by the population of the cities. Large cities have large numbers of bars and churches and appear near the upper right of the scatterplot. Small cities have small numbers of bars and churches and appear near the lower left.

Scatterplot smoothers are a relatively unstructured way to estimate  $g$ . Their output follows the data points more or less closely as the tuning parameter allows  $\hat{g}$  to be more or less wiggly. Sometimes an unstructured approach is appropriate, but not always. The rest of Chapter 3 presents more structured ways to estimate  $g$ .

## 3.2 Normal Linear Models

### 3.2.1 Introduction

In Section 1.3.4 we studied the Normal distribution, useful for continuous populations having a central tendency with roughly equally sized tails. In Section 3.2 we generalize to the case where there are many Normal distributions with different means which depend in a systematic way on another variable. We begin our study with an example in which there are three distinct distributions.

#### Example 3.3 (Hot Dogs, continued)

Figure 3.5 displays calorie data for three types of hot dogs. It appears that poultry hot dogs have, on average, slightly fewer calories than beef or meat hot dogs. How should we model these data?

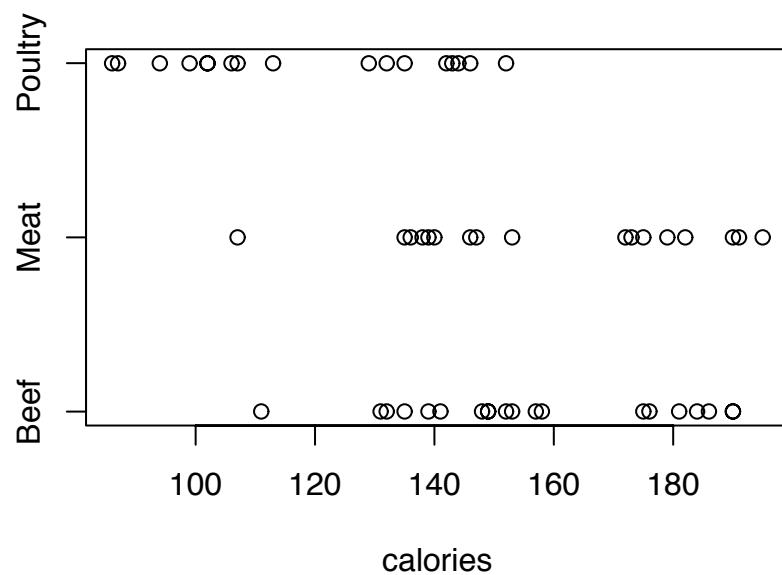


Figure 3.5: Calorie content of hot dogs

Figure 3.5 was produced with

```
stripchart (hotdogs$Calories ~ hotdogs>Type, pch=1,
 xlab="calories")
```

There are 20 Beef, 17 Meat and 17 Poultry hot dogs in the sample. We think of them as samples from much larger populations. Figure 3.6 shows density estimates of calorie content for the three types. For each type of hot dog, the calorie contents cluster around a central value and fall off to either side without a particularly long left or right tail. So it is reasonable, at least as a first attempt, to model the three distributions as Normal. Since the three distributions have about the same amount of spread we model them as all having the same SD. We adopt the model

$$\begin{aligned} B_1, \dots, B_{20} &\sim \text{i.i.d. } N(\mu_B, \sigma) \\ M_1, \dots, M_{17} &\sim \text{i.i.d. } N(\mu_M, \sigma) \\ P_1, \dots, P_{17} &\sim \text{i.i.d. } N(\mu_P, \sigma), \end{aligned} \tag{3.3}$$

where the  $B_i$ 's,  $M_i$ 's and  $P_i$ 's are the calorie contents of the Beef, Meat and Poultry hot dogs respectively. Figure 3.6 suggests

$$\mu_B \approx 150; \quad \mu_M \approx 160; \quad \mu_P \approx 120; \quad \sigma \approx 30.$$

An equivalent formulation is

$$\begin{aligned} B_1, \dots, B_{20} &\sim \text{i.i.d. } N(\mu, \sigma) \\ M_1, \dots, M_{17} &\sim \text{i.i.d. } N(\mu + \delta_M, \sigma) \\ P_1, \dots, P_{17} &\sim \text{i.i.d. } N(\mu + \delta_P, \sigma) \end{aligned} \tag{3.4}$$

Models 3.3 and 3.4 are mathematically equivalent. Each has three parameters for the population means and one for the SD. They describe exactly the same set of distributions and the parameters of either model can be written in terms of the other. The equivalence is shown in Table 3.3. For the purpose of further exposition we adopt Model 3.4.

We will see later how to carry out inferences regarding the parameters. For now we stop with the model.

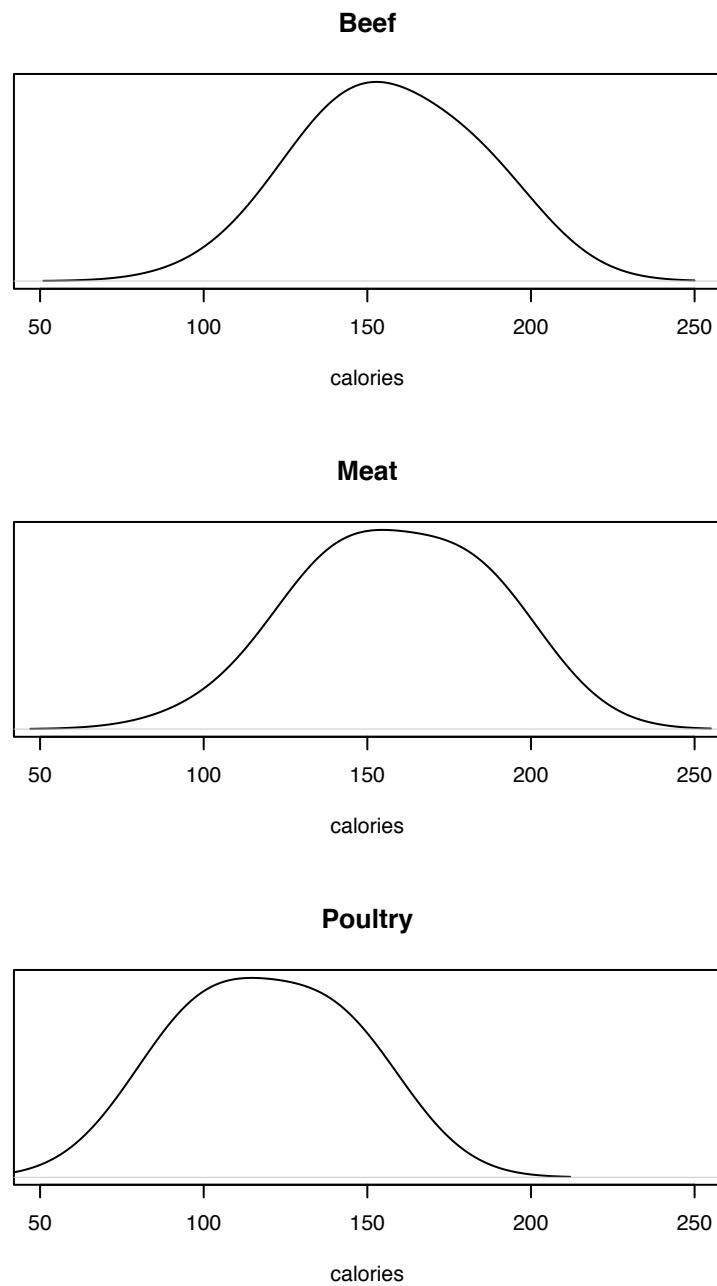


Figure 3.6: Density estimates of calorie contents of hot dogs

Figure 3.6 was produced with the following snippet.

```
par (mfrow=c(3,1))
plot (density (hotdogs$C[hotdogs$T=="Beef"] , bw=20) ,
 xlim=c(50,250) , yaxt="n" , ylab="" , xlab="calories" ,
 main="Beef")
plot (density (hotdogs$C[hotdogs$T=="Meat"] , bw=20) ,
 xlim=c(50,250) , yaxt="n" , ylab="" , xlab="calories" ,
 main="Meat")
plot (density (hotdogs$C[hotdogs$T=="Poultry"] , bw=20) ,
 xlim=c(50,250) , yaxt="n" , ylab="" , xlab="calories" ,
 main="Poultry")
```

- `hotdogs$C` and `hotdogs$T` illustrate a convenient feature of R, that components of a structure can be abbreviated. Instead of typing `hotdogs$Calories` and `hotdogs$type` we can use the abbreviations. The same thing applies to arguments of functions.
- `density ( ... , bw=20 )` specifies the *bandwidth* of the density estimate. Larger bandwidth gives a smoother estimate; smaller bandwidth gives a more wiggly estimate. Try different bandwidths to see what they do.

The `PlantGrowth` data set in R provides another example. As R explains, the data previously appeared in Dobson [1983] and are

“Results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions.”

The first several lines are

```
weight group
1 4.17 ctrl
2 5.58 ctrl
3 5.18 ctrl
```

Figure 3.7 shows the whole data set.

It appears that plants grown under different treatments tend to have different weights. In particular, plants grown under Treatment 1 appear to be smaller on average than plants grown under either the Control or Treatment 2. What statistical model should we adopt?

| Model 3.3       | Model 3.4        | Interpretation                                           | Approximate value |
|-----------------|------------------|----------------------------------------------------------|-------------------|
| $\mu_B$         | $\mu$            | mean calorie content of Beef hot dogs                    | 150               |
| $\mu_M$         | $\mu + \delta_M$ | mean calorie content of Meat hot dogs                    | 160               |
| $\mu_P$         | $\mu + \delta_P$ | mean calorie content of Poultry hot dogs                 | 120               |
| $\mu_M - \mu_B$ | $\delta_M$       | mean calorie difference between Beef and Meat hotdogs    | 10                |
| $\mu_P - \mu_B$ | $\delta_P$       | mean calorie difference between Beef and Poultry hotdogs | -30               |
| $\sigma$        | $\sigma$         | SD of calorie content within a single type of hot dog    | 30                |

Table 3.1: Correspondence between Models 3.3 and 3.4

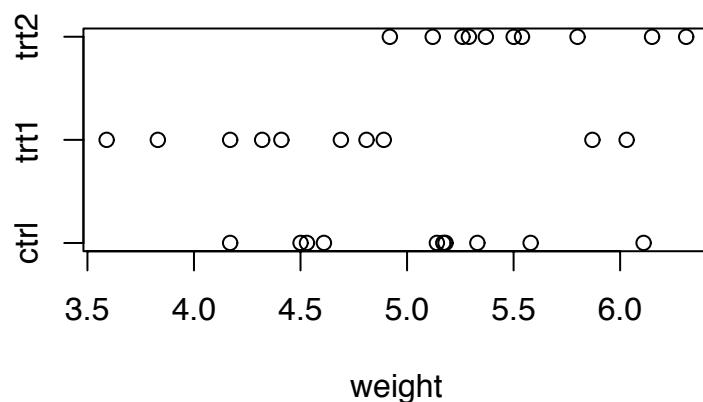


Figure 3.7: The PlantGrowth data

Figure 3.7 was produced with the following snippet.

```
stripchart (PlantGrowth$weight ~ PlantGrowth$group, pch=1,
 xlab="weight")
```

First, we think of the 10 plants grown under each condition as a sample from a much larger population of plants that could have been grown. Second, a look at the data suggests that the weights in each group are clustered around a central value, approximately symmetrically without an especially long tail in either direction. So we model the weights as having Normal distributions.

But we should allow for the possibility that the three populations have different means. (We do not address the possibility of different SD's here.) Let  $\mu$  be the population mean of plants grown under the Control condition,  $\delta_1$  and  $\delta_2$  be the extra weight due to Treatment 1 and Treatment 2 respectively, and  $\sigma$  be the SD. We adopt the model

$$\begin{aligned} W_{C,1}, \dots, W_{C,10} &\sim \text{i.i.d. } N(\mu, \sigma) \\ W_{T_1,1}, \dots, W_{T_1,10} &\sim \text{i.i.d. } N(\mu + \delta_1, \sigma) \\ W_{T_2,1}, \dots, W_{T_2,10} &\sim \text{i.i.d. } N(\mu + \delta_2, \sigma). \end{aligned} \tag{3.5}$$

There is a mathematical structure shared by 3.4, 3.5 and many other statistical models, and some common statistical notation to describe it. We'll use the hot dog data to illustrate.

#### Example 3.4 (Hot Dogs, continued)

Example 3.4 continues Example 2.2. First, there is the main variable of interest, often called the *response variable* and denoted  $Y$ . For the hot dog data  $Y$  is calorie content. (Another analysis could be made in which  $Y$  is sodium content.)

The distribution of  $Y$  is different under different circumstances. In this example,  $Y$  has a Normal distribution whose mean depends on the type of hot dog. In general, the distribution of  $Y$  will depend on some quantity of interest, called a *covariate*, *regressor*, or *explanatory variable*. Covariates are often called  $X$ .

The data consists of multiple data points, or *cases*. We write  $Y_i$  and  $X_i$  for the  $i$ 'th case. It is usual to represent the data as a matrix with one row for each case. One column is for  $Y$ ; the other columns are for explanatory variables. For the hot dog data the matrix is

| Type | Calories | Sodium |
|------|----------|--------|
| Beef | 186      | 495    |
| Beef | 181      | 477    |

|         |     |     |
|---------|-----|-----|
| ...     |     |     |
| Meat    | 140 | 428 |
| Meat    | 138 | 339 |
| Poultry | 129 | 430 |
| Poultry | 132 | 375 |
| ...     |     |     |

(For analysis of calories, the third column is irrelevant.)

Rewriting the data matrix in a slightly different form reveals some mathematical structure common to many models. There are 54 cases in the hotdog study. Let  $(Y_1, \dots, Y_{54})$  be their calorie contents. For each  $i$  from 1 to 54, define two new variables  $X_{1,i}$  and  $X_{2,i}$  by

$$X_{1,i} = \begin{cases} 1 & \text{if the } i\text{'th hot dog is Meat,} \\ 0 & \text{otherwise} \end{cases}$$

and

$$X_{2,i} = \begin{cases} 1 & \text{if the } i\text{'th hot dog is Poultry,} \\ 0 & \text{otherwise.} \end{cases}$$

$X_{1,i}$  and  $X_{2,i}$  are indicator variables. Two indicator variables suffice because, for the  $i$ 'th hot dog, if we know  $X_{1,i}$  and  $X_{2,i}$ , then we know what type it is. (More generally, if there are  $k$  populations, then  $k - 1$  indicator variables suffice.) With these new variables, Model 3.4 can be rewritten as

$$Y_i = \mu + \delta_M X_{1,i} + \delta_P X_{2,i} + \epsilon_i \quad (3.6)$$

for  $i = 1, \dots, 54$ , where

$$\epsilon_1, \dots, \epsilon_{54} \sim \text{i.i.d. } N(0, \sigma).$$

Equation 3.6 is actually 54 separate equations, one for each case. We can write them succinctly using vector and matrix notation. Let

$$\begin{aligned} \mathbf{Y} &= (Y_1, \dots, Y_{54})^t, \\ \mathbf{B} &= (\mu, \delta_M, \delta_P)^t, \\ \mathbf{E} &= (\epsilon_1, \dots, \epsilon_{54})^t, \end{aligned}$$

(The transpose is there because, by convention, vectors are column vectors.) and

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}$$

$\mathbf{X}$  is a  $54 \times 3$  matrix. The first 20 lines are for the Beef hot dogs; the next 17 are for the Meat hot dogs; and the final 17 are for the Poultry hot dogs. Equation 3.6 can be written

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \quad (3.7)$$

Equations similar to 3.6 and 3.7 are common to many statistical models. For the PlantGrowth data (page 211) let

$Y_i$  = weight of  $i$ 'th plant,

$$X_{1,i} = \begin{cases} 1 & \text{if } i\text{'th plant received treatment 1} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{2,i} = \begin{cases} 1 & \text{if } i\text{'th plant received treatment 2} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{Y} = (Y_1, \dots, Y_{30})^t$$

$$\mathbf{B} = (\mu, \delta_1, \delta_2)^t$$

$$\mathbf{E} = (\epsilon_1, \dots, \epsilon_{30})^t$$

and

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}$$

Then analogously to 3.6 and 3.7 we can write

$$Y_i = \mu + \delta_1 X_{1,i} + \delta_2 X_{2,i} + \epsilon_i \quad (3.8)$$

and

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \quad (3.9)$$

Notice that Equation 3.6 is nearly identical to Equation 3.8 and Equation 3.7 is identical to Equation 3.9. Their structure is common to many statistical models. Each  $Y_i$  is written as the sum of two parts. The first part,  $\mathbf{XB}$ , ( $\mu + \delta_M X_{1,i} + \delta_P X_{2,i}$  for the hot dogs;  $\mu + \delta_1 X_{1,i} + \delta_2 X_{2,i}$  for PlantGrowth) is called *systematic*, *deterministic*, or *signal* and represents the explainable differences between populations. The second part,  $\mathbf{E}$ , or  $\epsilon_i$ , is random, or noise, and represents the differences between hot dogs or plants within a single population. The  $\epsilon_i$ 's are called *errors*. In statistics, the word “error” does not indicate a mistake; it simply means the noise part of a model, or the part left unexplained by covariates. Modelling a response variable as

$$\text{response} = \text{signal} + \text{noise}$$

is a useful way to think and will recur throughout this book.

In 3.6 the signal  $\mu + \delta_M X_{1,i} + \delta_P X_{2,i}$  is a linear function of  $(\mu, \delta_M, \delta_P)$ . In 3.8 the signal  $\mu + \delta_1 X_{1,i} + \delta_2 X_{2,i}$  is a linear function of  $(\mu, \delta_1, \delta_2)$ . Models in which the signal is a linear function of the parameters are called *linear models*.

In our examples so far,  $X$  has been an indicator. For each of a finite number of  $X$ 's there has been a corresponding population of  $Y$ 's. As the next example illustrates, linear models can also arise when  $X$  is a continuous variable.

**Example 3.5** (Ice Cream Consumption)

This example comes from DASL, which says

"Ice cream consumption was measured over 30 four-week periods from March 18, 1951 to July 11, 1953. The purpose of the study was to determine if ice cream consumption depends on the variables price, income, or temperature. The variables Lag-temp and Year have been added to the original data."

You can download the data from

<http://lib.stat.cmu.edu/DASL/Datafiles/IceCream.html>.

The first few lines look like this:

| date | IC   | price | income | temp | Lag-temp | Year |
|------|------|-------|--------|------|----------|------|
| 1    | .386 | .270  | 78     | 41   | 56       | 0    |
| 2    | .374 | .282  | 79     | 56   | 63       | 0    |
| 3    | .393 | .277  | 81     | 63   | 68       | 0    |

The variables are

**date** Time period (1-30) of the study (from 3/18/51 to 7/11/53)

**IC** Ice cream consumption in pints per capita

**Price** Price of ice cream per pint in dollars

**Income** Weekly family income in dollars

**Temp** Mean temperature in degrees F

**Lag-temp** Temp variable lagged by one time period

**Year** Year within the study (0 = 1951, 1 = 1952, 2 = 1953)

Figure 3.8 is a plot of consumption versus temperature. It looks as though an equation of the form

$$\text{consumption} = \beta_0 + \beta_1 \text{temperature} + \text{error} \quad (3.10)$$

would describe the data reasonably well. This is a linear model, not because consumption is a linear function of temperature, but because it is a linear function of  $(\beta_0, \beta_1)$ . To write it in matrix form, let

$$\mathbf{Y} = (\mathbf{IC}_1, \dots, \mathbf{IC}_{30})^t$$

$$\mathbf{B} = (\beta_0, \beta_1)^t$$

$$\mathbf{E} = (\epsilon_1, \dots, \epsilon_{30})^t$$

and

$$\mathbf{X} = \begin{pmatrix} 1 & \text{temp}_1 \\ 1 & \text{temp}_2 \\ \vdots & \vdots \\ 1 & \text{temp}_{30} \end{pmatrix}$$

The model is

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \quad (3.11)$$

Equation 3.11 is a linear model, identical to Equations 3.7 and 3.9.

Equation 3.7 (equivalently, 3.9 or 3.11) is the basic form of all linear models. Linear models are extremely useful because they can be applied to so many kinds of data sets. Section 3.2.2 investigates some of their theoretical properties and R's functions for fitting them to data.

### 3.2.2 Inference for Linear Models

Section 3.2.1 showed some graphical displays of data that were eventually described by linear models. Section 3.2.2 treats more formal inference for linear models. We begin by deriving the likelihood function.

Linear models are described by Equation 3.7 (equivalently, 3.9 or 3.11) which we repeat here for convenience:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \quad (3.12)$$

In general there is an arbitrary number of cases, say  $n$ , and an arbitrary number of covariates, say  $p$ . Equation 3.12 is shorthand for the collection of univariate equations

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i} + \epsilon_i \quad (3.13)$$

or equivalently,

$$Y_i \sim N(\mu_i, \sigma)$$

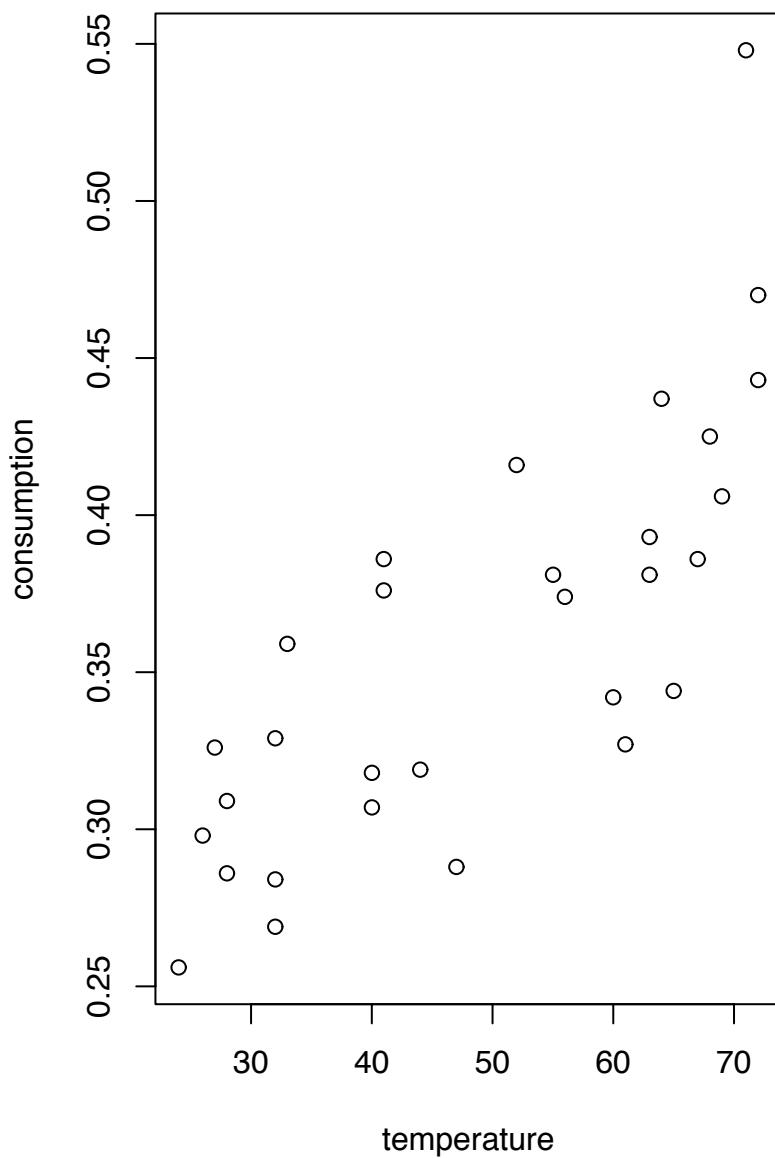


Figure 3.8: Ice cream consumption (pints per capita) versus mean temperature (°F)

for  $i = 1, \dots, n$  where  $\mu_i = \beta_0 + \sum_j \beta_j X_{j,i}$  and the  $\epsilon_i$ 's are i.i.d.  $N(0, \sigma)$ . There are  $p + 2$  parameters:  $(\beta_0, \dots, \beta_p, \sigma)$ . The likelihood function is

$$\begin{aligned}\ell(\beta_0, \dots, \beta_p, \sigma) &= \prod_{i=1}^n p(y_i | \beta_0, \dots, \beta_p, \sigma) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y_i-\mu_i}{\sigma})^2} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\sum_j \beta_j X_{j,i})}{\sigma}\right)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} \sum_i (y_i - (\beta_0 + \sum_j \beta_j X_{j,i}))^2}\end{aligned}\quad (3.14)$$

Likelihood 3.14 is a function of the  $p + 2$  parameters. To find the m.l.e.'s we could differentiate 3.14 with respect to each parameter in turn, set the derivatives equal to 0, and solve. But it is easier to take the log of 3.14 first, then differentiate and solve.

$$\log \ell(\beta_0, \dots, \beta_p, \sigma) = C - n \log \sigma - \frac{1}{2\sigma^2} \sum_i \left( y_i - (\beta_0 + \sum_j \beta_j X_{j,i}) \right)^2$$

for some irrelevant constant  $C$ , so we get the system of equations

$$\begin{aligned}\frac{1}{\hat{\sigma}^2} \sum_i (y_i - (\hat{\beta}_0 + \sum_j \hat{\beta}_j X_{j,i})) &= 0 \\ \frac{1}{\hat{\sigma}^2} \sum_i (y_i - (\hat{\beta}_0 + \sum_j \hat{\beta}_j X_{j,i})) X_{i,1} &= 0 \\ &\vdots \\ \frac{1}{\hat{\sigma}^2} \sum_i (y_i - (\hat{\beta}_0 + \sum_j \hat{\beta}_j X_{j,i})) X_{i,p} &= 0 \\ -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3} \sum_i (y_i - (\hat{\beta}_0 + \sum_j \hat{\beta}_j X_{j,i}))^2 &= 0\end{aligned}\quad (3.15)$$

Note the hat notation to indicate estimates. The m.l.e.'s  $(\hat{\beta}_0, \dots, \hat{\beta}_p, \hat{\sigma})$  are the values of the parameters that make the derivatives equal to 0 and therefore satisfy Equations 3.15. The first  $p + 1$  of these equations can be multiplied by  $\sigma^2$ , yielding  $p + 1$  linear equations in the  $p + 1$  unknown  $\beta$ 's. Because they're linear, they can be solved by linear algebra. The solution is

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

using the notation of Equation 3.12.

For each  $i \in \{1, \dots, n\}$ , let

$$\hat{y}_i = \hat{\beta}_0 + x_{1i}\hat{\beta}_1 + \dots + x_{pi}\hat{\beta}_p.$$

The  $\hat{y}_i$ 's are called *fitted* values. The residuals are

$$\begin{aligned} r_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 + x_{1i}\hat{\beta}_1 + \dots + x_{pi}\hat{\beta}_p \end{aligned}$$

and are estimates of the errors  $\epsilon_i$ . Finally, referring to the last line of Equation 3.15, the m.l.e.  $\hat{\sigma}$  is found from

$$\begin{aligned} 0 &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_i (y_i - (\beta_0 + \sum_j \beta_j X_{i,j}))^2 \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_i r_i^2 \end{aligned}$$

so

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i r_i^2$$

and

$$\hat{\sigma} = \left( \frac{\sum r_i^2}{n} \right)^{\frac{1}{2}} \tag{3.16}$$

In addition to the m.l.e.'s we often want to look at the likelihood function to judge, for example, how accurately each  $\beta$  can be estimated. The likelihood function for a single  $\beta_i$  comes from the Central Limit Theorem. We will not work out the math here but, fortunately, R will do all the calculations for us. We illustrate with the hot dog data.

### Example 3.6 (Hot Dogs, continued)

Estimating the parameters of a model is called *fitting a model to data*. R has built-in commands for fitting models. The following snippet fits Model 3.7 to the hot dog data. The syntax is similar for many model fitting commands in R, so it is worth spending some time to understand it.

```
hotdogs.fit <- lm (hotdogs$Calories ~ hotdogs>Type)
```

- `lm` stands for linear model.
- `~` stands for “is a function of”. It is used in many of R’s modelling commands. `y ~ x` is called a *formula* and means that `y` is modelled as a function of `x`. In the case at hand, Calories is modelled as a function of Type.
- `lm` specifies the type of model.
- R automatically creates the `X` matrix in Equation 3.12 and estimates the parameters.
- The result of fitting the model is stored in a new object called `hotdogs.fit`. Of course we could have called it anything we like.
- `lm` can have an argument `data`, which specifies a dataframe. So instead of

```
hotdogs.fit <- lm (hotdogs$Calories ~ hotdogs>Type)
```

we could have written

```
hotdogs.fit <- lm (Calories ~ Type, data=hotdogs)
```

You may want to try this to see how it works.

To see `hotdogs.fit`, use R’s `summary` function. Its use and the resulting output are shown in the following snippet.

```
> summary(hotdogs.fit)

Call:
lm(formula = hotdogs$Calories ~ hotdogs>Type)

Residuals:
 Min 1Q Median 3Q Max
-51.706 -18.492 -5.278 22.500 36.294

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.0000 1.0000 10.000 0.0000000 ***
Type -1.0000 0.5000 -2.000 0.0468750 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Intercept) 156.850 5.246 29.901 < 2e-16 ***
hotdogs$TypeMeat 1.856 7.739 0.240 0.811
hotdogs$TypePoultry -38.085 7.739 -4.921 9.4e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.46 on 51 degrees of freedom
Multiple R-Squared: 0.3866, Adjusted R-squared: 0.3626
F-statistic: 16.07 on 2 and 51 DF, p-value: 3.862e-06
```

The most important part of the output is the table labelled **Coefficients**:. There is one row of the table for each coefficient. Their names are on the left. In this table the names are **Intercept**, **hotdogs\$TypeMeat**, and **hotdogs\$TypePoultry**. The first column is labelled **Estimate**. Those are the m.l.e.'s. R has fit the model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$

where  $X_1$  and  $X_2$  are indicator variables for the type of hot dog. The model implies

$$\begin{aligned} Y_i &= \beta_0 + \epsilon_i && \text{for beef hotdogs} \\ Y_i &= \beta_0 + \beta_1 + \epsilon_i && \text{for meat hotdogs} \\ Y_i &= \beta_0 + \beta_2 + \epsilon_i && \text{for poultry hotdogs} \end{aligned}$$

Therefore the names mean

$$\begin{aligned} \beta_0 &= \text{Intercept} && = \text{mean calorie content of beef hot dogs} \\ \beta_1 &= \text{hotdogs$TypeMeat} && = \text{mean difference between beef and meat hot dogs} \\ \beta_2 &= \text{hotdogs$TypePoultry} && = \text{mean difference between} \\ &&& \text{beef and poultry hot dogs} \end{aligned}$$

From the **Coefficients** table the estimates are

$$\begin{aligned} \hat{\beta}_0 &= 156.850 \\ \hat{\beta}_1 &= 1.856 \\ \hat{\beta}_2 &= -38.085 \end{aligned}$$

The next column of the table is labelled **Std. Error**. It contains the SD's of the estimates. In this case,  $\hat{\beta}_0$  has an SD of about 5.2;  $\hat{\beta}_1$  has an SD of about 7.7, and  $\hat{\beta}_2$

also has an SD of about 7.7. The Central Limit Theorem says that approximately, in large samples

$$\begin{aligned}\hat{\beta}_0 &\sim N(\beta_0, \sigma_{\beta_0}) \\ \hat{\beta}_1 &\sim N(\beta_1, \sigma_{\beta_1}) \\ \hat{\beta}_2 &\sim N(\beta_2, \sigma_{\beta_2})\end{aligned}$$

The SD's in the table are estimates of the SD's in the Central Limit Theorem.

Figure 3.9 plots the likelihood functions. The interpretation is that  $\beta_0$  is likely somewhere around 157, plus or minus about 10 or so;  $\beta_1$  is somewhere around 2, plus or minus about 15 or so; and  $\beta_2$  is somewhere around -38, plus or minus about 15 or so. (Compare to Table 3.3.) In particular, there is no strong evidence that Meat hot dogs have, on average, more or fewer calories than Beef hot dogs; but there is quite strong evidence that Poultry hot dogs have considerably fewer.

Figure 3.9 was produced with the following snippet.

```
m <- c(156.85, 1.856, -38.085)
s <- c(5.246, 7.739, 7.739)

par(mfrow=c(2,2))

x <- seq(m[1]-3*s[1], m[1]+3*s[1], length=40)
plot(x, dnorm(x,m[1],s[1]), type="l",
 xlab=expression(mu), ylab="likelihood", yaxt="n")

x <- seq(m[2]-3*s[2], m[2]+3*s[2], length=40)
plot(x, dnorm(x,m[2],s[2]), type="l",
 xlab=expression(delta[M]), ylab="likelihood", yaxt="n")

x <- seq(m[3]-3*s[3], m[3]+3*s[3], length=40)
plot(x, dnorm(x,m[3],s[3]), type="l",
 xlab=expression(delta[P]), ylab="likelihood", yaxt="n")
```

The summary also gives an estimate of  $\sigma$ . The estimate is labelled Residual

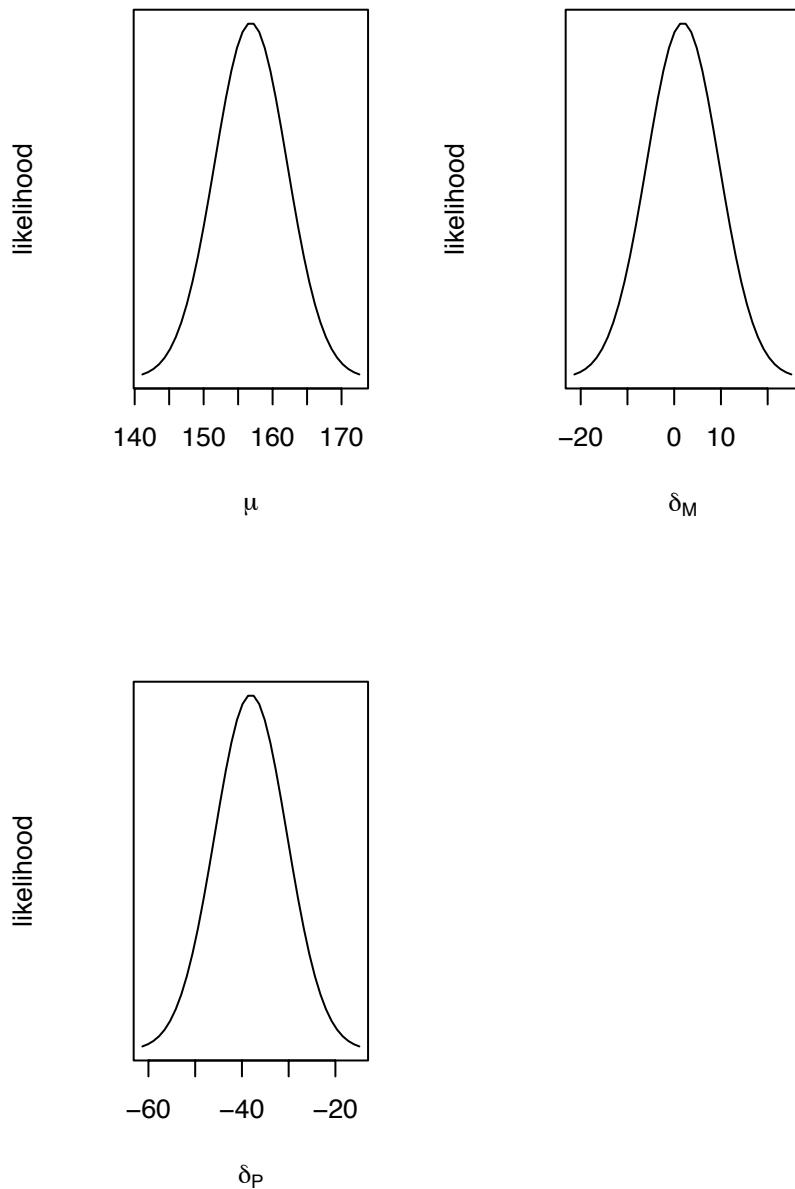


Figure 3.9: Likelihood functions for  $(\mu, \delta_M, \delta_P)$  in the Hot Dog example.

standard error. In this case,  $\hat{\sigma} \approx 23.46$ .<sup>1</sup> So our model says that for each type of hot dog, the calorie contents have approximately a Normal distribution with SD about 23 or so. Compare to Figure 3.5 to see whether the 23.46 makes sense.

Regression is sometimes used in an exploratory setting, when scientists want to find out which variables are related to which other variables. Often there is a response variable  $Y$ , (imagine, for example, performance in school) and they want to know which other variables affect  $Y$  (imagine, for example, poverty, amount of television watching, computer in the home, parental involvement, etc.) Example 3.7 illustrates the process.

### Example 3.7 (mtcars)

This example uses linear regression to explore the R data set `mtcars` (See Figure 3.1, panel (c)) more thoroughly with the goal of modelling `mpg` (miles per gallon) as a function of the other variables. As usual, type `data(mtcars)` to load the data into R and `help(mtcars)` for an explanation. As R explains:

“The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).”

In an exploratory exercise such as this, it often helps to begin by looking at the data. Accordingly, Figure 3.10 is a pairs plot of the data, using just the continuous variables.

Figure 3.10 was produced by

```
pairs (mtcars[,c(1,3:7)])
```

Clearly, `mpg` is related to several of the other variables. Weight is an obvious and intuitive example. The figure suggests that the linear model

$$\text{mpg} = \beta_0 + \beta_1 \text{wt} + \epsilon \quad (3.17)$$

is a good start to modelling the data. Figure 3.11(a) is a plot of `mpg` vs. weight plus the fitted line. The estimated coefficients turn out to be  $\hat{\beta}_0 \approx 37.3$  and  $\hat{\beta}_1 \approx -5.34$ . The interpretation is that `mpg` decreases by about 5.34 for every 1000 pounds of weight. Note: this does not mean that if you put a 1000 pound weight in your car your `mpg`

---

<sup>1</sup>R, like most statistical software, does not report the m.l.e. but reports instead  $\hat{\sigma} \equiv (\sum r_i^2/(n-p-1))^{1/2}$ . Compare to Equation 3.16 for the m.l.e. in which the denominator is  $n$ . The situation is similar to the sample SD on page 97. When  $n \gg p$  there is little difference between the two estimates.

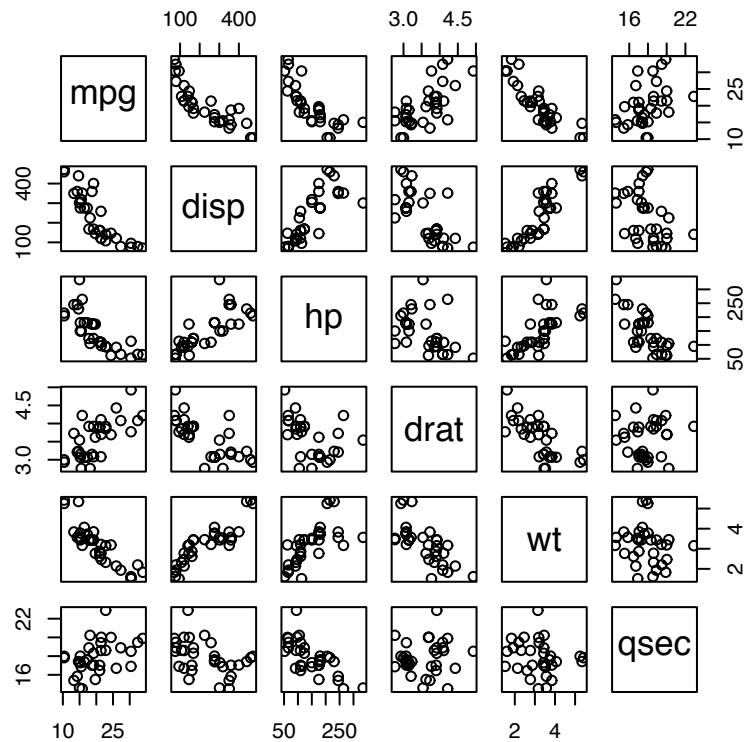


Figure 3.10: `pairs` plot of the `mtcars` data. Type `help(mtcars)` in R for an explanation.

will decrease by 5.34. It means that if car A weighs about 1000 pounds less than car B, then we expect car A to get an extra 5.34 miles per gallon. But there are likely many differences between A and B besides weight. The 5.34 accounts for all of those differences, on average.

We could just as easily have begun by fitting mpg as a function of horsepower with the model

$$\text{mpg} = \gamma_0 + \gamma_1 \text{hp} + \epsilon \quad (3.18)$$

We use  $\gamma$ 's to distinguish the coefficients in Equation 3.18 from those in Equation 3.17. The m.l.e.'s turn out to be  $\hat{\gamma}_0 \approx 30.1$  and  $\hat{\gamma}_1 \approx -0.069$ . Figure 3.11(b) shows the corresponding scatterplot and fitted line. Which model do we prefer? Choosing among different possible models is a major area of statistical practice with a large literature that can be highly technical. In this book we show just a few considerations.

One way to judge models is through *residual plots*, which are plots of residuals versus either  $X$  variables or fitted values. If models are adequate, then residual plots should show no obvious patterns. Patterns in residual plots are clues to model inadequacy and ways to improve models. Figure 3.11(c) and (d) are residual plots for `mpg.fit1` (mpg vs. wt) and `mpg.fit2` (mpg vs. hp). There are no obvious patterns in panel (c). In panel (d) there is a suggestion of curvature. For fitted values between about 15 and 23, residuals tend to be low but for fitted values less than about 15 or greater than about 23, residuals tend to be high (The same pattern might have been noted in panel (b).) suggesting that mpg might be better fit as a nonlinear function of hp. We do not pursue that suggestion further at the moment, merely noting that there may be a minor flaw in `mpg.fit2` and we therefore slightly prefer `mpg.fit1`.

Another thing to note from panels (c) and (d) is the overall size of the residuals. In (c), they run from about -4 to about +6, while in (d) they run from about -6 to about +6. That is, the residuals from `mpg.fit2` tend to be slightly larger in absolute value than the residuals from `mpg.fit1`, suggesting that wt predicts mpg slightly better than does hp. That impression can be confirmed by getting the summary of both fits and checking  $\hat{\sigma}$ . From `mpg.fit1`  $\hat{\sigma} \approx 3.046$  while from `mpg.fit2`  $\hat{\sigma} \approx 3.863$ . I.e., from wt we can predict mpg to within about 6 or so (two SD's) while from hp we can predict mpg only to within about 7.7 or so. For this reason too, we slightly prefer `mpg.fit1` to `mpg.fit2`.

What about the possibility of using both weight and horsepower to predict mpg? Consider

```
mpg.fit3 <- lm (mpg ~ wt + hp, data=mtcars)
```

- The formula  $y \sim x_1 + x_2$  means fit  $y$  as a function of both  $x_1$  and  $x_2$ . In our example that means

$$\text{mpg} = \delta_0 + \delta_1 \text{wt}_1 + \delta_2 \text{hp}_2 + \epsilon \quad (3.19)$$

A residual plot from model 3.19 is shown in Figure 3.11 (e). The m.l.e.'s are  $\hat{\delta}_0 \approx 37.2$ ,  $\hat{\delta}_1 \approx -3.88$ ,  $\hat{\delta}_2 \approx -0.03$ , and  $\hat{\sigma} \approx 2.6$ . Since the residual plot looks curved, Model 3.17 has residuals about as small as Model 3.19, and Model 3.17 is more parsimonious than Model 3.19 we slightly prefer Model 3.17.

Figure 3.11 (a) was produced with

```
plot (mtcars$wt, mtcars$mpg, xlab="weight", ylab="mpg")
mpg.fit1 <- lm (mpg ~ wt, data=mtcars)
abline (coef(mpg.fit1))
```

Figure 3.11, panels (c) and (d) were produced with

```
panel c
plot (fitted(mpg.fit1), resid(mpg.fit1), main="(c)",
 xlab="fitted values from fit1", ylab="resid")
panel d
plot (fitted(mpg.fit2), resid(mpg.fit2),
 xlab="fitted values from fit2", ylab="resid",
 main="(d)")
```

In Example 3.7 we fit three models for mpg, repeated here with their original equation numbers.

$$\text{mpg} = \beta_0 + \beta_1 \text{wt} + \epsilon \quad (3.17)$$

$$\text{mpg} = \gamma_0 + \gamma_1 \text{hp} + \epsilon \quad (3.18)$$

$$\text{mpg} = \delta_0 + \delta_1 \text{wt}_1 + \delta_2 \text{hp} + \epsilon \quad (3.19)$$

What is the connection between, say,  $\beta_1$  and  $\delta_1$ , or between  $\gamma_1$  and  $\delta_2$ ?  $\beta_1$  is the average mpg difference between two cars whose weights differ by 1000 pounds. Since heavier cars tend to be different than lighter cars in many ways, not just in weight,  $\beta_1$  captures

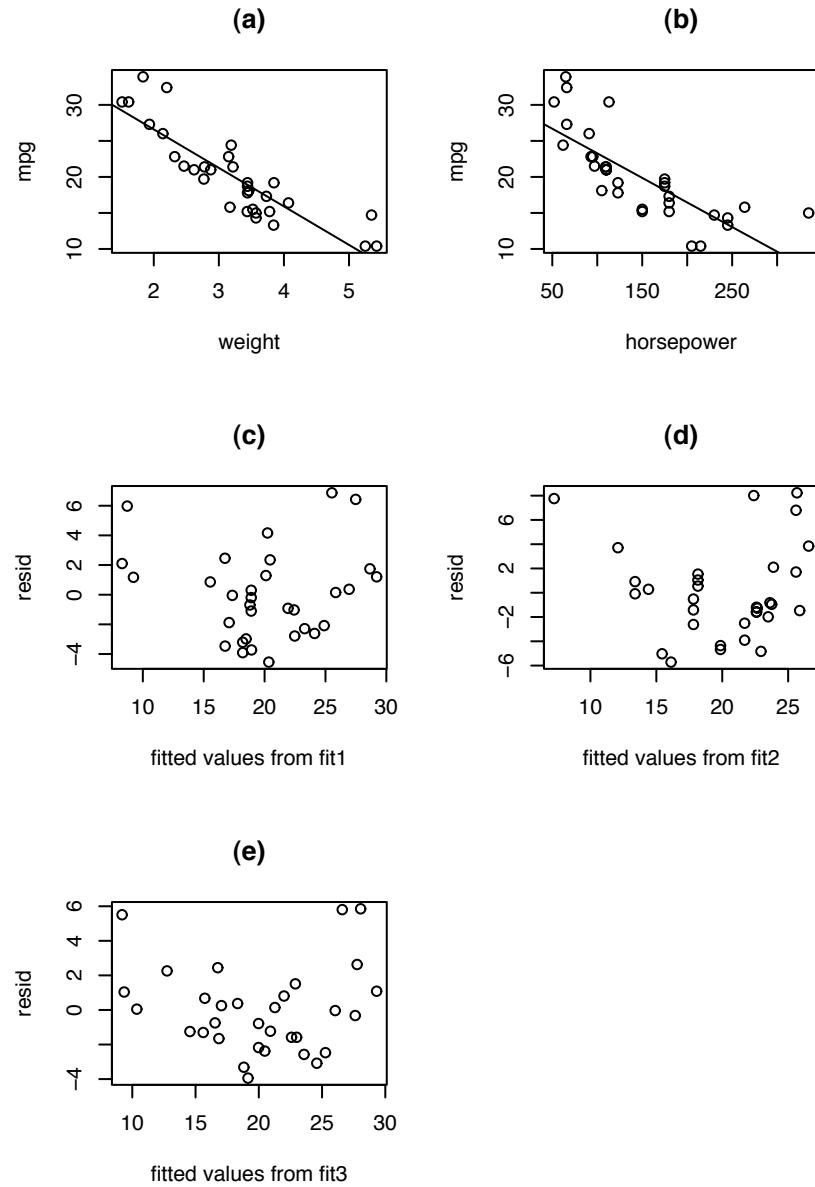


Figure 3.11: `mtcars` — (a): `mpg` vs. `wt`; (b): `mpg` vs. `hp`; (c): residual plot from `mpg ~ wt`; (d): residual plot from `mpg ~ hp`; (e): residual plot from `mpg ~ wt+hp`

the net effect on mpg of all those differences. On the other hand,  $\delta_1$  is the average mpg difference between two cars of identical horsepower but whose weights differ by 1000 pounds. Figure 3.12 shows the likelihood functions of these four parameters. The evidence suggests that  $\beta_1$  is probably in the range of about -7 to about -4, while  $\delta_1$  is in the range of about -6 to -2. It's possible that  $\beta_1 \approx \delta_1$ . On the other hand,  $\gamma_1$  is probably in the interval  $(-.1, -.04)$  while  $\delta_2$  is probably in the interval  $(-.05, 0)$ . It's quite likely that  $\gamma_1 \not\approx \delta_2$ . Scientists sometimes ask the question “What is the effect of variable  $X$  on variable  $Y$ ?” That question does not have an unambiguous answer; the answer depends on which other variables are accounted for and which are not.

Figure 3.12 was produced with

```
par (mfrow=c(2,2))
x <- seq (-8, -1.5, len=60)
plot (x, dnorm(x,-5.3445,.5591), type="l",
 xlab=expression(beta[1]), ylab="", yaxt="n")
x <- seq (-.1, 0, len=60)
plot (x, dnorm(x,-.06823,.01012), type="l",
 xlab=expression(gamma[1]), ylab="", yaxt="n")
x <- seq (-8, -1.5, len=60)
plot (x, dnorm(x,-3.87783,.63273), type="l",
 xlab=expression(delta[1]), ylab="", yaxt="n")
x <- seq (-.1, 0, len=60)
plot (x, dnorm(x,-.03177,.00903), type="l",
 xlab=expression(delta[2]), ylab="", yaxt="n")
```

## 3.3 Generalized Linear Models

### 3.3.1 Logistic Regression

Look again at panel (c) in Figure 3.1 on page 200. The dependent variable is binary, as opposed to the continuous dependent variables in panels (a), (b) and (d). In (a), (b) and (d) we modelled  $Y|X$  as having a Normal distribution; regression was a model for  $\mathbb{E}[Y|X]$ , the mean of that Normal distribution as a function of  $X$ . In (c)  $Y|X$  has a Binomial distribution. We still use the term “regression” for a model of  $\mathbb{E}[Y|X]$ . When  $Y$  is binary, regression is a model for the probability of success  $\theta$  as a function of  $X$ .

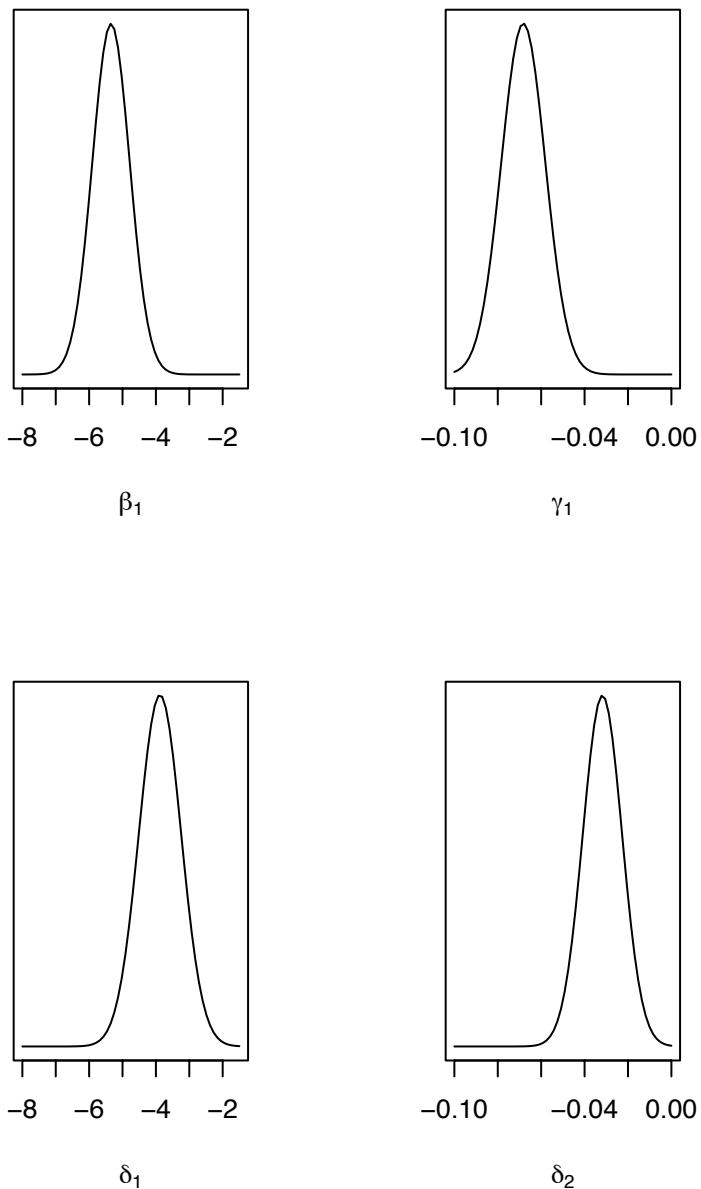


Figure 3.12: likelihood functions for  $\beta_1$ ,  $\gamma_1$ ,  $\delta_1$  and  $\delta_2$  in the `mtcars` example.

Figure 3.13 shows two more scatterplots where  $Y$  is a binary variable. The data are described in the next two examples.

**Example 3.8** (FACE, continued)

Refer to Examples 1.12 and 2.11 about the FACE experiment to assess the effects of excess  $\text{CO}_2$  on the growth of the forest. To describe the size of trees, ecologists sometimes use Diameter at Breast Height, or DBH. DBH was recorded every year for each loblolly pine tree in the FACE experiment. One potential effect of elevated  $\text{CO}_2$  is for the trees to reach sexual maturity and hence be able to reproduce earlier than otherwise. If they do mature earlier, ecologists would like to know whether that's due only to their increased size, or whether trees will reach maturity not just at younger ages, but also at smaller sizes. Sexually mature trees can produce pine cones but immature trees cannot. So to investigate sexual maturity, a graduate student counted the number of pine cones on each tree. For each tree let  $X$  be its DBH and  $Y$  be either 1 or 0 according to whether the tree has pine cones.

Figure 3.13(a) is a plot of  $Y$  versus  $X$  for all the trees in Ring 1. It does appear that larger trees are more likely to have pine cones.

**Example 3.9** (O-rings)

"On January 28, 1986 America was shocked by the destruction of the space shuttle Challenger, and the death of its seven crew members." So begins the website [HTTP://WWW.FAS.ORG/SPP/51L.HTML](http://www.fas.org/spp/51L.html) of the Federation of American Scientist's *Space Policy Project*.

Up until 1986 the space shuttle orbiter was lifted into space by a pair of booster rockets, one on each side of the shuttle, that were comprised of four sections stacked vertically on top of each other. The joints between the sections were sealed by O-rings. On January 28, 1986 the temperature at launch time was so cold that the O-rings became brittle and failed to seal the joints, allowing hot exhaust gas to come into contact with unburned fuel. The result was the Challenger disaster. An investigation ensued. The website [HTTP://SCIENCE.KSC.NASA.GOV/SHUTTLE/MISSIONS/51-L/DOCS](http://science.ksc.nasa.gov/shuttle/missions/51-L/docs) contains links to

1. a description of the event,
2. a report (*Kerwin*) on the initial attempt to determine the cause,
3. a report (*rogers-commission*) of the presidential investigative commission that finally did determine the cause, and
4. a transcript of the operational recorder voice tape.

One of the issues was whether NASA could or should have foreseen that cold weather might diminish performance of the O-rings.

After launch the booster rockets detach from the orbiter and fall into the ocean where they are recovered by NASA, taken apart and analyzed. As part of the analysis NASA records whether any of the O-rings were damaged by contact with hot exhaust gas. If the probability of damage is greater in cold weather then, in principle, NASA might have foreseen the possibility of the accident which occurred during a launch much colder than any previous launch.

Figure 3.13(b) plots  $Y$  = presence of damage against  $X$  = temperature for the launches prior to the Challenger accident. The figure does suggest that colder launches are more likely to have damaged O-rings. What is wanted is a model for probability of damage as a function of temperature, and a prediction for probability of damage at 37°F, the temperature of the Challenger launch.

Fitting straight lines to Figure 3.13 doesn't make sense. In panel (a) what we need is a curve such that

1.  $\mathbb{E}[Y|X] = P[Y = 1|X]$  is close to 0 when  $X$  is smaller than about 10 or 12 cm., and
2.  $\mathbb{E}[Y|X] = P[Y = 1|X]$  is close to 1 when  $X$  is larger than about 25 or 30 cm.

In panel (b) we need a curve that goes in the opposite direction.

The most commonly adopted model in such situations is

$$\mathbb{E}[Y|X] = P[Y = 1|X] = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (3.20)$$

Figure 3.14 shows the same data as Figure 3.13 with some curves added according to Equation 3.20. The values of  $\beta_0$  and  $\beta_1$  are in Table 3.3.1.

|     |        | $\beta_0$ | $\beta_1$ |
|-----|--------|-----------|-----------|
| (a) | solid  | -8        | .45       |
|     | dashed | -7.5      | .36       |
|     | dotted | -5        | .45       |
| (b) | solid  | 20        | -.3       |
|     | dashed | 15        | -.23      |
|     | dotted | 18        | -.3       |

Table 3.2:  $\beta$ 's for Figure 3.14

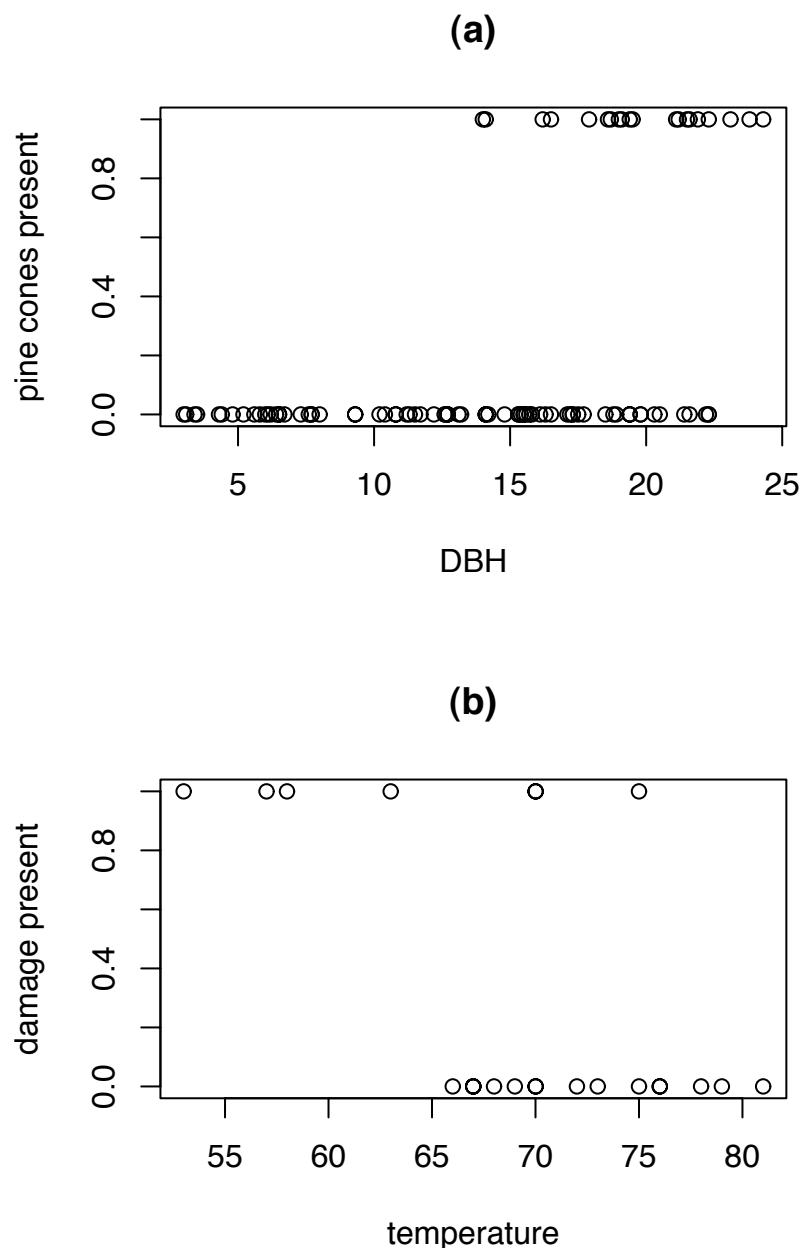


Figure 3.13: (a): pine cone presence/absence vs. dbh. (b): O-ring damage vs. launch temperature

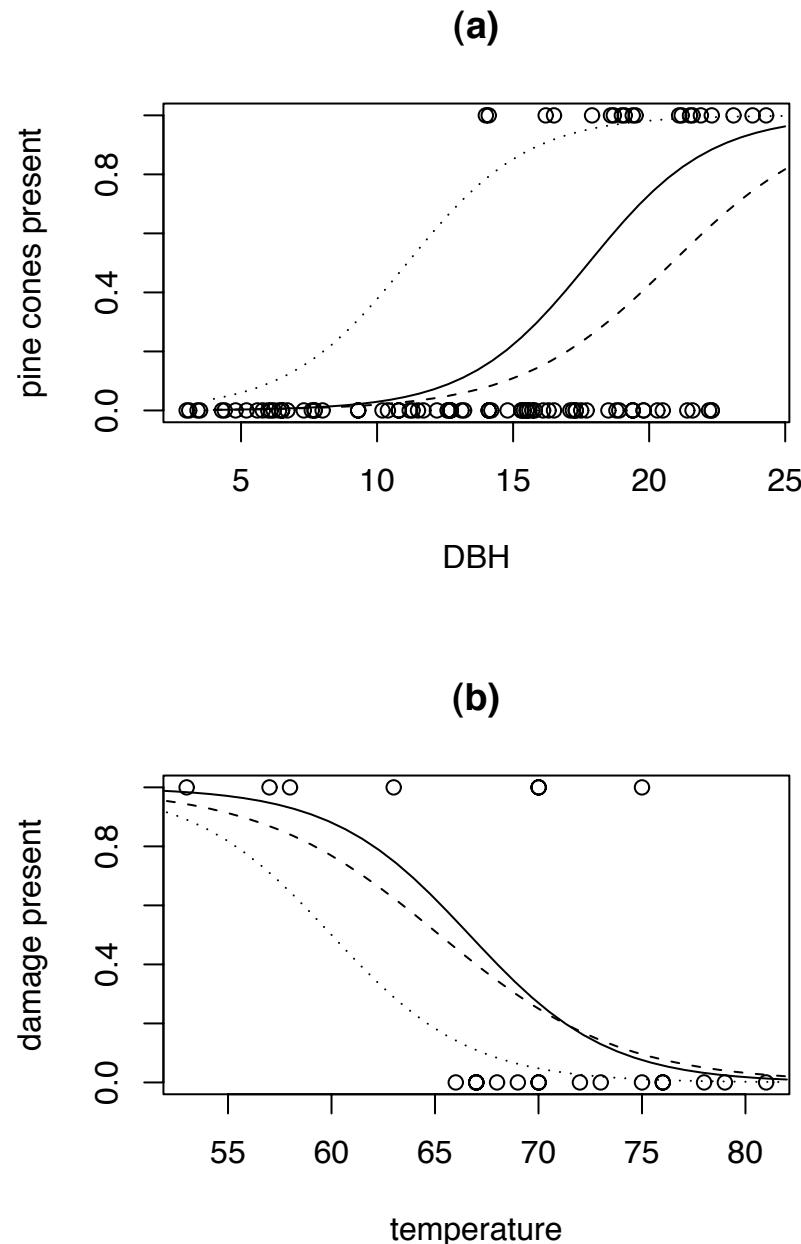


Figure 3.14: (a): pine cone presence/absence vs. dbh. (b): O-ring damage vs. launch temperature, with some logistic regression curves

Figure 3.14 was produced by the following snippet.

```

par (mfrow=c(2,1))
plot (cones$dbh[ring1], mature[ring1], xlab="DBH",
 ylab="pine cones present", main="(a)")
x <- seq (4, 25, length=40)
b0 <- c(-8, -7.5, -5)
b1 <- c (.45, .36, .45)
for (i in 1:3)
 lines (x, exp(b0[i] + b1[i]*x)/(1 + exp(b0[i] + b1[i]*x)) ,
 lty=i)

plot (orings$temp, orings$damage>0, xlab="temperature",
 ylab="damage present", main="(b)")
x <- seq (50, 82, length=40)
b0 <- c(20, 15, 18)
b1 <- c (-.3, -.23, -.3)
for (i in 1:3)
 lines (x, exp(b0[i] + b1[i]*x)/(1 + exp(b0[i] + b1[i]*x)) ,
 lty=i)

```

Model 3.20 is known as *logistic regression*. Let the  $i$ 'th observation have covariate  $x_i$  and probability of success  $\theta_i = \mathbb{E}[Y_i | x_i]$ . Define

$$\phi_i \equiv \log\left(\frac{\theta_i}{1 - \theta_i}\right).$$

$\phi_i$  is called the *logit* of  $\theta_i$ . The inverse transformation is

$$\theta_i = \frac{e^{\phi_i}}{1 + e^{\phi_i}}.$$

The logistic regression model is

$$\phi_i = \beta_0 + \beta_1 x_i.$$

This is called a *generalized linear model* or *glm* because it is a linear model for  $\phi$ , a transformation of  $\mathbb{E}(Y|x)$  rather than for  $\mathbb{E}(Y|x)$  directly. The quantity  $\beta_0 + \beta_1 x$  is called the *linear predictor*. If  $\beta_1 > 0$ , then as  $x \rightarrow +\infty$ ,  $\theta \rightarrow 1$  and as  $x \rightarrow -\infty$ ,  $\theta \rightarrow 0$ . If  $\beta_1 < 0$  the situation is reversed.  $\beta_0$  is like an intercept; it controls how far to the left or

right the curve is.  $\beta_1$  is like a slope; it controls how quickly the curve moves between its two asymptotes.

Logistic regression and, indeed, all generalized linear models differ from linear regression in two ways: the regression function is nonlinear and the distribution of  $Y|x$  is not Normal. These differences imply that the methods we used to analyze linear models are not correct for generalized linear models. We need to derive the likelihood function and find new calculational algorithms.

The likelihood function is derived from first principles.

$$\begin{aligned}
 p(Y_1, \dots, Y_n | x_1, \dots, x_n, \beta_0, \beta_1) &= \prod_i p(Y_i | x_i, \beta_0, \beta_1) \\
 &= \prod_i \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \\
 &= \prod_{i:y_i=1} \theta_i \prod_{i:y_i=0} (1 - \theta_i) \\
 &= \prod_{i:y_i=1} \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \prod_{i:y_i=0} \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}
 \end{aligned}$$

This is a rather complicated function of the two variables  $(\beta_0, \beta_1)$ . However, a Central Limit Theorem applies to give a likelihood function for  $\beta_0$  and  $\beta_1$  that is accurate when  $n$  is reasonable large. The theory is beyond the scope of this book, but R will do the calculations for us. We illustrate with the pine cone data from Example 3.8. Figure 3.15 shows the likelihood function.

Figure 3.15 was produced by the following snippet.

```

mature <- cones$X2000[ring1] > 0
b0 <- seq (-11, -4, length=60)
b1 <- seq (.15, .5, length=60)
lik <- matrix (NA, 60, 60)
for (i in 1:60)
for (j in 1:60) {
 linpred <- b0[i] + b1[j]*cones$dbh[ring1]
 theta <- exp(linpred) / (1+exp(linpred))
 lik[i,j] <- prod (theta^mature * (1-theta)^(1-mature))
}
lik <- lik/max(lik)
contour (b0, b1, lik, xlab=expression(beta[0]),

```

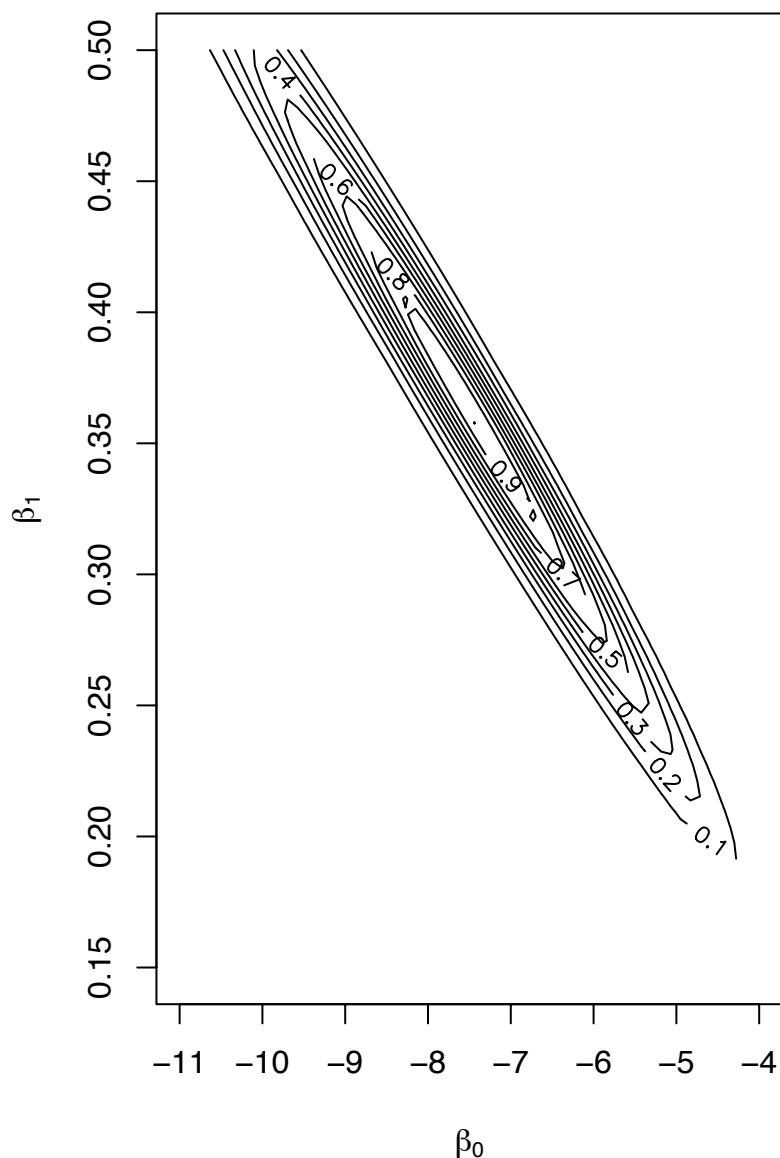


Figure 3.15: Likelihood function for the pine cone data

```
ylab=expression(beta[1]))
```

- `mature` is an indicator variable for whether a tree has at least one pine cone.
- The lines `b0 <- ...` and `b1 <- ...` set some values of  $(\beta_0, \beta_1)$  at which to evaluate the likelihood. They were chosen after looking at the output from fitting the logistic regression model.
- `lik <- ...` creates a matrix to hold values of the likelihood function.
- `linpred` is the linear predictor. Because `cones$dbh[ring1]` is a vector, `linpred` is also a vector. Therefore `theta` is also a vector, as is `theta^mature * (1-theta)^(1-mature)`. It will help your understanding of R to understand what these vectors are.

One notable feature of Figure 3.15 is the diagonal slope of the contour ellipses. The meaning is that we do not have independent information about  $\beta_0$  and  $\beta_1$ . For example if we thought, for some reason, that  $\beta_0 \approx -9$ , then we could be fairly confident that  $\beta_1$  is in the neighborhood of about .4 to about .45. But if we thought  $\beta_0 \approx -6$ , then we would believe that  $\beta_1$  is in the neighborhood of about .25 to about .3. More generally, if we knew  $\beta_0$ , then we could estimate  $\beta_1$  to within a range of about .05. But since we don't know  $\beta_0$ , we can only say that  $\beta_1$  is likely to be somewhere between about .2 and .6. The dependent information for  $(\beta_0, \beta_1)$  means that our marginal information for  $\beta_1$  is much less precise than our conditional information for  $\beta_1$  given  $\beta_0$ . That imprecise marginal information is reflected in the output from R, shown in the following snippet which fits the model and summarizes the result.

```
cones <- read.table ("data/pinecones.dat", header=T)
ring1 <- cones$ring == 1
mature <- cones$X2000[ring1] > 0

fit <- glm (mature ~ cones$dbh[ring1], family=binomial)
summary (fit)

...
Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.46684 1.76004 -4.242 2.21e-05 ***
cones$dbh[ring1] 0.36151 0.09331 3.874 0.000107 ***
```

...

- `cones` ... reads in the data. There is one line for each tree. The first few lines look like this.

| ring | ID      | xcoor | ycoor | spec | dbh  | X1998 | X1999 | X2000 |
|------|---------|-------|-------|------|------|-------|-------|-------|
| 1    | 1 11003 | 0.71  | 0.53  | pita | 19.4 | 0     | 0     | 0     |
| 2    | 1 11004 | 1.26  | 2.36  | pita | 14.1 | 0     | 0     | 4     |
| 3    | 1 11011 | 1.44  | 6.16  | pita | 19.4 | 0     | 6     | 0     |

`ID` is a unique identifying number for each tree; `xcoor` and `ycoor` are coordinates in the plane; `spec` is the species; `pita` stands for *pinus taeda* or loblolly pine, `X1998`, `X1999` and `X2000` are the numbers of pine cones each year.

- `ring1` ... is an indicator variable for trees in Ring 1.
- `mature` ... indicates whether the tree had any cones at all in 2000. It is not a precise indicator of maturity.
- `fit` ... fits the logistic regression. `glm` fits a generalized linear model. The argument `family=binomial` tells R what kind of data we have. In this case it's binomial because `y` is either a success or failure.
- `summary(fit)` shows that  $(\hat{\beta}_0, \hat{\beta}_1) \approx (-7.5, 0.36)$ . The SD's are about 1.8 and .1. These values guided the choice of `b0` and `b1` in creating Figure 3.15. It's the SD of about .1 that says we can estimate  $\beta_1$  to within an interval of about .4, or about  $\pm 2\text{SD}$ 's.

### 3.3.2 Poisson Regression

Section 3.3.1 dealt with the case where the response variable  $Y$  was Bernoulli. Another common situation is where the response  $Y$  is a count. In that case it is natural to adopt, at least provisionally, a model in which  $Y$  has a Poisson distribution:  $Y \sim \text{Poi}(\lambda)$ . When there are covariates  $X$ , then  $\lambda$  may depend on  $X$ . It is common to adopt the regression

$$\log \lambda = \beta_0 + \beta_1 x \tag{3.21}$$

Model 3.21 is another example of a generalized linear model. Example 3.10 illustrates its use.

**Example 3.10** (Seedlings, continued)

Several earlier examples have discussed data from the Ceweeta LTER on the emergence and survival of red maple (*acer rubrum*) seedlings. Example 3.2 showed that the arrival rate of seedlings seemed to vary by quadrat. Refer especially to Figure 3.4. Example 3.10 follows up that observation more quantitatively.

Roughly speaking, New seedlings arise in a two-step process. First, a seed falls out of the sky, then it germinates and emerges from the ground. We may reasonably assume that the emergence of one seedling does not affect the emergence of another (They're too small to interfere with each other.) and hence that the number of New seedlings has a  $\text{Poi}(\lambda)$  distribution. Let  $Y_{ij}$  be the number of New seedlings observed in quadrat  $i$  and year  $j$ . Here are two fits in R, one in which  $\lambda$  varies by quadrat and one in which it doesn't.

```
new <- data.frame (count=count,
 quadrat=as.factor(quadrat),
 year=as.factor(year)
)
fit0 <- glm (count ~ 1, family=poisson, data=new)
fit1 <- glm (count ~ quadrat, family=poisson, data=new)
```

- The command `data.frame` creates a `dataframe`. R describes `dataframes` as

“tightly coupled collections of variables which share many of the properties of matrices and of lists, used as the fundamental data structure by most of R’s modeling software.”

We created a `dataframe` called `new`, having three columns called `count`, `quadrat`, and `year`. Each row of `new` contains a `count` (of New seedlings), a quadrat number and a year. There are as many rows as there are observations.

- The command `as.factor` turns its argument into a `factor`. That is, instead of treating `quadrat` and `year` as numerical variables, we treat them as indicator variables. That’s because we don’t want a quadrat variable running from 1 to 60 implying that the 60th quadrat has 60 times as much of something as the 1st quadrat. We want the quadrat numbers to act as labels, not as numbers.
- `glm` stands for generalized linear model. The `family=poisson` argument says what kind of data we’re modelling. `data=new` says the data are to be found in a `dataframe` called `new`.

- The formula `count ~ 1` says to fit a model with only an intercept, no covariates.
- The formula `count ~ quadrat` says to fit a model in which `quadrat` is a covariate. Of course that's really 59 new covariates, indicator variables for 59 of the 60 quadrats.

To examine the two fits and see which we prefer, we plotted actual versus fitted values and residuals versus fitted values in Figure 3.16. Panels **(a)** and **(b)** are from `fit0`. Because there may be overplotting, we jittered the points and replotted them in panels **(c)** and **(d)**. Panels **(e)** and **(f)** are jittered values from `fit1`. Comparison of panels **(c)** to **(e)** and **(d)** to **(f)** shows that `fit1` predicts more accurately and has smaller residuals than `fit0`. That's consistent with our reading of Figure 3.4. So we prefer `fit1`.

Figure 3.17 continues the story. Panel **(a)** shows residuals from `fit1` plotted against year. There is a clear difference between years. Years 1, 3, and 5 are high while years 2 and 4 are low. So perhaps we should use year as a predictor. That's done by

```
fit2 <- glm (count ~ quadrat+year, family=poisson,
 data=new)
```

Panels **(b)** and **(c)** show diagnostic plots for `fit2`. Compare to similar panels in Figure 3.16 to see whether using year makes an appreciable difference to the fit.

Figure 3.16 was created with the following snippet.

```
par (mfrow=c(3,2))
plot (fitted(fit0), new$count, xlab="fitted values",
 ylab="actual values", main="(a)")
abline (0, 1)
plot (fitted(fit0), residuals(fit0), xlab="fitted values",
 ylab="residuals", main="(b)")
plot (jitter(fitted(fit0)), jitter(new$count),
 xlab="fitted values", ylab="actual values",
 main="(c)")
abline (0, 1)
plot (jitter(fitted(fit0)), jitter(residuals(fit0)),
 xlab="fitted values", ylab="residuals", main="(d)")
plot (jitter(fitted(fit1)), jitter(new$count),
```

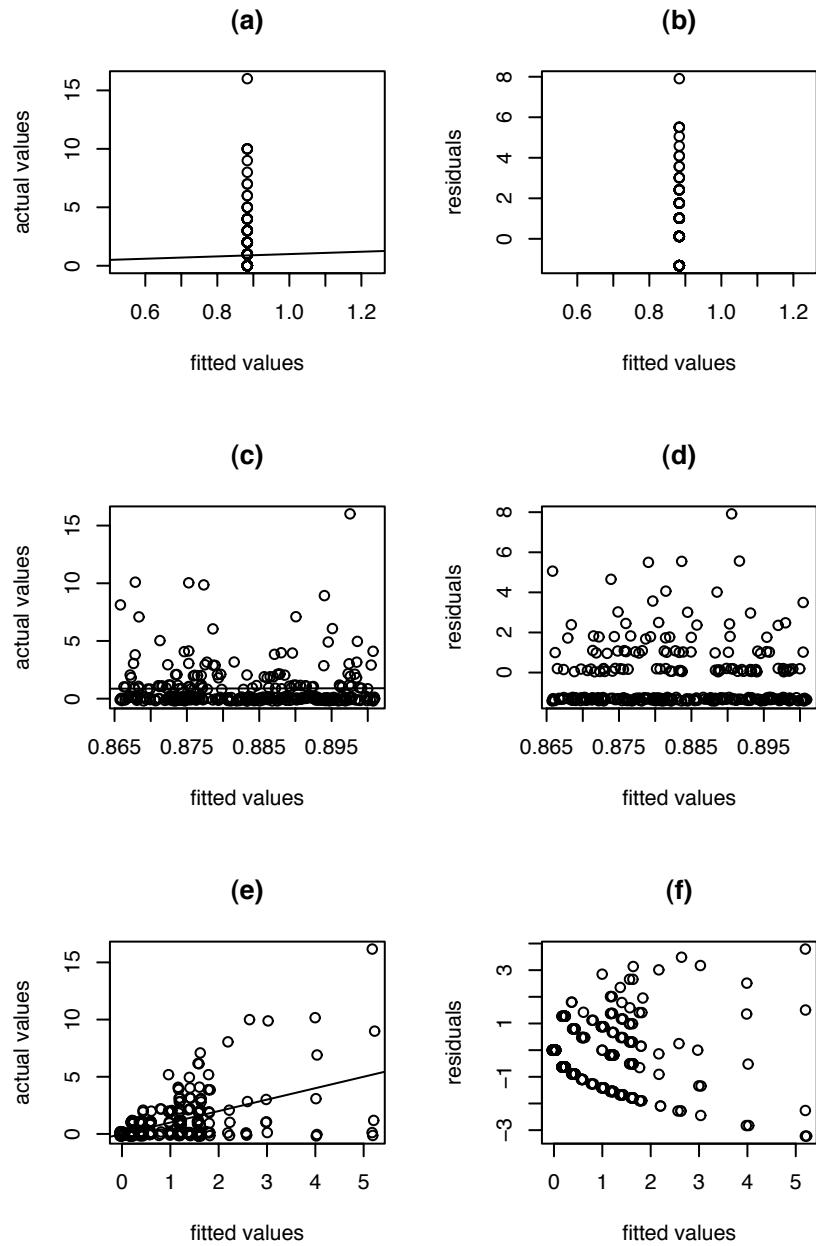


Figure 3.16: Actual vs. fitted and residuals vs. fitted for the New seedling data. (a) and (b): `fit0`. (c) and (d): jittered values from `fit0`. (e) and (f): jittered values from `fit1`.

```

 xlab="fitted values", ylab="actual values",
 main="(e)")
abline (0, 1)
plot (jitter(fitted(fit1)), jitter(residuals(fit1)),
 xlab="fitted values", ylab="residuals", main="(f)")

```

The following snippet shows how Figure 3.17 was made in R.

```

par (mfrow=c(2,2))
plot (new$year, residuals(fit1),
 xlab="year", ylab="residuals", main="(a)")
plot (jitter(fitted(fit2)), jitter(new$count),
 xlab="fitted values", ylab="actual values",
 main="(b)")
abline (0, 1)
plot (jitter(fitted(fit2)), jitter(residuals(fit2)),
 xlab="fitted values", ylab="residuals", main="(c)")

```

## 3.4 Predictions from Regression

From a regression equation, if we have estimates of the  $\beta$ 's we can

1. plug in the values of the  $x$ 's we have to get *fitted values*, and
2. plug in the values of  $x$  for new or future cases to get *predicted values*.

We illustrate with the `mtcars` data.

### Example 3.11 (`mtcars`, continued)

Example 3.7 concluded with a comparison of three models for mpg. Here we continue that comparison by seeing whether the models make substantially different predictions for any of the cars in the data set. For each car we know its weight and horsepower and we have estimates of all the parameters in Equations 3.17, 3.18, and 3.19, so we

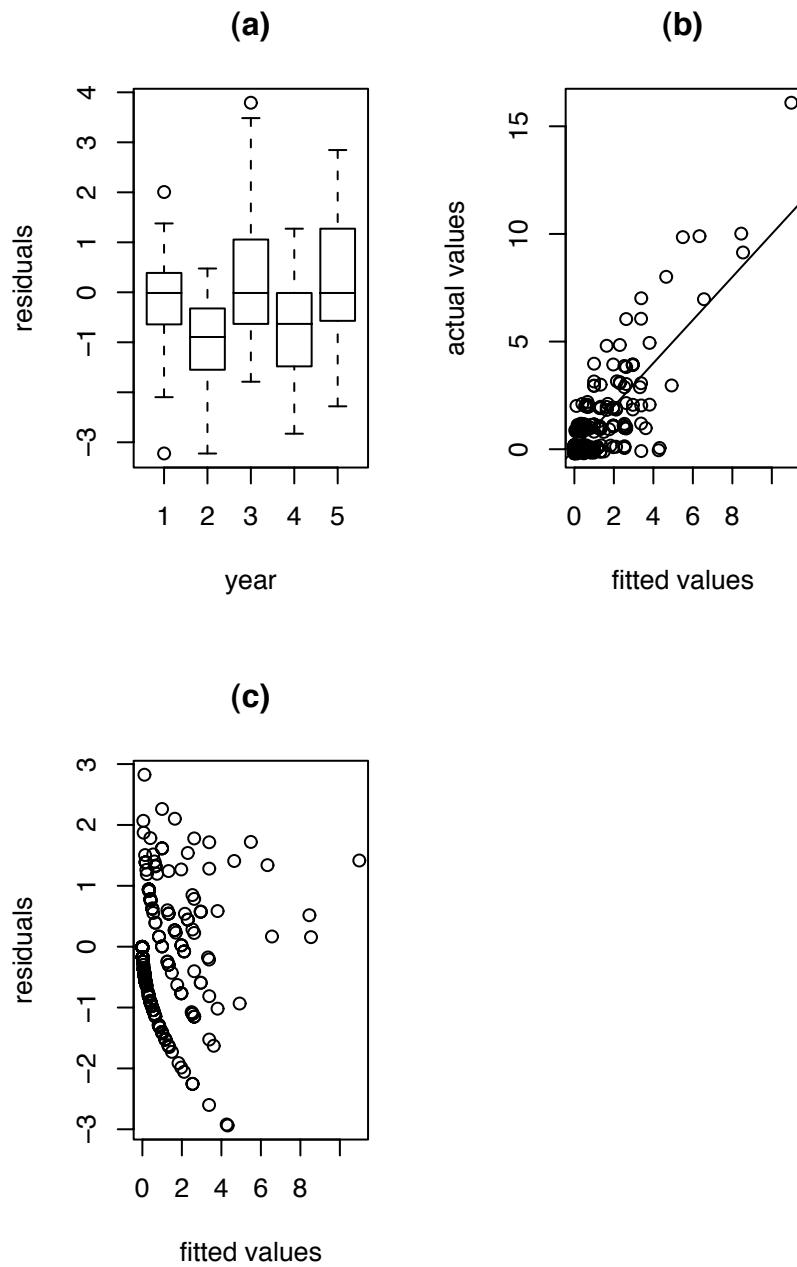


Figure 3.17: New seedling data. (a): residuals from `fit1` vs. year. (b): actual vs. fitted from `fit2`. (c): residuals vs. fitted from `fit2`.

can compute its fitted values from all three models. In symbols,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{wt}_i \quad (\text{from 3.17})$$

$$\hat{y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \text{hp}_i \quad (\text{from 3.18})$$

$$\hat{y}_i = \hat{\delta}_0 + \hat{\delta}_1 \text{wt}_i + \hat{\delta}_2 \text{hp}_i \quad (\text{from 3.19})$$

We plot the fitted values against each other to see whether there are any noticeable differences. Figure 3.18 displays the result. Figure 3.18 shows that the `mpg.fit1` and `mpg.fit3` produce fitted values substantially similar to each other and agreeing fairly well with actual values, while `mpg.fit2` produces fitted values that differ somewhat from the others and from the actual values, at least for a few cars. This is another reason to prefer `mpg.fit1` and `mpg.fit3` to `mpg.fit2`. In Example 3.7 this lack of fit showed up as a higher  $\hat{\sigma}$  for `mpg.fit2` than for `mpg.fit1`.

Figure 3.18 was made with the following snippet.

```
fitted.mpg <- cbind(fitted(mpg.fit1), fitted(mpg.fit2),
 fitted(mpg.fit3), mtcars$mpg)
pairs(fitted.mpg, labels = c("fitted from wt",
 "fitted from hp", "fitted from both", "actual mpg"))
```

- `fitted(xyz)` extracts fitted values. `xyz` can be any model previously fitted by `lm`, `glm`, or other R functions to fit models.

In Example 3.5 we posited model 3.10:

$$y = \beta_0 + \beta_1 x + \epsilon \quad (3.22)$$

where  $x$  was mean temperature during the week and  $y$  was ice cream consumption during the week. Now we want to fit the model to the data and use the fit to predict consumption. In addition, we want to say how accurate the predictions are. Let  $x_f$  be the predicted mean temperature for some future week and  $y_f$  be consumption.  $x_f$  is known;  $y_f$  is not. Our model says

$$y_f \sim N(\mu_f, \sigma)$$

where

$$\mu_f = \beta_0 + \beta_1 x_f$$

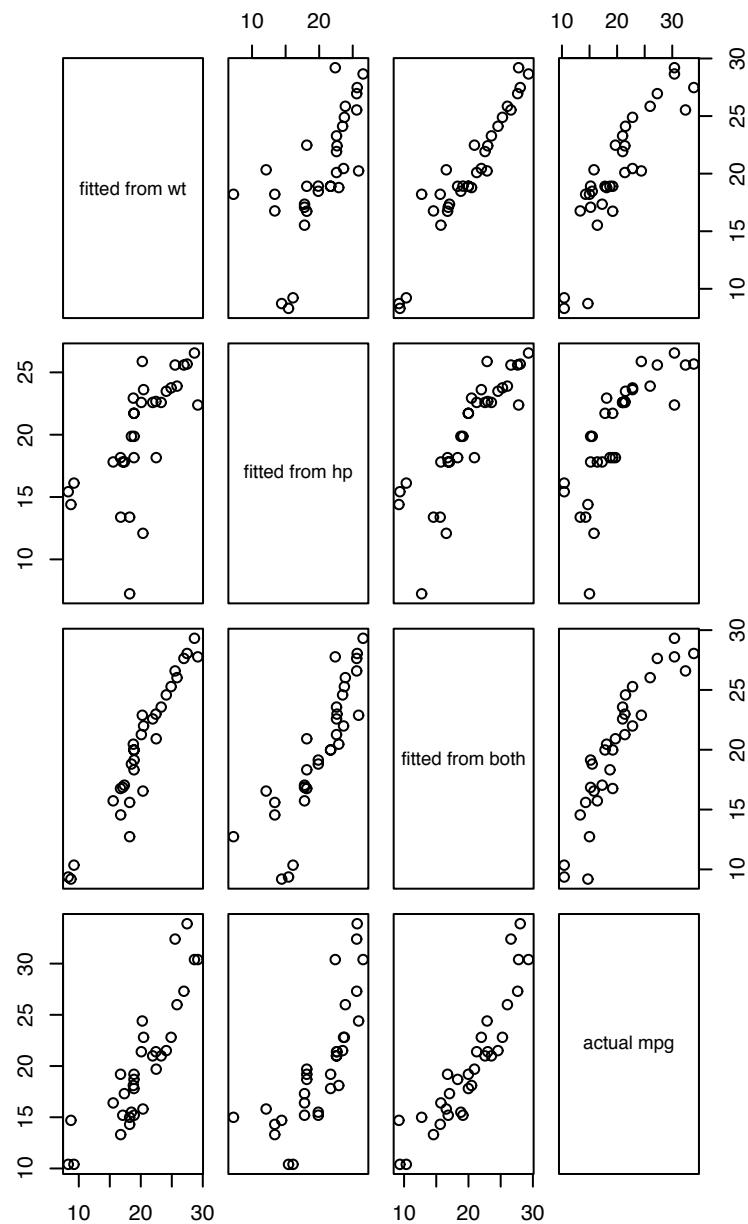


Figure 3.18: Actual mpg and fitted values from three models

$\mu_f$  is unknown because  $\beta_0$  and  $\beta_1$  are unknown. But  $(\beta_0, \beta_1)$  can be estimated from the data, so we can form an estimate

$$\hat{\mu}_f = \hat{\beta}_0 + \hat{\beta}_1 x_f$$

How accurate is  $\hat{\mu}_f$  as an estimate of  $\mu_f$ ? The answer depends on how accurate  $(\hat{\beta}_0, \hat{\beta}_1)$  are as estimates of  $(\beta_0, \beta_1)$ . Advanced theory about Normal distributions, beyond the scope of this book, tells us

$$\hat{\mu}_f \sim N(\mu_f, \sigma_{\text{fit}})$$

for some  $\sigma_{\text{fit}}$  which may depend on  $x_f$ ; we have omitted the dependency from the notation.

$\mu_f$  is the average ice cream consumption in all weeks whose mean temperature is  $x_f$ . So  $\hat{\mu}_f$  is also an estimator of  $y_f$ . But in any particular week the actual consumption won't exactly equal  $\mu_f$ . Our model says

$$y_f = \mu_f + \epsilon$$

where  $\epsilon \sim N(0, \sigma)$ . So in any given week  $y_f$  will differ from  $\mu_f$  by an amount up to about  $\pm 2\sigma$  or so.

Thus the uncertainty  $\sigma_{\text{fit}}$  in estimating  $y_f$  has two components: (1) the uncertainty of  $\mu_f$  which comes because we don't know  $(\beta_0, \beta_1)$  and (2) the variability  $\sigma$  due to  $\epsilon$ . We can't say in advance which component will dominate. Sometimes it will be the first, sometimes the second. What we can say is that as we collect more and more data, we learn  $(\beta_0, \beta_1)$  more accurately, so the first component becomes negligible and the second component dominates. When that happens, we won't go far wrong by simply ignoring the first component.

## 3.5 Exercises

1. (a) Use the `attenu`, `airquality` and `faithful` datasets to reproduce Figures 3.1 (a), (b) and (d).  
 (b) Add `lowess` and `supsmu` fits.  
 (c) Figure out how to use the tuning parameters and try out several different values.  
 (Use the `help` or `help.start` functions.)
2. With the `mtcars` dataset, use a scatterplot smoother to plot the relationship between weight and displacement. Does it matter which we think of as  $X$  and which as  $Y$ ? Is one way more natural than the other?

3. Download the 1970 draft data from DASL and reproduce Figure 3.3. Use the tuning parameters ( $f$  for lowess;  $span$  for `supsmu`) to draw smoother and wigglier scatterplot smoothers.
4. How could you test whether the draft numbers in Example 3.1 were generated uniformly? What would  $H_0$  be? What would be a good test statistic  $w$ ? How would estimate the distribution of  $w$  under  $H_0$ ?
5. Using the information in Example 3.6 estimate the mean calorie content of meat and poultry hot dogs.
6. Refer to Examples 2.2, 3.4, and 3.6.
  - (a) Formulate statistical hypotheses for testing whether the mean calorie content of Poultry hot dogs is equal to the mean calorie content of Beef hot dogs.
  - (b) What statistic will you use?
  - (c) What should that statistic be if  $H_0$  is true?
  - (d) How many SD's is it off?
  - (e) What do you conclude?
  - (f) What about Meat hot dogs?
7. Refer to Examples 2.2, 3.4, and 3.6. Figure 3.5 shows plenty of overlap in the calorie contents of Beef and Poultry hot dogs. I.e., there are many Poultry hot dogs with more calories than many Beef hot dogs. But Figure 3.9 shows very little support for values of  $\delta_P$  near 0. Can that be right? Explain?
8. Examples 2.2, 3.4, and 3.6 analyze the calorie content of Beef, Meat, and Poultry hot dogs. Create a similar analysis, but for sodium content. Your analysis should cover at least the following steps.
  - (a) A stripchart similar to Figure 3.5 and density estimates similar to Figure 3.6.
  - (b) A model similar to Model 3.4, including definitions of the parameters.
  - (c) Indicator variables analogous to those in Equation 3.6.
  - (d) A model similar to Model 3.7, including definitions of all the terms.
  - (e) A fit in R, similar to that in Example 3.6.
  - (f) Parameter estimates and SD's.
  - (g) Plots of likelihood functions, analogous to Figure 3.9.

- (h) Interpretation.
9. Analyze the **PlantGrowth** data from page 211. State your conclusion about whether the treatments are effective. Support you conclusion with analysis.
  10. Analyze the **Ice Cream** data from Example 3.5. Write a model similar to Model 3.7, including definitions of all the terms. Use R to fit the model. Estimate the coefficients and say how accurate your estimates are. If temperature increases by about 5 °F, about how much would you expect ice cream consumption to increase? Make a plot similar to Figure 3.8, but add on the line implied by Equation 3.10 and your estimates of  $\beta_0$  and  $\beta_1$ .
  11. Verify the claim that for Equation 3.18  $\hat{\gamma}_0 \approx 30$ ,  $\hat{\gamma}_1 \approx -.07$  and  $\hat{\sigma} \approx 3.9$ .
  12. Does a football filled with helium travel further than one filled with air? DASL has a data set that attempts to answer the question. Go to DASL, [HTTP://LIB.STAT.CMU.EDU/DASL](http://lib.stat.cmu.edu/DASL), download the data set **Helium football** and read the story. Use what you know about linear models to analyze the data and reach a conclusion. You must decide whether to include data from the first several kicks and from kicks that appear to be flubbed. Does your decision affect your conclusion?
  13. Use the **PlantGrowth** data from R. Refer to page 211 and Equation 3.5.
    - (a) Estimate  $\mu_C$ ,  $\mu_{T1}$ ,  $\mu_{T2}$  and  $\sigma$ .
    - (b) Test the hypothesis  $\mu_{T1} = \mu_C$ .
    - (c) Test the hypothesis  $\mu_{T1} = \mu_{T2}$ .
  14. Jack and Jill, two Duke University sophomores, have to choose their majors. They both love poetry so they might choose to be English majors. Then their futures would be full of black clothes, black coffee, low paying jobs, and occasional volumes of poetry published by independent, non-commercial presses. On the other hand, they both see the value of money, so they could choose to be Economics majors. Then their futures would be full of power suits, double cappuccinos, investment banking and, at least for Jack, membership in the Augusta National golf club. But which would make them more happy?

To investigate, they conduct a survey. Not wanting to embarrass their friends and themselves, Jack and Jill leave Duke's campus and go up Chapel Hill to interview poets and investment bankers at UNC-CH. In all of Chapel Hill there are 90 poets but only 10 investment bankers. J&J interview them all. From the interviews

J&J compute the *Happiness Quotient* or HQ of each subject. The HQ's are in Figure 3.19. J&J also record two indicator variables for each person:  $P_i = 1$  or 0 (for poets and bankers);  $B_i = 1$  or 0 (for bankers and poets).

Jill and Jack each write a statistical model:

$$\text{Jill: } \text{HQ}_i = \alpha_0 + \alpha_1 P_i + \epsilon_i \quad \text{Jack: } \text{HQ}_i = \beta_1 P_i + \beta_2 B_i + \epsilon_i$$

- (a) Say in words what are  $\alpha_0$ ,  $\alpha_1$ ,  $\beta_1$  and  $\beta_2$ .
- (b) Express  $\beta_1$  and  $\beta_2$  in terms of  $\alpha_0$  and  $\alpha_1$ .
- (c) In their data set J&J find  $\overline{HQ} = 43$  among poets,  $\overline{HQ} = 44$  among bankers and  $\hat{\sigma}^2 = 1$ . (Subjects report disappointment with their favorite basketball team as the primary reason for low HQ.) Find sensible numerical estimates of  $\alpha_0$ ,  $\alpha_1$ ,  $\beta_1$  and  $\beta_2$ .

15. Is poverty related to academic performance in school? The file

`schools_poverty`

at this text's website contains relevant data from the Durham, NC school system in 2001. The first few lines are

|   | pfl | eog | type |
|---|-----|-----|------|
| 1 | 66  | 65  | e    |
| 2 | 32  | 73  | m    |
| 3 | 65  | 65  | e    |

Each school in the Durham public school system is represented by one line in the file. The variable `pfl` stands for *percent free lunch*. It records the percentage of the school's student population that qualifies for a free lunch program. It is an indicator of poverty. The variable `eog` stands for *end of grade*. It is the school's average score on end of grade tests and is an indicator of academic success. Finally, `type` indicates the type of school — e, m, or h for elementary, middle or high school, respectively. You are to investigate whether `pfl` is predictive of `eog`.

- (a) Read the data into R and plot it in a sensible way. Use different plot symbols for the three types of schools.

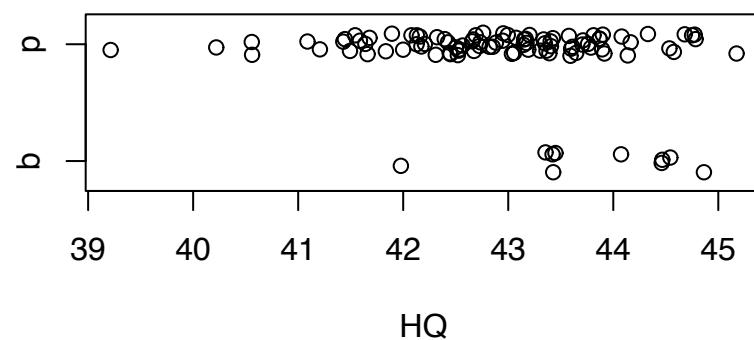


Figure 3.19: Happiness Quotient of bankers and poets

- (b) Does there appear to be a relationship between `pfl` and `eog`? Is the relationship the same for the three types of schools? Decide whether the rest of your analysis should include all types of schools, or only one or two.
- (c) Using the types of schools you think best, remake the plot and add a regression line. Say in words what the regression line means.
- (d) During the 2000-2001 school year Duke University, in Durham, NC, sponsored a tutoring program in one of the elementary schools. Many Duke students served as tutors. From looking at the plot, and assuming the program was successful, can you figure out which school it was?
16. Load `mtcars` into an R session. Use R to find the m.l.e.'s  $(\hat{\beta}_0, \hat{\beta}_1)$ . Confirm that they agree with the line drawn in Figure 3.11(a). Starting from Equation 3.17, derive the m.l.e.'s for  $\beta_0$  and  $\beta_1$ .
17. Get more current data similar to `mtcars`. Carry out a regression analysis similar to Example 3.7. Have relationships among the variables changed over time? What are now the most important predictors of `mpg`?
18. Repeat the logistic regression of `am` on `wt`, but use `hp` instead of `wt`.
19. A researcher randomly selects cities in the US. For each city she records the number of bars  $y_i$  and the number of churches  $z_i$ . In the regression equation  $z_i = \beta_0 + \beta_1 y_i$  do you expect  $\beta_1$  to be positive, negative, or around 0?
20. (a) Jane writes the following R code:

```
x <- runif (60, -1, 1)
```

Describe `x`. Is it a number, a vector, or a matrix? What is in it?

- (b) Now she writes

```
y <- x + rnorm (60)
myfit <- lm (y ~ x)
```

Make an intelligent guess of what she found for  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

- (c) Using advanced statistical theory she calculates

$$\begin{aligned} \text{SD}(\hat{\beta}_0) &= .13 \\ \text{SD}(\hat{\beta}_1) &= .22 \end{aligned}$$

Finally she writes

```
in0 <- 0
in1 <- 0
for (i in 1:100) {
 x <- runif (60, -1, 1)
 y <- x + rnorm (60)
 fit <- lm (y ~ x)
 if (abs(fit$coef[1]) <= .26) in0 <- in0 + 1
 if (abs(fit$coef[2]-1) <= .44) in1 <- in1 + 1
}
```

Make an intelligent guess of `in0` and `in1` after Jane ran this code.

21. The Army is testing a new mortar. They fire a shell up at an angle of  $60^\circ$  and track its progress with a laser. Let  $t_1, t_2, \dots, t_{100}$  be equally spaced times from  $t_1 =$  (time of firing) to  $t_{100} =$  (time when it lands). Let  $y_1, \dots, y_{100}$  be the shell's heights and  $z_1, \dots, z_{100}$  be the shell's distance from the howitzer (measured horizontally along the ground) at times  $t_1, t_2, \dots, t_{100}$ . The  $y_i$ 's and  $z_i$ 's are measured by the laser. The measurements are not perfect; there is some measurement error. In answering the following questions you may assume that the shell's horizontal speed remains constant until it falls to ground.

- (a) **True or False:** The equation

$$y_i = \beta_0 + \beta_1 t_i + \epsilon_i$$

should fit the data well.

- (b) **True or False:** The equation

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \epsilon_i \quad (3.23)$$

should fit the data well.

- (c) **True or False:** The equation

$$z_i = \beta_0 + \beta_1 t_i + \epsilon_i \quad (3.24)$$

should fit the data well.

- (d) **True or False:** The equation

$$z_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \epsilon_i \quad (3.25)$$

should fit the data well.

- (e) **True or False:** The equation

$$y_i = \beta_0 + \beta_1 z_i + \epsilon_i \quad (3.26)$$

should fit the data well.

- (f) **True or False:** The equation

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \epsilon_i \quad (3.27)$$

should fit the data well.

- (g) Approximately what value did the Army find for  $\hat{\beta}_0$  in Part (b)?

- (h) Approximately what value did the Army find for  $\hat{\beta}_2$  in Part (d)?

22. Some nonstatisticians (not readers of this book, we hope) do statistical analyses based almost solely on numerical calculations and don't use plots. R comes with the data set `anscombe` which demonstrates the value of plots. Type `data(anscombe)` to load the data into your R session. It is an 11 by 8 `dataframe`. The variable names are `x1`, `x2`, `x3`, `x4`, `y1`, `y2`, `y3`, and `y4`.

- (a) Start with `x1` and `y1`. Use `lm` to model `y1` as a function of `x1`. Print a summary of the regression so you can see  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and  $\hat{\sigma}$ .
- (b) Do the same for the other pairs: `x2` and `y2`, `x3` and `y3`, `x4` and `y4`.
- (c) What do you conclude so far?
- (d) Plot `y1` versus `x1`. Repeat for each pair. You may want to put all four plots on the same page. (It's not necessary, but you should know how to draw the regression line on each plot. Do you?)
- (e) What do you conclude?
- (f) Are any of these pairs well described by linear regression? How would you describe the others? If the others were not artificially constructed data, but were real, how would you analyze them?

23. Here's some R code:

```
x <- rnorm (1, 2, 3)
y <- -2*x + 1 + rnorm (1, 0, 1)
```

- (a) What is the marginal distribution of  $x$ ?
- (b) Write down the marginal density of  $x$ .
- (c) What is the conditional distribution of  $y$  given  $x$ ?
- (d) Write down the conditional density of  $y$  given  $x$ .
- (e) Write down the joint density of  $(x, y)$ .

Here's more R code:

```
N.sim <- 1000
w <- rep (NA, N.sim)
for (i in 1:N.sim) {
 x <- rnorm (50, 2, 3)
 y <- -2*x + 1 + rnorm (50, 0, 1)
 fit <- lm (y ~ x)
 w[i] <- fit$coef[2]
}
z1 <- mean (w)
z2 <- sqrt (var (w + 2))
```

What does  $z1$  estimate? What does  $z2$  estimate? Why did the code writer write `sqrt ( var ( w + 2 ) )` instead of `sqrt ( var ( w ) )`? Does it matter?

24. A statistician thinks the regression equation  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  fits her data well. She would like to learn  $\beta_1$ . She is able to measure the  $y_i$ 's accurately but can measure the  $x_i$ 's only approximately. In fact, she can measure  $w_i = x_i + \delta_i$  where  $\delta_i \sim N(0, .1)$ . So she can fit the regression equation  $y_i = \beta_0^* + \beta_1^* w_i + \epsilon_i^*$ . Note that  $(\beta_0^*, \beta_1^*)$  might be different than  $(\beta_0, \beta_1)$  because they're for the  $w_i$ 's, not the  $x_i$ 's. So the statistician writes the following R code.

```
N.sim <- 1000
b.0 <- -10:10
b.1 <- -10:10
n <- 50
```

```

for (i in 1:21)
for (j in 1:21) {
 val <- rep (NA, N.sim)
 for (k in 1:N.sim) {
 x <- rnorm (n)
 w <- x + rnorm (n, 0, sqrt(.1))
 y <- b.0[i] + x * b.1[j] + rnorm (n, 0, 3)
 fit <- lm (y ~ w)
 val[k] <- fit$coef[2]
 }
 m <- mean (val)
 sd <- sqrt (var (val))
 print (c (m, sd))
}

```

What is she trying to do? The last time through the loop, the print statement yields  
`[1] 9.0805986 0.5857724`. What does this show?

25. The purpose of this exercise is to familiarize yourself with plotting logistic regression curves and getting a feel for the meaning of  $\beta_0$  and  $\beta_1$ .
  - (a) Choose some values of  $x$ . You will want between about 20 and 100 evenly spaced values. These will become the abscissa of your plot.
  - (b) Choose some values of  $\beta_0$  and  $\beta_1$ . You are trying to see how different values of the  $\beta$ 's affect the curve. So you might begin with a single value of  $\beta_1$  and several values of  $\beta_0$ , or *vice versa*.
  - (c) For each choice of  $(\beta_0, \beta_1)$  calculate the set of  $\theta_i = e^{\beta_0 + \beta_1 x_i} / (1 + e^{\beta_0 + \beta_1 x_i})$  and plot  $\theta_i$  versus  $x_i$ . You should get sigmoidal shaped curves. These are logistic regression curves.
  - (d) You may find that the particular  $x$ 's and  $\beta$ 's you chose do not yield a visually pleasing result. Perhaps all your  $\theta$ 's are too close to 0 or too close to 1. In that case, go back and choose different values. You will have to play around until you find  $x$ 's and  $\beta$ 's compatible with each other.
26. Carry out a logistic regression analysis of the O-ring data. What does your analysis say about the probability of O-ring damage at 36°F, the temperature of the Challenger launch. How relevant should such an analysis have been to the decision of whether to postpone the launch?

27. This exercise refers to Example 3.10.

- (a) Why are the points lined up vertically in Figure 3.16, panels **(a)** and **(b)**?
- (b) Why do panels **(c)** and **(d)** appear to have more points than panels **(a)** and **(b)**?
- (c) If there were no jittering, how many distinct values would there be on the abscissa of panels **(c)** and **(d)**?
- (d) Download the seedling data. Fit a model in which year is a predictor but quadrat is not. Compare to `fit1`. Which do you prefer? Which variable is more important: quadrat or year? Or are they both important?

## CHAPTER 4

# MORE PROBABILITY

## 4.1 More Probability Density

Section 1.2 on page 6 introduced probability densities. Section 4.1 discusses them further and gives a formal definition.

Let  $X$  be a continuous random variable with cdf  $F_X$ . Equation 1.2 on page 8 implies that  $f_X(x) = \frac{d}{db}F_X(b)\Big|_{b=x}$  and therefore that we can define the pdf by  $f_X(x) \equiv F'_X(x) = \frac{d}{db}F_X(b)\Big|_{b=x}$ . In fact, this definition is a little too restrictive. The key property of pdf's is that the probability of a set  $A$  is given by the integral of the pdf. I.e.,

$$P[X \in A] = \int_A f_X(x) dx$$

But if  $f^*$  is a function that differs from  $f_X$  at only countably many points then, for any set  $A$ ,  $\int_A f^* = \int_A f_X$ , so we could just as well have defined

$$P[X \in A] = \int_A f^*(x) dx$$

There are infinitely many functions having the same integrals as  $f_X$  and  $f^*$ . These functions differ from each other on “sets of measure zero”, terminology beyond our scope but defined in books on measure theory. For our purposes we can think of sets of measure zero as sets containing at most countably many points. In effect, the pdf of  $X$  can be arbitrarily changed on sets of measure zero. It does not matter which of the many equivalent functions we use as the probability density of  $X$ . Thus, we define

**Definition 4.1.** Any function  $f$  such that, for all intervals  $A$ ,

$$P[X \in A] = \int_A f(x) dx$$

is called a *probability density function*, or *pdf*, for the random variable  $X$ . Any such function may be denoted  $f_X$ .

Definition 4.1 can be used in an alternate proof of Theorem 1.1 on page 12. The central step in the proof is just a change-of-variable in an integral, showing that Theorem 1.1 is, in essence, just a change of variables. For convenience we restate the theorem before reproving it.

**Theorem 1.1** *Let  $X$  be a random variable with pdf  $p_X$ . Let  $g$  be a differentiable, monotonic, invertible function and define  $Z = g(X)$ . Then the pdf of  $Z$  is*

$$p_Z(t) = p_X(g^{-1}(t)) \left| \frac{d g^{-1}(t)}{dt} \right|$$

*Proof.* For any set  $A$ ,  $P[Z \in g(A)] = P[X \in A] = \int_A p_X(x) dx$ . Let  $z = g(x)$  and change variables in the integral to get

$$P[Z \in g(A)] = \int_{g(A)} p_X(g^{-1}(z)) \left| \frac{dx}{dz} \right| dz$$

I.e.,  $P[Z \in g(A)] = \int_{g(A)} \text{something } dz$ . Therefore something must be  $p_Z(z)$ . Hence,  $p_Z(z) = p_X(g^{-1}(z)) |dx/dz|$ .  $\square$

## 4.2 Random Vectors

It is often useful, even essential, to talk about several random variables simultaneously. We have seen many examples throughout the text beginning with Section 1.5 on joint, marginal, and conditional probabilities. Section 4.2 reviews the basics and sets out new probability theory for multiple random variables.

Let  $X_1, \dots, X_n$  be a set of  $n$  random variables. The  $n$ -dimensional vector  $\vec{X} = (X_1, \dots, X_n)$  is called a multivariate random variable or *random vector*. As explained below,  $\vec{X}$  has a pdf or pmf, a cdf, an expected value, and a covariance matrix, all analogous to univariate random variables.

### 4.2.1 Densities of Random Vectors

When  $X_1, \dots, X_n$  are continuous then  $\vec{X}$  has a pdf, written

$$p_{\vec{X}}(x_1, \dots, x_n).$$

As in the univariate case, the pdf is any function whose integral yields probabilities. That is, if  $A$  is a region in  $\mathbb{R}^n$  then

$$P[\vec{X} \in A] = \int_A \cdots \int p_{\vec{X}}(x_1, \dots, x_n) dx_1 \dots dx_n$$

For example, let  $X_1 \sim \text{Exp}(1)$ ;  $X_2 \sim \text{Exp}(1/2)$ ;  $X_1 \perp X_2$ ; and  $\vec{X} = (X_1, X_2)$  and suppose we want to find  $P[|X_1 - X_2| \leq 1]$ . Our plan for solving this problem is to find the joint density  $p_{\vec{X}}$ , then integrate  $p_{\vec{X}}$  over the region  $A$  where  $|X_1 - X_2| \leq 1$ . Because  $X_1 \perp X_2$ , the joint density is

$$p_{\vec{X}}(x_1, x_2) = p_{X_1}(x_1)p_{X_2}(x_2) = e^{-x_1} \times \frac{1}{2}e^{-x_2/2}$$

To find the region  $A$  over which to integrate, it helps to plot the  $X_1$ - $X_2$  plane. Making the plot is left as an exercise.

$$\begin{aligned} P[|X_1 - X_2| \leq 1] &= \iint_A p_{\vec{X}}(x_1, x_2) dx_1 dx_2 \\ &= \frac{1}{2} \int_0^1 \int_0^{x_1+1} e^{-x_1} e^{-x_2/2} dx_2 dx_1 + \frac{1}{2} \int_1^\infty \int_{x_1-1}^{x_1+1} e^{-x_1} e^{-x_2/2} dx_2 dx_1 \\ &\approx 0.47 \quad (4.1) \end{aligned}$$

The random variables  $(X_1, \dots, X_n)$  are said to be *mutually independent* or *jointly independent* if

$$p_{\vec{X}}(x_1, \dots, x_n) = p_{X_1}(x_1) \times \cdots \times p_{X_n}(x_n)$$

for all vectors  $(x_1, \dots, x_n)$ .

Mutual independence implies pairwise independence. I.e., if  $(X_1, \dots, X_n)$  are mutually independent, then any pair  $(X_i, X_j)$  are also independent. The proof is left as an exercise. It is curious but true that pairwise independence does not imply joint independence. For an example, consider the discrete three-dimensional distribution on  $\vec{X} = (X_1, X_2, X_3)$  with

$$\begin{aligned} P[(X_1, X_2, X_3) = (0, 0, 0)] \\ = P[(X_1, X_2, X_3) = (1, 0, 1)] \\ = P[(X_1, X_2, X_3) = (0, 1, 1)] \\ = P[(X_1, X_2, X_3) = (1, 1, 0)] = 1/4 \end{aligned} \quad (4.2)$$

It is easily verified that  $X_1 \perp X_2$ ,  $X_1 \perp X_3$ , and  $X_2 \perp X_3$  but that  $X_1$ ,  $X_2$ , and  $X_3$  are not mutually independent. See Exercise 6.

### 4.2.2 Moments of Random Vectors

When  $\vec{X}$  is a random vector, its expected value is also a vector.

$$\mathbb{E}[\vec{X}] \equiv (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])$$

When  $\vec{X} \equiv (X_1, \dots, X_n)$  is a random vector, instead of a variance it has a *covariance matrix*. The  $ij$ 'th entry of the covariance matrix is  $\text{Cov}(X_i, X_j)$ . The notation is

$$\text{Cov}(\vec{X}) \equiv \Sigma_{\vec{X}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

where  $\sigma_{ij} = \text{Cov}(X_i, X_j)$  and  $\sigma_i^2 = \text{Var}(X_i)$ . Sometimes  $\sigma_i^2$  is also denoted  $\sigma_{ii}$ .

### 4.2.3 Functions of Random Vectors

Section 4.2.3 considers functions of random vectors. If  $g$  is an arbitrary function that maps  $\vec{X}$  to  $\mathbb{R}$  then

$$\mathbb{E}[g(\vec{X})] = \int \cdots \int g(x_1, \dots, x_n) p_{\vec{X}}(x_1, \dots, x_n) dx_1 \cdots dx_n$$

but it's hard to say much in general about the variance of  $g(\vec{X})$ . When  $g$  is a linear function we can go farther, but first we need a lemma.

**Lemma 4.1.** *Let  $X_1$  and  $X_2$  be random variables and  $Y = X_1 + X_2$ . Then*

1.  $\mathbb{E}[Y] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$
2.  $\text{Var}(Y) = \text{Var}(X_1) + \text{Var}(X_2) + 2 \text{Cov}(X_1, X_2)$

*Proof.* Left as exercise. □

Now we can deal with linear combinations of random vectors.

**Theorem 4.2.** *Let  $\vec{a} = (a_1, \dots, a_n)$  be an  $n$ -dimensional vector and define  $Y = \vec{a}' \vec{X} = \sum a_i X_i$ . Then,*

1.  $\mathbb{E}[Y] = \mathbb{E}[\sum a_i X_i] = \sum a_i \mathbb{E}[X_i]$
2.  $\text{Var}(Y) = \sum a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \text{Cov}(X_i, X_j) = \vec{a}' \Sigma_{\vec{X}} \vec{a}$

*Proof.* Use Lemma 4.1 and Theorems 1.3 (pg. 40) and 1.4 (pg. 40). See Exercise 8.  $\square$

The next step is to consider several linear combinations simultaneously. For some  $k \leq n$ , and for each  $i = 1, \dots, k$ , let

$$Y_i = a_{i1}X_1 + \dots + a_{in}X_n = \sum_j a_{ij}X_j = \vec{d}_i^T \vec{X}$$

where the  $a_{ij}$ 's are arbitrary constants and  $\vec{d}_i = (a_{i1}, \dots, a_{in})$ . Let  $\vec{Y} = (Y_1, \dots, Y_k)$ . In matrix notation,

$$\vec{Y} = A\vec{X}$$

where  $A$  is the  $k \times n$  matrix of elements  $a_{ij}$ . Covariances of the  $Y_i$ 's are given by

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= \text{Cov}(\vec{d}_i^T \vec{X}, \vec{d}_j^T \vec{X}) \\ &= \sum_{k=1}^n \sum_{\ell=1}^n \text{Cov}(a_{ik}X_k, a_{j\ell}X_\ell) \\ &= \sum_{k=1}^n a_{ik}a_{jk}\sigma_k^2 + \sum_{k=1}^{n-1} \sum_{\ell=k+1}^n (a_{ik}a_{j\ell} + a_{jk}a_{i\ell})\sigma_{k\ell} \\ &= \vec{d}_i^T \Sigma_{\vec{X}} \vec{d}_j \end{aligned}$$

Combining the previous result with Theorem 4.2 yields Theorem 4.3.

**Theorem 4.3.** *Let  $\vec{X}$  be a random vector of dimension  $n$  with mean  $\mathbb{E}[\vec{X}] = \mu$  and covariance matrix  $\text{Cov}(\vec{X}) = \Sigma$ ; let  $A$  be a  $k \times n$  matrix of rank  $k$ ; and let  $\vec{Y} = A\vec{X}$ . Then*

1.  $\mathbb{E}[\vec{Y}] = A\mu$ , and

2.  $\text{Cov}(\vec{Y}) = A\Sigma A'$

Finally, we take up the question of multivariate transformations, extending the univariate version, Theorem 1.1 (pg. 12). Let  $\vec{X} = (X_1, \dots, X_n)$  be an  $n$ -dimensional continuous random vector with pdf  $f_{\vec{X}}$ . Define a new  $n$ -dimensional random vector  $\vec{Y} = (Y_1, \dots, Y_n) = (g_1(\vec{X}), \dots, g_n(\vec{X}))$  where the  $g_i$ 's are differentiable functions and where the transformation  $g : \vec{X} \mapsto \vec{Y}$  is invertible. What is  $f_{\vec{Y}}$ , the pdf of  $\vec{Y}$ ?

Let  $J$  be the so-called *Jacobian* matrix of partial derivatives.

$$J = \begin{pmatrix} \frac{\partial Y_1}{\partial X_1} & \dots & \frac{\partial Y_1}{\partial X_n} \\ \frac{\partial Y_2}{\partial X_1} & \dots & \frac{\partial Y_2}{\partial X_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial Y_n}{\partial X_1} & \dots & \frac{\partial Y_n}{\partial X_n} \end{pmatrix}$$

and  $|J|$  be the absolute value of the determinant of  $J$ .

**Theorem 4.4.**

$$f_{\vec{Y}}(\vec{y}) = f_{\vec{X}}(g^{-1}(\vec{y}))|J|^{-1}$$

*Proof.* The proof follows the alternate proof of Theorem 1.1 on page 261. For any set  $A$ ,  $P[\vec{Y} \in g(A)] = P[\vec{X} \in A] = \int \cdots \int_A p_{\vec{X}}(\vec{x}) dx_1 \cdots dx_n$ . Let  $\vec{y} = g(\vec{x})$  and change variables in the integral to get

$$P[\vec{Y} \in g(A)] = \int \cdots \int_{g(A)} p_{\vec{X}}(g^{-1}(\vec{y})) |J|^{-1} dy_1 \cdots dy_n$$

I.e.,  $P[\vec{Y} \in g(A)] = \int \cdots \int_{g(A)} \text{something} dy_1 \cdots dy_n$ . Therefore **something** must be  $p_{\vec{Y}}(\vec{y})$ . Hence,  $p_{\vec{Y}}(\vec{y}) = p_{\vec{X}}(g^{-1}(\vec{y}))|J|^{-1}$ .  $\square$

To illustrate the use of Theorem 4.4 we solve again an example previously given on page 262, which we restate here. Let  $X_1 \sim \text{Exp}(1)$ ;  $X_2 \sim \text{Exp}(2)$ ;  $X_1 \perp X_2$ ; and  $\vec{X} = (X_1, X_2)$  and suppose we want to find  $P[|X_1 - X_2| \leq 1]$ . We solved this problem previously by finding the joint density of  $\vec{X} = (X_1, X_2)$ , then integrating over the region where  $|X_1 - X_2| \leq 1$ . Our strategy this time is to define new variables  $Y_1 = X_1 - X_2$  and  $Y_2$ , which is essentially arbitrary, find the joint density of  $\vec{Y} = (Y_1, Y_2)$ , then integrate over the region where  $|Y_1| \leq 1$ . We define  $Y_1 = X_1 - X_2$  because that's the variable we're interested in. We need a  $Y_2$  because Theorem 4.4 is for full rank transformations from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ . The precise definition of  $Y_2$  is unimportant, as long as the transformation from  $\vec{X}$  to  $\vec{Y}$  is differentiable and invertible. For convenience, we define  $Y_2 = X_2$ . With these definitions,

$$\begin{aligned} J &= \begin{pmatrix} \frac{\partial Y_1}{\partial X_1} & \frac{\partial Y_1}{\partial X_2} \\ \frac{\partial Y_2}{\partial X_1} & \frac{\partial Y_2}{\partial X_2} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \\ |J| &= 1 \\ X_1 &= Y_1 + Y_2 \end{aligned}$$

and

$$X_2 = Y_2$$

From the solution on page 262 we know  $p_{\vec{X}}(x_1, x_2) = e^{-x_1} \times \frac{1}{2}e^{-x_2/2}$ , so  $p_{\vec{Y}}(y_1, y_2) =$

$e^{-(y_1+y_2)} \times \frac{1}{2}e^{-y_2/2} = \frac{1}{2}e^{-y_1}e^{-3y_2/2}$ . Figure 4.1 shows the region over which to integrate.

$$\begin{aligned}
 P[|X_1 - X_2| \leq 1] &= P[|Y_1| \leq 1] = \iint_A p_{\vec{Y}}(y_1, y_2) dy_1 dy_2 \\
 &= \frac{1}{2} \int_{-1}^0 e^{-y_1} \int_{-y_1}^{\infty} e^{-3y_2/2} dy_2 dy_1 + \frac{1}{2} \int_0^1 e^{-y_1} \int_0^{\infty} e^{-3y_2/2} dy_2 dy_1 \\
 &= \frac{1}{3} \int_{-1}^0 e^{-y_1} \left[ -e^{-3y_2/2} \right]_{-y_1}^{\infty} dy_1 + \frac{1}{3} \int_0^1 e^{-y_1} \left[ -e^{-3y_2/2} \right]_0^{\infty} dy_1 \\
 &= \frac{1}{3} \int_{-1}^0 e^{y_1/2} dy_1 + \frac{1}{3} \int_0^1 e^{-y_1} dy_1 \\
 &= \frac{2}{3} e^{y_1/2} \Big|_{-1}^0 - \frac{1}{3} e^{-y_1} \Big|_0^1 \\
 &= \frac{2}{3} [1 - e^{-1/2}] + \frac{1}{3} [1 - e^{-1}] \\
 &\approx 0.47 \quad (4.3)
 \end{aligned}$$

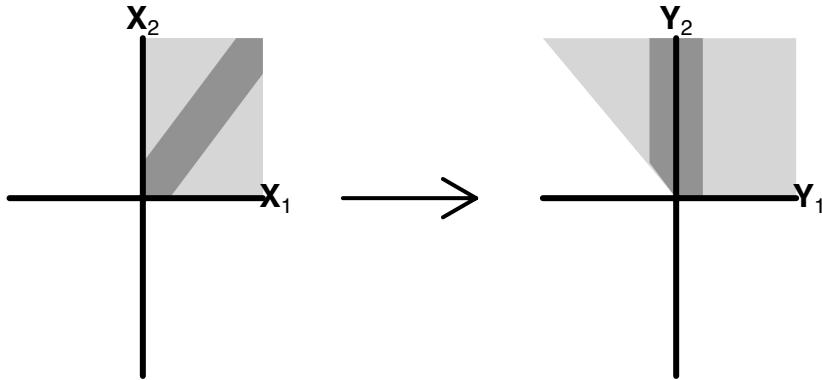


Figure 4.1: The  $(X_1, X_2)$  plane and the  $(Y_1, Y_2)$  plane. The light gray regions are where  $\vec{X}$  and  $\vec{Y}$  live. The dark gray regions are where  $|X_1 - X_2| \leq 1$ .

Figure 4.1 was produced by the following snippet.

```
par (mar=c(0,0,0,0))
plot (c(0,6), c(0,2), type="n", xlab="", ylab="", xaxt="n",
 yaxt="n", bty="n")
polygon (c(1,1.9,1.9,1), c(1,1,1.9,1.9), col=gray(.8),
 border=NA)
polygon (c(1,1.2,1.9,1.9,1.7,1), c(1,1,1.7,1.9,1.9,1.2),
 col=gray(.5), border=NA)
segments (0, 1, 1.9, 1, lwd=3) # x1 axis
segments (1, 0, 1, 1.9, lwd=3) # x2 axis
text (c(2,1), c(1,2), c(expression(bold(X[1])),
 expression(bold(X[2]))))
polygon (c(5,5.9,5.9,4), c(1,1,1.9,1.9), col=gray(.8),
 border=NA)
polygon (c(5,5.2,5.2,4.8,4.8), c(1,1,1.9,1.9,1.2),
 col=gray(.5), border=NA)
segments (4, 1, 5.9, 1, lwd=3) # y1 axis
segments (5, 0, 5, 1.9, lwd=3) # y2 axis
text (c(6,5), c(1,2), c(expression(bold(Y[1])),
 expression(bold(Y[2]))))

arrows (2.5, 1, 3.5, 1, length=.2, lwd=2)
```

The point of the example, of course, is the method, not the answer. Functions of random variables and random vectors are common in statistics and probability. There are many methods to deal with them. The method of transforming the pdf is one that is often useful.

## 4.3 Representing Distributions

We usually describe a random variable  $Y$  through  $p_Y$  — its pmf if  $Y$  is discrete or its pdf if  $Y$  is continuous. But there are at least two alternatives. First, any random variable  $Y$  can be described by its *cumulative distribution function*, or cdf,  $F_Y$  which is defined by

$$F_Y(c) \equiv P[Y \leq c] = \begin{cases} \sum_{y=-\infty}^c P[Y = y] & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^c p(y) dy & \text{if } Y \text{ is continuous.} \end{cases} \quad (4.4)$$

Equation 4.4 defines the cdf in terms of the pmf or pdf. It is also possible to go the other way. If  $Y$  is continuous, then for any number  $b \in \mathbb{R}$

$$P(Y \leq b) = F(b) = \int_{-\infty}^b p(y) dy$$

which shows by the Fundamental Theorem of Calculus that  $p(y) = F'(y)$ . On the other hand, if  $Y$  is discrete, Then  $P[Y = y] = P[Y \leq y] - P[Y < y] = F_Y(y) - F_Y(y^-)$ . (We use the notation  $F_Y(y^-)$  to mean the limit of  $F_Y(z)$  as  $z$  approaches  $y$  from below. It is also written  $\lim_{\epsilon \uparrow 0} F_Y(y - \epsilon)$ ). Thus the reverse of Equation 4.4 is

$$p_Y(y) = \begin{cases} F_Y(y) - F_Y(y^-) & \text{if } Y \text{ is discrete} \\ F'_Y(y) & \text{if } Y \text{ is continuous} \end{cases} \quad (4.5)$$

Equation 4.5 is correct except in one case which seldom arises in practice. It is possible that  $F_Y(y)$  is a continuous but nondifferentiable function, in which case  $Y$  is a continuous random variable, but  $Y$  does not have a density. In this case there is a cdf  $F_Y$  without a corresponding pmf or pdf.

Figure 4.2 shows the pmf and cdf of the  $\text{Bin}(10, .7)$  distribution and the pdf and cdf of the  $\text{Exp}(1)$  distribution.

Figure 4.2 was produced by the following snippet.

```
par (mfrow=c(2,2))
y <- seq (-1, 11, by=1)
plot (y, dbinom (y, 10, .7), type="p", ylab="pmf",
 main="Bin (10, .7)")
plot (y, pbinom (y, 10, .7), type="p", pch=16,
 ylab="cdf", main="Bin (10, .7)")
segments (-1:10, pbinom (-1:10, 10, .7),
 0:11, pbinom (-1:10, 10, .7))
y <- seq (0, 5, len=50)
plot (y, dexp (y, 1), type="l", ylab="pdf", main="Exp(1)")
plot (y, pexp (y, 1), type="l", ylab="cdf", main="Exp(1)")
```

- `segments ( x0, y0, x1, y1)` draws line segments. The line segments run from  $(x0, y0)$  to  $(x1, y1)$ . The arguments may be vectors.

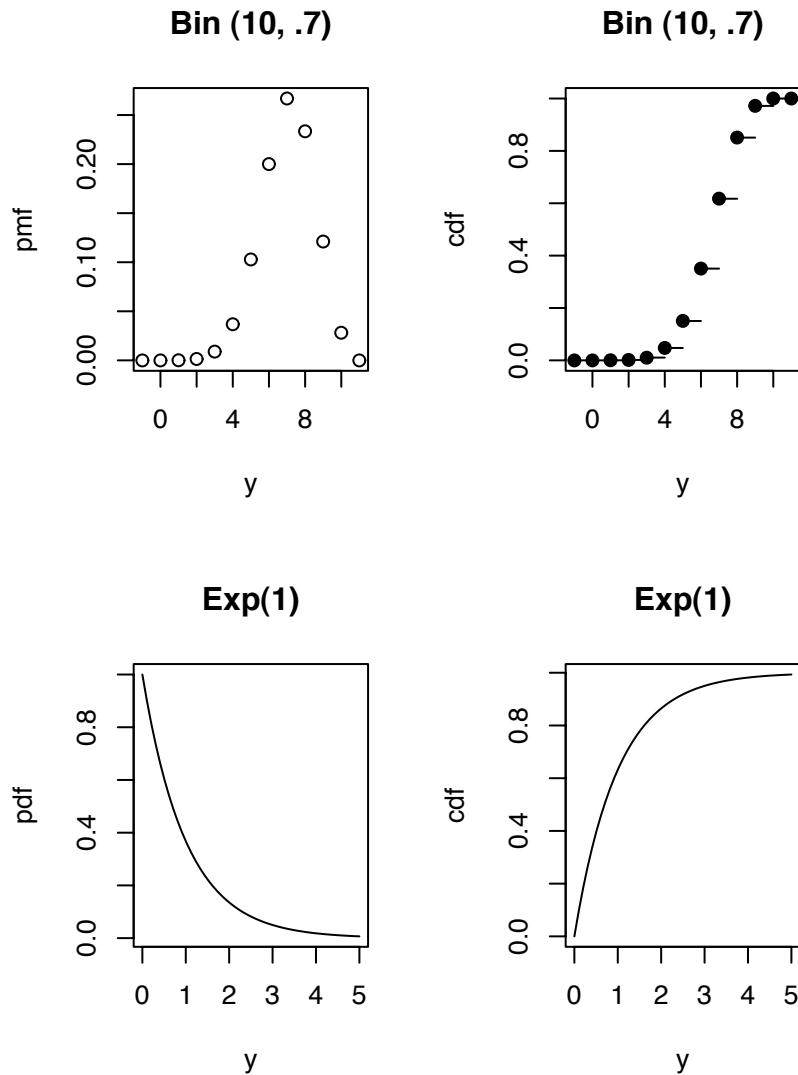


Figure 4.2: pmf's, pdf's, and cdf's

The other alternative representation for  $Y$  is its *moment generating function* or *mgf*  $M_Y$ . The moment generating function is defined as

$$M_Y(t) = \mathbb{E}[e^{tY}] = \begin{cases} \sum_y e^{ty} p_Y(y) & \text{if } Y \text{ is discrete} \\ \int e^{ty} p_Y(y) dy & \text{if } Y \text{ is continuous} \end{cases} \quad (4.6)$$

$M_Y$  is also known as the *Laplace transform* of  $p_Y$ .

Because we define the mgf as a sum or integral there is the question of whether the sum or integral is finite and hence whether the mgf is well defined. In Equation 4.6, the mgf is always defined at  $t = 0$ . (See Exercise 12.) But even if  $M_Y(t)$  is not well defined (the integral or sum is not absolutely convergent) for large  $t$ , what matters for statistical practice is whether  $M_Y(t)$  is well defined in a neighborhood of 0, i.e. whether there exists a  $\delta > 0$  such that  $M_Y(t)$  exists for  $t \in (-\delta, \delta)$ . The moment generating function gets its name from the following theorem.

**Theorem 4.5.** *If  $Y$  has mgf  $M_Y$  defined in a neighborhood of 0, then*

$$\mathbb{E}[Y^n] = M_Y^{(n)}(0) \equiv \frac{d^n}{dt^n} M_Y(t) \Big|_0$$

*Proof.* We provide the proof for the case  $n = 1$ . The proof for larger values of  $n$  is similar.

$$\begin{aligned} \frac{d}{dt} M_Y(t) \Big|_0 &= \frac{d}{dt} \int e^{ty} p_Y(y) dy \Big|_0 \\ &= \int \frac{d}{dt} e^{ty} \Big|_0 p_Y(y) dy \\ &= \int y e^{ty} \Big|_0 p_Y(y) dy \\ &= \int y p_Y(y) dy \\ &= \mathbb{E}[Y] \end{aligned}$$

□

The second line of the proof has the form

$$\frac{d}{dt} \int f(t, y) dy = \int \frac{d}{dt} f(t, y) dy,$$

an equality which is not necessarily true. It is true for “nice” functions  $f$ ; but establishing exactly what “nice” means requires measure theory and is beyond the scope of this book. We will continue to use the equality without thorough justification.

One could, if one wished, calculate and plot  $M_Y(t)$ , though there is usually little point in doing so. The main purpose of moment generating functions is in proving theorems and not, as their name might suggest, in deriving moments. And mgf's are useful in proving theorems mostly because of the following two results.

**Theorem 4.6.** *Let  $X$  and  $Y$  be two random variables with moment generating functions (assumed to exist)  $M_X$  and  $M_Y$ . If  $M_X(t) = M_Y(t)$  for all  $t$  in some neighborhood of 0, then  $F_X = F_Y$ ; i.e.,  $X$  and  $Y$  have the same distribution.*

**Theorem 4.7.** *Let  $Y_1, \dots$  be a sequence of random variables with moment generating functions (assumed to exist)  $M_{Y_1}, \dots$ . Define  $M(t) = \lim_{n \rightarrow \infty} M_{Y_n}(t)$ . If the limit exists for all  $t$  in a neighborhood of 0, and if  $M(t)$  is a moment generating function, then there is a unique cdf  $F$  such that*

1.

$$F(y) = \lim_{n \rightarrow \infty} F_{Y_n}(y)$$

for all  $y$  where  $F$  is continuous and

2.  $M$  is the mgf of  $F$ .

Theorems 4.6 and 4.7 both assume that the necessary mgf's exist. It is inconvenient that not all distributions have mgf's. One can avoid the problem by using *characteristic functions* (also known as Fourier transforms) instead of moment generating functions. The characteristic function is defined as

$$C_Y(t) = \mathbb{E}[e^{itY}]$$

where  $i = \sqrt{-1}$ . All distributions have characteristic functions, and the characteristic function completely characterizes the distribution, so characteristic functions are ideal for our purpose. However, dealing with complex numbers presents its own inconveniences. We shall not pursue this topic further. Proofs of Theorems 4.6 and 4.7 and similar results for characteristic functions are omitted but may be found in more advanced books.

Two more useful results are theorems 4.8 and 4.9.

**Theorem 4.8.** *Let  $X$  be a random variable,  $a, b$  be constants, and define  $Y = aX + b$ . Then  $M_Y(t) = e^{bt}M_X(at)$ .*

*Proof.*

$$\begin{aligned} M_Y(t) &= \mathbb{E}[e^{(aX+b)t}] \\ &= e^{bt}\mathbb{E}[e^{atX}] \\ &= e^{bt}M_X(at) \end{aligned}$$

□

**Theorem 4.9.** Let  $X$  and  $Y$  be independent random variables. Define  $Z = X + Y$ . Then

$$M_Z(t) = M_X(t)M_Y(t)$$

*Proof.*

$$M_Z(t) = \mathbb{E}[e^{(X+Y)t}] = \mathbb{E}[e^{Xt}e^{Yt}] = \mathbb{E}[e^{Xt}]\mathbb{E}[e^{Yt}] = M_X(t)M_Y(t)$$

□

**Corollary 4.10.** Let  $Y_1, \dots, Y_n$  be a collection of i.i.d. random variables each with mgf  $M_Y$ . Define  $X = Y_1 + \dots + Y_n$ . Then

$$M_X(t) = [M_Y(t)]^n$$

## 4.4 Exercises

1. Refer to Equation 4.1 on page 262.
  - (a) To help visualize the joint density  $p_{\vec{X}}$ , make a contour plot. You will have to choose some values of  $x_1$ , some values of  $x_2$ , and then evaluate  $p_{\vec{X}}(x_1, x_2)$  on all pairs  $(x_1, x_2)$  and save the values in a matrix. Finally, pass the values to the `contour` function. Choose values of  $x_1$  and  $x_2$  that help you visualize  $p_{\vec{X}}$ . You may have to choose values by trial and error.
  - (b) Draw a diagram that illustrates how to find the region  $A$  and the limits of integration in Equation 4.1.
  - (c) Supply the missing steps in Equation 4.1. Make sure you understand them. Verify the answer.
  - (d) Use R to verify the answer to Equation 4.1 by simulation.
2. Refer to Example 1.6 on page 43 on tree seedlings where  $N$  is the number of New seedlings that emerge in a given year and  $X$  is the number that survive to the next year. Find  $P[X \geq 1]$ .
3.  $(X_1, X_2)$  have a joint distribution that is uniform on the unit disk. Find  $p_{(X_1, X_2)}$ .
4. The random vector  $(X, Y)$  has pdf  $p_{(X,Y)}(x, y) \propto ky$  for some  $k > 0$  and  $(x, y)$  in the triangular region bounded by the points  $(0, 0)$ ,  $(-1, 1)$ , and  $(1, 1)$ .
  - (a) Find  $k$ .
  - (b) Find  $P[Y \leq 1/2]$ .

- (c) Find  $P[X \leq 0]$ .  
 (d) Find  $P[|X - Y| \leq 1/2]$ .
5. Prove the assertion on page 262 that mutual independence implies pairwise independence.
- (a) Begin with the case of three random variables  $\vec{X} = (X_1, X_2, X_3)$ . Prove that if  $X_1, X_2, X_3$  are mutually independent, then any two of them are independent.
  - (b) Generalize to the case  $\vec{X} = (X_1, \dots, X_n)$ .
6. Refer to Equation 4.2 on page 262. Verify that  $X_1 \perp X_2$ ,  $X_1 \perp X_3$ , and  $X_2 \perp X_3$  but that  $X_1, X_2$ , and  $X_3$  are not mutually independent.
7. Prove Lemma 4.1
8. Fill in the proof of Theorem 4.2 on page 263.
9.  $X$  and  $Y$  are uniformly distributed in the rectangle whose corners are  $(1, 0)$ ,  $(0, 1)$ ,  $(-1, 0)$ , and  $(0, -1)$ .
- (a)
    - i. Find  $p(x, y)$ .
    - ii. Are  $X$  and  $Y$  independent?
    - iii. Find the marginal densities  $p(x)$  and  $p(y)$ .
    - iv. Find the conditional densities  $p(x|y)$  and  $p(y|x)$ .
    - v. Find  $E[X]$ ,  $E[X|Y = .5]$ , and  $E[X|Y = -.5]$ .
  - (b) Let  $U = X + Y$  and  $V = X - Y$ .
    - i. Find the region where  $U$  and  $V$  live.
    - ii. Find the joint density  $p(u, v)$ .
    - iii. Are  $U$  and  $V$  independent?
    - iv. Find the marginal densities  $p(u)$  and  $p(v)$ .
    - v. Find the conditional densities  $p(u|v)$  and  $p(v|u)$ .
    - vi. Find  $E[U]$ ,  $E[U|V = .5]$ , and  $E[U|V = -.5]$ .
10. Let the random vector  $(U, V)$  be distributed uniformly on the unit square. Let  $X = UV$  and  $Y = U/V$ .
- (a) Draw the region of the  $X-Y$  plane where the random vector  $(X, Y)$  lives.
  - (b) Find the joint density of  $(X, Y)$ .

- (c) Find the marginal density of  $X$ .  
 (d) Find the marginal density of  $Y$ .  
 (e) Find  $P[Y > 1]$ .  
 (f) Find  $P[X > 1]$ .  
 (g) Find  $P[Y > 1/2]$ .  
 (h) Find  $P[X > 1/2]$ .  
 (i) Find  $P[XY > 1]$ .  
 (j) Find  $P[XY > 1/2]$ .
11. (a) Let  $(X_1, X_2)$  be distributed uniformly on the disk where  $X_1^2 + X_2^2 \leq 1$ . Let  $R = \sqrt{X_1^2 + X_2^2}$  and  $\Theta = \arctan(X_1/X_2)$ . Hint: it may help to draw a picture.
- i. What is the joint density  $p(x_1, x_2)$ ?
  - ii. Are  $X_1$  and  $X_2$  independent? Explain.
  - iii. Find the joint density  $p(r, \theta)$ .
  - iv. Are  $R$  and  $\Theta$  independent? Explain.
- (b) Let  $(X_1, X_2)$  be i.i.d.  $N(0,1)$ . Let  $R = \sqrt{X_1^2 + X_2^2}$  and  $\Theta = \arctan(X_1/X_2)$ .
- i. What is the joint density  $p(x_1, x_2)$ ?
  - ii. Find the joint density  $p(r, \theta)$ .
  - iii. Are  $R$  and  $\Theta$  independent? Explain.
  - iv. Find the marginal density  $p(r)$ .
  - v. Let  $V = R^2$ . Find the density  $p(v)$ .
- (c) Let  $(X_1, X_2)$  be distributed uniformly on the square whose corners are  $(1, 1)$ ,  $(-1, 1)$ ,  $(-1, -1)$ , and  $(1, -1)$ . Let  $R = \sqrt{X_1^2 + X_2^2}$  and  $\Theta = \arctan(X_1/X_2)$ .
- i. What is the joint density  $p(x_1, x_2)$ ?
  - ii. Are  $X_1$  and  $X_2$  independent? Explain.
  - iii. Are  $R$  and  $\Theta$  independent? Explain.
12. Just below Equation 4.6 is the statement “the mgf is always defined at  $t = 0$ .” For any random variable  $Y$ , find  $M_Y(0)$ .
13. Provide the proof of Theorem 4.5 for the case  $n = 2$ .
14. Refer to Theorem 4.9. Where in the proof is the assumption  $X \perp Y$  used?

## CHAPTER 5

# SPECIAL DISTRIBUTIONS

Statisticians often make use of standard *parametric families* of probability distributions. A parametric family is a collection of probability distributions distinguished by, or indexed by, a *parameter*. An example is the Binomial distribution introduced in Section 1.3.1. There were  $N$  trials. Each had a probability  $\theta$  of success. Usually  $\theta$  is unknown and could be any number in  $(0, 1)$ . There is one  $\text{Bin}(N, \theta)$  distribution for each value of  $\theta$ ;  $\theta$  is a parameter; the set of probability distributions

$$\{\text{Bin}(N, \theta) : \theta \in (0, 1)\}$$

is a parametric family of distributions.

We have already seen four parametric families — the Binomial (Section 1.3.1), Poisson (Section 1.3.2), Exponential (Section 1.3.3), and Normal (Section 1.3.4) distributions. Chapter 5 examines these in more detail and introduces several others.

## 5.1 The Binomial and Negative Binomial Distributions

**The Binomial Distribution** Statisticians often deal with situations in which there is a collection of *trials* performed under identical circumstances; each trial results in either *success* or *failure*. Typical examples are coin flips (Heads or Tails), medical trials (cure or not), voter polls (Democrat or Republican), basketball free throws (make or miss). Conditions for the Binomial Distribution are

1. the number of trials  $n$  is fixed in advance,
2. the probability of success  $\theta$  is the same for each trial, and

3. trials are conditionally independent of each other, given  $\theta$ .

Let the random variable  $X$  be the number of successes in such a collection of trials. Then  $X$  is said to have the Binomial distribution with parameters  $(n, \theta)$ , written  $X \sim \text{Bin}(n, \theta)$ . The possible values of  $X$  are the integers  $0, 1, \dots, n$ . Figure 1.5 shows examples of Binomial pmf's for several combinations of  $n$  and  $\theta$ . Usually  $\theta$  is unknown and the trials are performed in order to learn about  $\theta$ .

Obviously, large values of  $X$  are evidence that  $\theta$  is large and small values of  $X$  are evidence that  $\theta$  is small. But to evaluate the evidence quantitatively we must be able to say more. In particular, once a particular value  $X = x$  has been observed we want to quantify how well it is explained by different possible values of  $\theta$ . That is, we want to know  $p(x | \theta)$ .

**Theorem 5.1.** *If  $X \sim \text{Bin}(n, \theta)$  then*

$$p_X(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

for  $x = 0, 1, \dots, n$ .

*Proof.* When the  $n$  trials of a Binomial experiment are carried out there will be a sequence of successes (1's) and failures (0's) such as 1000110...100. Let  $\mathbb{S} = \{0, 1\}^n$  be the set of such sequences and, for each  $x \in \{0, 1, \dots, n\}$ , let  $\mathbb{S}_x$  be the subset of  $\mathbb{S}$  consisting of sequences with  $x$  1's and  $n - x$  0's. If  $s \in \mathbb{S}_x$  then  $\Pr(s) = \theta^x (1 - \theta)^{n-x}$ . In particular, all  $s$ 's in  $\mathbb{S}_x$  have the same probability. Therefore,

$$\begin{aligned} p_X(x) &= \Pr(X = x) = \Pr(\mathbb{S}_x) \\ &= (\text{size of } \mathbb{S}_x) \cdot (\theta^x (1 - \theta)^{n-x}) \\ &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \end{aligned}$$

□

The special case  $n = 1$  is important enough to have its own name. When  $n = 1$  then  $X$  is said to have a *Bernoulli* distribution with parameter  $\theta$ . We write  $X \sim \text{Bern}(\theta)$ . If  $X \sim \text{Bern}(\theta)$  then  $p_X(x) = \theta^x (1 - \theta)^{1-x}$  for  $x \in \{0, 1\}$ . Experiments that have two possible outcomes are called Bernoulli trials.

Suppose  $X_1 \sim \text{Bin}(n_1, \theta)$ ,  $X_2 \sim \text{Bin}(n_2, \theta)$  and  $X_1 \perp X_2$ . Let  $X_3 = X_1 + X_2$ . What is the distribution of  $X_3$ ? Logic suggests the answer is  $X_3 \sim \text{Bin}(n_1 + n_2, \theta)$  because (1) there are  $n_1 + n_2$  trials, (2) the trials all have the same probability of success  $\theta$ , (3) the trials are independent of each other (the reason for the  $X_1 \perp X_2$  assumption) and (4)  $X_3$  is the total number of successes. Theorem 5.3 shows a formal proof of this proposition. But first we need to know the moment generating function.

**Theorem 5.2.** Let  $X \sim \text{Bin}(n, \theta)$ . Then

$$M_X(t) = [\theta e^t + (1 - \theta)]^n$$

*Proof.* Let  $Y \sim \text{Bern}(\theta)$ . Then

$$M_Y(t) = \mathbb{E}[e^{tY}] = \theta e^t + (1 - \theta).$$

Now let  $X = \sum_{i=1}^n Y_i$  where the  $Y_i$ 's are i.i.d.  $\text{Bern}(\theta)$  and apply Corollary 4.10.  $\square$

**Theorem 5.3.** Suppose  $X_1 \sim \text{Bin}(n_1, \theta)$ ;  $X_2 \sim \text{Bin}(n_2, \theta)$ ; and  $X_1 \perp X_2$ . Let  $X_3 = X_1 + X_2$ . Then  $X_3 \sim \text{Bin}(n_1 + n_2, \theta)$ .

*Proof.*

$$\begin{aligned} M_{X_3}(t) &= M_{X_1}(t)M_{X_2}(t) \\ &= [\theta e^t + (1 - \theta)]^{n_1} [\theta e^t + (1 - \theta)]^{n_2} \\ &= [\theta e^t + (1 - \theta)]^{n_1+n_2} \end{aligned}$$

The first equality is by Theorem 4.9; the second is by Theorem 5.2. We recognize the last expression as the mgf of the  $\text{Bin}(n_1 + n_2, \theta)$  distribution. So the result follows by Theorem 4.6.  $\square$

The mean of the Binomial distribution was calculated in Equation 1.11. Theorem 5.4 restates that result and gives the variance and standard deviation.

**Theorem 5.4.** Let  $X \sim \text{Bin}(n, \theta)$ . Then

1.  $\mathbb{E}[X] = n\theta$ .
2.  $\text{Var}(X) = n\theta(1 - \theta)$ .
3.  $\text{SD}(X) = \sqrt{n\theta(1 - \theta)}$ .

*Proof.* The proof for  $E[X]$  was given earlier. If  $X \sim \text{Bin}(n, \theta)$ , then  $X = \sum_{i=1}^n X_i$  where  $X_i \sim \text{Bern}(\theta)$  and the  $X_i$ 's are mutually independent. Therefore, by Theorem 1.9,  $\text{Var}(X) = n \text{Var}(X_i)$ . But

$$\text{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \theta - \theta^2 = \theta(1 - \theta).$$

So  $\text{Var}(X) = n\theta(1 - \theta)$ . The result for  $\text{SD}(X)$  follows immediately.

Exercise 1 asks you to prove Theorem 5.4 by moment generating functions.  $\square$

R comes with built-in functions for working with Binomial distributions. You can get the following information by typing `help(dbinom)`, `help(pbinom)`, `help(qbinom)`, or `help(rbinom)`. There are similar functions for working with other distributions, but we won't repeat their help pages here.

**Usage:**

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

**Arguments:**

**x, q:** vector of quantiles.

**p:** vector of probabilities.

**n:** number of observations. If ‘`length(n) > 1`’, the length is taken to be the number required.

**size:** number of trials.

**prob:** probability of success on each trial.

**log, log.p:** logical; if TRUE, probabilities p are given as  $\log(p)$ .

**lower.tail:** logical; if TRUE (default), probabilities are  $P[X \leq x]$ , otherwise,  $P[X > x]$ .

**Details:**

The binomial distribution with ‘`size`’ = n and ‘`prob`’ = p has density

$$p(x) = \text{choose}(n, x) p^x (1-p)^{n-x}$$

for  $x = 0, \dots, n$ .

If an element of ‘x’ is not integer, the result of ‘dbinom’ is zero, with a warning.  $p(x)$  is computed using Loader’s algorithm, see the reference below.

The quantile is defined as the smallest value  $x$  such that  $F(x) \geq p$ , where  $F$  is the distribution function.

**Value:**

‘dbinom’ gives the density, ‘pbinom’ gives the distribution function, ‘qbinom’ gives the quantile function and ‘rbinom’ generates random deviates.

If ‘size’ is not an integer, ‘NaN’ is returned.

**References:**

Catherine Loader (2000). Fast and Accurate Computation of Binomial Probabilities; manuscript available from <URL: <http://cm.bell-labs.com/cm/ms/departments/sia/catherine/dbinom>>

**See Also:**

‘dnbinom’ for the negative binomial, and ‘dpois’ for the Poisson distribution.

**Examples:**

```
Compute P(45 < X < 55) for X Binomial(100, 0.5)
sum(dbinom(46:54, 100, 0.5))

Using "log = TRUE" for an extended range :
n <- 2000
k <- seq(0, n, by = 20)
plot (k, dbinom(k, n, pi/10, log=TRUE), type='l', ylab="log density",
 main = "dbinom(*, log=TRUE) is better than log(dbinom(*))")
lines(k, log(dbinom(k, n, pi/10)), col='red', lwd=2)
extreme points are omitted since dbinom gives 0.
```

```
mtext("dbinom(k, log=TRUE)", adj=0)
mtext("extended range", adj=0, line = -1, font=4)
mtext("log(dbinom(k))", col="red", adj=1)
```

Figure 5.1 shows the Binomial pmf for several values of  $x$ ,  $n$ , and  $p$ . Note that for a fixed  $p$ , as  $n$  gets larger the pmf looks increasingly like a Normal pdf. That's the Central Limit Theorem. Let  $Y_1, \dots, Y_n \sim \text{i.i.d. Bern}(p)$ . Then the distribution of  $X$  is the same as the distribution of  $\sum Y_i$  and the Central Limit Theorem tells us that  $\sum Y_i$  looks increasingly Normal as  $n \rightarrow \infty$ .

Also, for a fixed  $n$ , the pmf looks more Normal when  $p = .5$  than when  $p = .05$ . And that's because convergence under the Central Limit Theorem is faster when the distribution of each  $Y_i$  is more symmetric.

Figure 5.1 was produced by

```
par (mfrow=c(3,2))
n <- 5
p <- .05
x <- 0:5
plot (x, dbinom(x,n,p), ylab="p(x)", main="n=5, p=.05")
...
```

**The Negative Binomial Distribution** Rather than fix in advance the number of trials, experimenters will sometimes continue the sequence of trials until a prespecified number of successes  $r$  has been achieved. In this case the total number of failures  $N$  is the random variable and is said to have the Negative Binomial distribution with parameters  $(r, \theta)$ , written  $N \sim \text{NegBin}(r, \theta)$ . (Warning: some authors say that the total number of trials,  $N + r$ , has the Negative Binomial distribution.) One example is a gambler who decides to play the daily lottery until she wins. The prespecified number of successes is  $r = 1$ . The number of failures  $N$  until she wins is random. In this case, and whenever  $r = 1$ ,  $N$  is said to have a Geometric distribution with parameter  $\theta$ ; we write  $N \sim \text{Geo}(\theta)$ . Often,  $\theta$  is unknown. Large values of  $N$  are evidence that  $\theta$  is small; small values of  $N$  are evidence

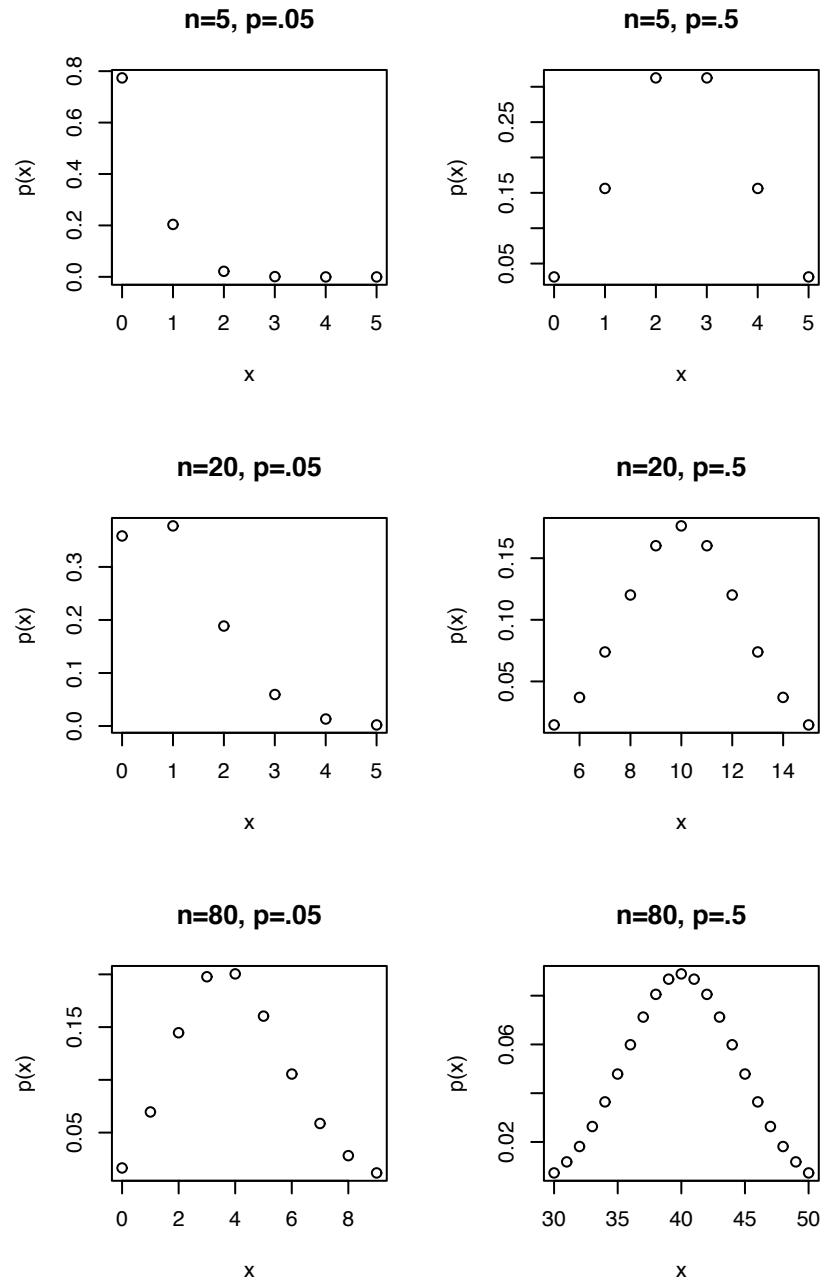


Figure 5.1: The Binomial pmf

that  $\theta$  is large. The probability function is

$$\begin{aligned} p_N(k) &= P(N = k) \\ &= P(r - 1 \text{ successes in the first } k + r - 1 \text{ trials} \\ &\quad \text{and } k + r\text{'th trial is a success}) \\ &= \binom{k+r-1}{r-1} \theta^r (1-\theta)^k \end{aligned}$$

for  $k = 0, 1, \dots$ .

Let  $N_1 \sim \text{NegBin}(r_1, \theta), \dots, N_t \sim \text{NegBin}(r_t, \theta)$ , and  $N_1, \dots, N_t$  be independent of each other. Then one can imagine a sequence of trials of length  $\sum(N_i + r_i)$  having  $\sum r_i$  successes.  $N_1$  is the number of failures before the  $r_1$ 'th success;  $\dots$ ;  $N_1 + \dots + N_t$  is the number of failures before the  $r_1 + \dots + r_t$ 'th success. It is evident that  $N \equiv \sum N_i$  is the number of failures before the  $r \equiv \sum r_i$ 'th success occurs and therefore that  $N \sim \text{NegBin}(r, \theta)$ .

**Theorem 5.5.** *If  $Y \sim \text{NegBin}(r, \theta)$  then  $E[Y] = r(1 - \theta)/\theta$  and  $\text{Var}(Y) = r(1 - \theta)/\theta^2$ .*

*Proof.* It suffices to prove the result for  $r = 1$ . Then the result for  $r > 1$  will follow by the foregoing argument and Theorems 1.7 and 1.9. For  $r = 1$ ,

$$\begin{aligned} E[N] &= \sum_{n=0}^{\infty} n P[N = n] \\ &= \sum_{n=1}^{\infty} n(1 - \theta)^n \theta \\ &= \theta(1 - \theta) \sum_{n=1}^{\infty} n(1 - \theta)^{n-1} \\ &= -\theta(1 - \theta) \sum_{n=1}^{\infty} \frac{d}{d\theta} (1 - \theta)^n \\ &= -\theta(1 - \theta) \frac{d}{d\theta} \sum_{n=1}^{\infty} (1 - \theta)^n \\ &= -\theta(1 - \theta) \frac{d}{d\theta} \frac{1 - \theta}{\theta} \\ &= -\theta(1 - \theta) \frac{-1}{\theta^2} \\ &= \frac{1 - \theta}{\theta} \end{aligned}$$

The trick of writing each term as a derivative, then switching the order of summation and derivative is occasionally useful. Here it is again.

$$\begin{aligned}
 \mathbb{E}(N^2) &= \sum_{n=0}^{\infty} n^2 P[N = n] \\
 &= \theta(1 - \theta) \sum_{n=1}^{\infty} (n(n-1) + n)(1 - \theta)^{n-1} \\
 &= \theta(1 - \theta) \sum_{n=1}^{\infty} n(1 - \theta)^{n-1} + \theta(1 - \theta)^2 \sum_{n=1}^{\infty} n(n-1)(1 - \theta)^{n-2} \\
 &= \frac{1 - \theta}{\theta} + \theta(1 - \theta)^2 \sum_{n=1}^{\infty} \frac{d^2}{\theta^2} (1 - \theta)^n \\
 &= \frac{1 - \theta}{\theta} + \theta(1 - \theta)^2 \frac{d^2}{\theta^2} \sum_{n=1}^{\infty} (1 - \theta)^n \\
 &= \frac{1 - \theta}{\theta} + \theta(1 - \theta)^2 \frac{d^2}{\theta^2} \frac{1 - \theta}{\theta} \\
 &= \frac{1 - \theta}{\theta} + 2 \frac{\theta(1 - \theta)^2}{\theta^3} \\
 &= \frac{2 - 3\theta + \theta^2}{\theta^2}
 \end{aligned}$$

Therefore,

$$\text{Var}(N) = E[N^2] - (E[N])^2 = \frac{1 - \theta}{\theta^2}.$$

□

The R functions for working with the negative Binomial distribution are `dnb`, `pnb`, `qnb`, and `rnb`. Figure 5.2 displays the Negative Binomial pdf and illustrates the use of `qnb`.

Figure 5.2 was produced with the following snippet.

```

r <- c(1, 5, 30)
p <- c(.1, .5, .8)
par(mfrow=c(3,3))
for (i in seq(along=r))
 for (j in seq(along=p)) {
 lo <- qnb(.01,r[i],p[j])
 }

```

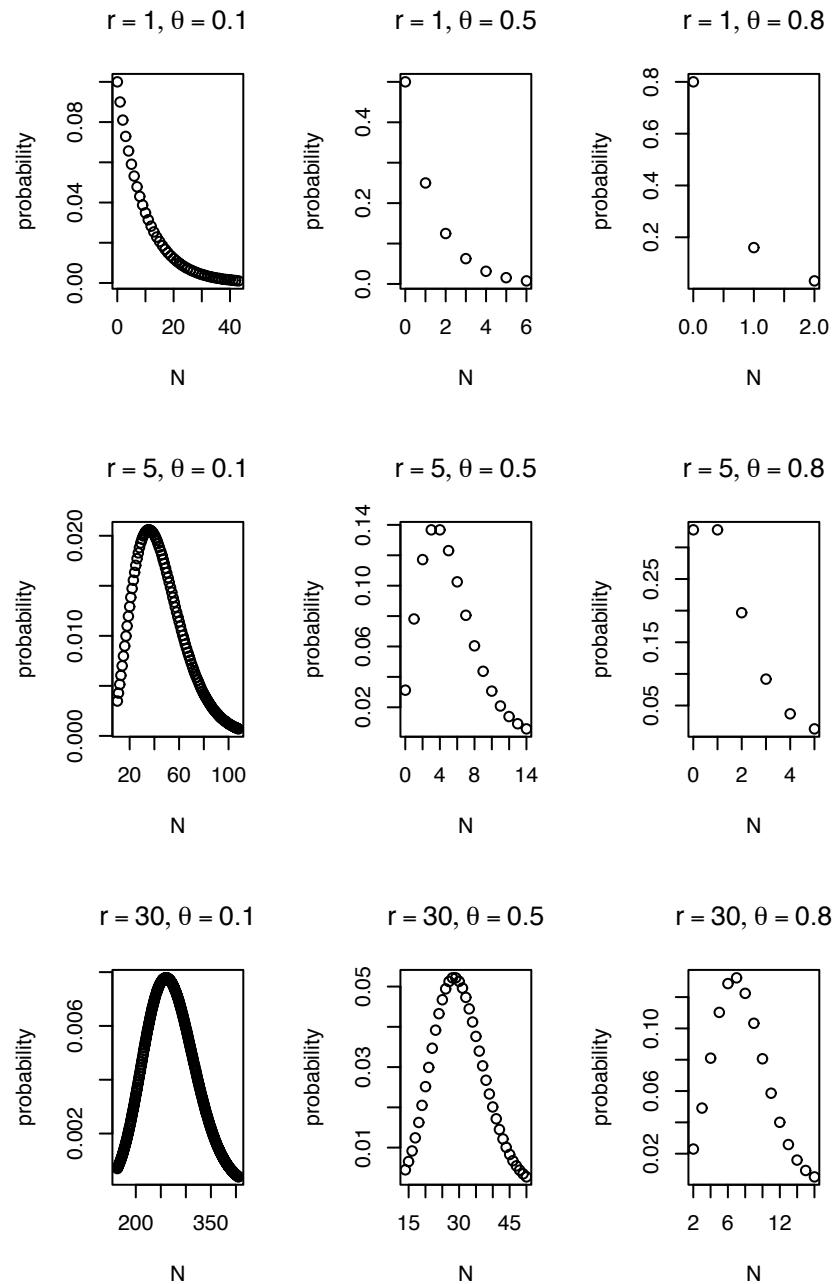


Figure 5.2: The Negative Binomial pmf

```

hi <- qnbinom(.99,r[i],p[j])
x <- lo:hi
plot (x, dnbinom(x,r[i],p[j]), ylab="probability", xlab="N",
 main = substitute (list (r == a, theta == b),
 list(a=i,b=j)))
}

```

- `lo` and `hi` are the limits on the x-axis of each plot. The use of `qbinom` ensures that each plot shows at least 98% of its distribution.

## 5.2 The Multinomial Distribution

The multinomial distribution generalizes the binomial distribution in the following way. The binomial distribution applies when the outcome of a trial has two possible values; the multinomial distribution applies when the outcome of a trial has more than two possible outcomes. Some examples are

**Clinical Trials** In clinical trials, each patient is administered a treatment, usually an experimental treatment or a standard, control treatment. Later, each patient may be scored as either success, failure, or censored. Censoring occurs because patients don't show up for their appointments, move away, or can't be found for some other reason.

**Craps** After the come-out roll, each successive roll is either a win, loss, or neither.

**Genetics** Each gene comes in several variants. Every person has two copies of the gene, one maternal and one paternal. So the person's status can be described by a pair like  $\{a, c\}$  meaning that she has one copy of type  $a$  and one copy of type  $c$ . The pair is called the person's *genotype*. Each person in a sample can be considered a trial. Geneticists may count how many people have each genotype.

**Political Science** In an election, each person prefers either the Republican candidate, the Democrat, the Green, or is undecided.

In this case we count the number of outcomes of each type. If there are  $k$  possible outcomes then the result is a vector  $y_1, \dots, y_k$  where  $y_i$  is the number of times that outcome  $i$  occurred and  $y_1 + \dots + y_k = n$  is the number of trials.

Let  $p \equiv (p_1, \dots, p_k)$  be the probabilities of the  $k$  categories and  $n$  be the number of trials. We write  $Y \sim \text{Mult}(n, p)$ . In particular,  $Y \equiv (Y_1, \dots, Y_k)$  is a vector of length  $k$ . Because  $Y$  is a vector, so is its expectation

$$\mathbb{E}[Y] = \mu = (\mu_1, \dots, \mu_k) \equiv (\mathbb{E}[Y_1], \dots, \mathbb{E}[Y_k]) = (np_1, \dots, np_k).$$

The  $i$ 'th coordinate,  $Y_i$ , is a random variable in its own right. Because  $Y_i$  counts the number of times outcome  $i$  occurred in  $n$  trials, its distribution is

$$Y_i \sim \text{Bin}(n, p_i). \quad (\text{See Exercise 20.}) \quad (5.1)$$

Although the  $Y_i$ 's are all Binomial, they are not independent. After all, if  $Y_1 = n$ , then  $Y_2 = \dots = Y_k = 0$ , so the  $Y_i$ 's must be dependent. What is their joint pmf? What is the conditional distribution of, say,  $Y_2, \dots, Y_k$  given  $Y_1$ ? The next two theorems provide the answers.

**Theorem 5.6.** *If  $Y \sim \text{Mult}(n, p)$  then*

$$f_Y(y_1, \dots, y_k) = \binom{n}{y_1 \dots y_k} p_1^{y_1} \cdots p_k^{y_k}$$

where  $\binom{n}{y_1 \dots y_k}$  is the multinomial coefficient

$$\binom{n}{y_1 \dots y_k} = \frac{n!}{\prod y_i!}$$

*Proof.* When the  $n$  trials of a multinomial experiment are carried out, there will be a sequence of outcomes such as  $abkdbg \cdots f$ , where the letters indicate the outcomes of individual trials. One such sequence is

$$\underbrace{a \cdots a}_{y_1 \text{ times}} \underbrace{b \cdots b}_{y_2 \text{ times}} \cdots \underbrace{k \cdots k}_{y_k \text{ times}}$$

The probability of this particular sequence is  $\prod p_i^{y_i}$ . Every sequence with  $y_1$   $a$ 's,  $\dots$ ,  $y_k$   $k$ 's has the same probability. So

$$f_Y(y_1, \dots, y_k) = (\text{number of such sequences}) \times \prod p_i^{y_i} = \binom{n}{y_1 \dots y_k} \prod p_i^{y_i}.$$

□

**Theorem 5.7.** If  $Y \sim \text{Mult}(n, p)$  then

$$(Y_2, \dots, Y_k \mid Y_1 = y_1) \sim \text{Mult}(n - y_1, (p_2^*, \dots, p_k^*))$$

where  $p_i^* = p_i / (1 - p_1)$  for  $i = 2, \dots, k$ .

*Proof.* See Exercise 19. □

R's functions for the multinomial distribution are `rmultinom` and `dmultinom`. `rmultinom(m, n, p)` draws a sample of size  $m$ .  $p$  is a vector of probabilities. The result is a  $k \times m$  matrix. Each column is one draw, so each column sums to  $n$ . The user does not specify  $k$ ; it is determined by `k = length(p)`.

## 5.3 The Poisson Distribution

The Poisson distribution is used to model counts in the following situation.

- There is a domain of study, usually a block of space or time.
- Events arise at seemingly random locations in the domain.
- There is an underlying rate at which events arise.
- The rate does not vary over the domain.
- The occurrence of an event at any location  $\ell_1$  is independent of the occurrence of an event at any other location  $\ell_2$ .

Let  $y$  be the total number of events that arise in the domain.  $Y$  has a Poisson distribution with rate parameter  $\lambda$ , written  $Y \sim \text{Poi}(\lambda)$ . The pmf is

$$p_Y(y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad \text{for } y = 0, 1, \dots$$

The mean was derived in Chapter 1, Exercise 18A. It is

$$\mathbb{E}[Y] = \lambda.$$

**Theorem 5.8.** Let  $Y \sim \text{Poi}(\lambda)$ . Then

$$M_Y(t) = e^{\lambda(e^t - 1)}$$

*Proof.*

$$\begin{aligned}
 M_Y(t) &= \mathbb{E}[e^{tY}] = \sum_{y=0}^{\infty} e^{ty} p_Y(y) = \sum_{y=0}^{\infty} e^{ty} \frac{e^{-\lambda} \lambda^y}{y!} \\
 &= \sum_{y=0}^{\infty} \frac{e^{-\lambda} (\lambda e^t)^y}{y!} = \frac{e^{-\lambda}}{e^{-\lambda e^t}} \sum_{y=0}^{\infty} \frac{e^{-\lambda e^t} (\lambda e^t)^y}{y!} \\
 &= e^{\lambda(e^t - 1)}
 \end{aligned}$$

□

**Theorem 5.9.** Let  $Y \sim Poi(\lambda)$ . Then

$$\text{Var}(Y) = \lambda$$

*Proof.* Just for fun (!) we will prove the theorem two ways — first directly and then with moment generating functions.

**Proof 1.**

$$\begin{aligned}
 \mathbb{E}[Y^2] &= \sum_{y=0}^{\infty} y^2 \frac{e^{-\lambda} \lambda^y}{y!} \\
 &= \sum_{y=0}^{\infty} y(y-1) \frac{e^{-\lambda} \lambda^y}{y!} + \sum_{y=0}^{\infty} y \frac{e^{-\lambda} \lambda^y}{y!} \\
 &= \sum_{y=2}^{\infty} y(y-1) \frac{e^{-\lambda} \lambda^y}{y!} + \lambda \\
 &= \sum_{z=0}^{\infty} \frac{e^{-\lambda} \lambda^{z+2}}{z!} + \lambda \\
 &= \lambda^2 + \lambda
 \end{aligned}$$

So  $\text{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \lambda$ .

**Proof 2.**

$$\begin{aligned}
\mathbb{E}[Y^2] &= \frac{d^2}{dt^2} M_Y(t) \Big|_{t=0} \\
&= \frac{d^2}{dt^2} e^{\lambda(e^t-1)} \Big|_{t=0} \\
&= \frac{d}{dt} \lambda e^t e^{\lambda(e^t-1)} \Big|_{t=0} \\
&= [\lambda e^t e^{\lambda(e^t-1)} + \lambda^2 e^{2t} e^{\lambda(e^t-1)}] \Big|_{t=0} \\
&= \lambda + \lambda^2
\end{aligned}$$

So  $\text{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \lambda$ . □

**Theorem 5.10.** Let  $Y_i \sim \text{Poi}(\lambda_i)$  for  $i = 1, \dots, n$  and let the  $Y_i$ 's be mutually independent. Let  $Y = \sum Y_i$  and  $\lambda = \sum \lambda_i$ . Then  $Y \sim \text{Poi}(\lambda)$ .

*Proof.* Using Theorems 4.9 and 5.8 we have

$$M_Y(t) = \prod M_{Y_i}(t) = \prod e^{\lambda_i(e^t-1)} = e^{\lambda(e^t-1)}$$

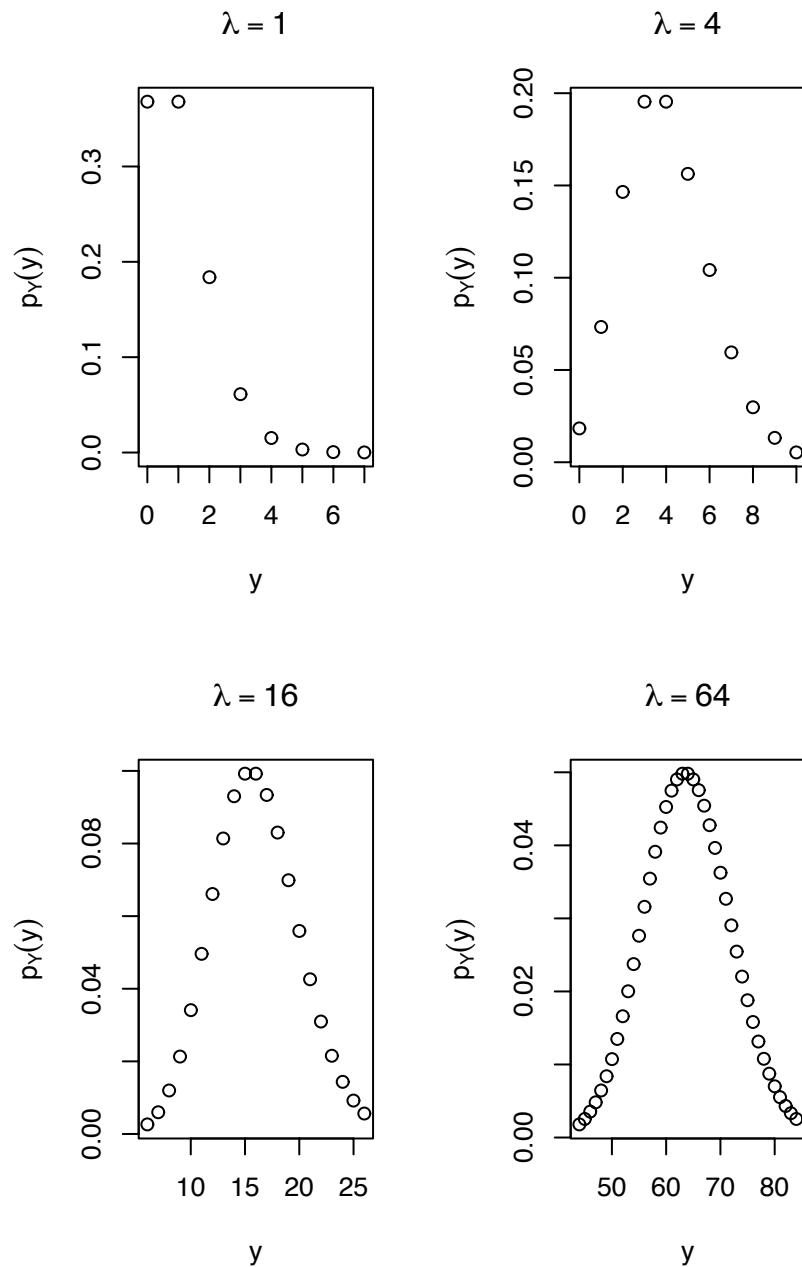
which is the mgf of the  $\text{Poi}(\lambda)$  distribution. □

Suppose, for  $i = 1, \dots, n$ ,  $Y_i$  is the number of events occurring on a domain  $D_i$ ;  $Y_i \sim \text{Poi}(\lambda_i)$ . Suppose the  $D_i$ 's are disjoint and the  $Y_i$ 's are independent. Let  $Y = \sum Y_i$  be the number of events arising on  $D = \cup D_i$ . The logic of the situation suggests that  $Y \sim \text{Poi}(\lambda)$  where  $\lambda = \sum \lambda_i$ . Theorem 5.10 assures us that everything works correctly; that  $Y$  does indeed have the  $\text{Poi}(\lambda)$  distribution. Another way to put it: If  $Y \sim \text{Poi}(\lambda)$ , and if the individual events that  $Y$  counts are randomly divided into two types  $Y_1$  and  $Y_2$  according to a binomial distribution with parameter  $\theta$ , then (1)  $Y_1 \sim \text{Poi}(\lambda\theta)$  and  $Y_2 \sim \text{Poi}(\lambda(1-\theta))$  and (2)  $Y_1 \perp Y_2$ .

Figure 5.3 shows the Poisson pmf for  $\lambda = 1, 4, 16, 64$ . As  $\lambda$  increases the pmf looks increasingly Normal. That's a consequence of Theorem 5.10 and the Central Limit Theorem. When  $Y \sim \text{Poi}(\lambda)$  Theorem 5.10 tells us we can think of  $Y$  as  $Y = \sum_{i=1}^{\lambda} Y_i$  where each  $Y_i \sim \text{Poi}(1)$ . ( $\lambda$  must be an integer for this to be precise.) Then the Central Limit Theorem tells us that  $Y$  will be approximately Normal when  $\lambda$  is large.

Figure 5.3 was produced with the following snippet.

```
y <- 0:7
plot (y, dpois(y,1), xlab="y", ylab=expression(p[Y](y)),
```

Figure 5.3: Poisson pmf for  $\lambda = 1, 4, 16, 64$

```

main=expression(lambda==1))
y <- 0:10
plot (y, dpois(y,4), xlab="y", ylab=expression(p[Y](y)),
 main=expression(lambda==4))
y <- 6:26
plot (y, dpois(y,16), xlab="y", ylab=expression(p[Y](y)),
 main=expression(lambda==16))
y <- 44:84
plot (y, dpois(y,64), xlab="y", ylab=expression(p[Y](y)),
 main=expression(lambda==64))

```

One of the early uses of the Poisson distribution was in *The probability variations in the distribution of  $\alpha$  particles* by RUTHERFORD AND GEIGER 1910. (An  $\alpha$  particle is a Helium nucleus, or two protons and two neutrons.)

### Example 5.1 (Rutherford and Geiger)

The phenomenon of radioactivity was beginning to be understood in the early 20<sup>th</sup> century. In their 1910 article, Rutherford and Geiger write

“In counting the  $\alpha$  particles emitted from radioactive substances . . . [it] is of importance to settle whether . . . variations in distribution are in agreement with the laws of probability, *i.e.* whether the distribution of  $\alpha$  particles on an average is that to be anticipated if the  $\alpha$  particles are expelled at random both in regard to space and time. It might be conceived, for example, that the emission of an  $\alpha$  particle might precipitate the disintegration of neighbouring atoms, and so lead to a distribution of  $\alpha$  particles at variance with the simple probability law.”

So Rutherford and Geiger are going to do three things in their article. They’re going to count  $\alpha$  particle emissions from some radioactive substance; they’re going to derive the distribution of  $\alpha$  particle emissions according to theory; and they’re going to compare the actual and theoretical distributions.

Here they describe their experimental setup.

“The source of radiation was a small disk coated with polonium, which was placed inside an exhausted tube, closed at one end by a zinc sulphide screen. The scintillations were counted in the usual way . . . the number of scintillations . . . corresponding to 1/8 minute intervals were counted . . .”

"The following example is an illustration of the result obtained. The numbers, given in the horizontal lines, correspond to the number of scintillations for successive intervals of 7.5 seconds.

|              |                           |   |   |   |   |   |    |        | Total per minute. |
|--------------|---------------------------|---|---|---|---|---|----|--------|-------------------|
| 1st minute : | 3                         | 7 | 4 | 4 | 2 | 3 | 2  | 0 .... | 25                |
| 2nd "        | 5                         | 2 | 5 | 4 | 3 | 5 | 4  | 2 .... | 30                |
| 3rd "        | 5                         | 4 | 1 | 3 | 3 | 1 | 5  | 2 .... | 24                |
| 4th "        | 8                         | 2 | 2 | 2 | 3 | 4 | 2  | 6 .... | 31                |
| 5th "        | 7                         | 4 | 2 | 6 | 4 | 5 | 10 | 4 .... | 42                |
|              |                           |   |   |   |   |   |    | —      |                   |
|              | Average for 5 minutes ... |   |   |   |   |   |    |        | 30.4              |
|              | True average .....        |   |   |   |   |   |    |        | 31.0              |

And here they describe their theoretical result.

"The distribution of  $\alpha$  particles according to the law of probability was kindly worked out for us by Mr. Bateman. The mathematical theory is appended as a note to this paper. Mr. Bateman has shown that if  $x$  be the true average number of particles for any given interval falling on the screen from a constant source, the probability that  $n \alpha$  particles are observed in the same interval is given by  $\frac{x^n}{n!} e^{-x}$ .  $n$  is here a whole number, which may have all positive values from 0 to  $\infty$ . The value of  $x$  is determined by counting a large number of scintillations and dividing by the number of intervals involved. The probability for  $n \alpha$  particles in the given interval can then at once be calculated from the theory."

Refer to BATEMAN [1910] for his derivation. Table 5.1 shows their data. As Rutherford and Geiger explain:

"For convenience the tape was measured up in four parts, the results of which are given separately in horizontal columns I. to IV.

"For example (see column I.), out of 792 intervals of 1/8 minute, in which 3179  $\alpha$  particles were counted, the number of intervals 3  $\alpha$  particles was 152. Combining the four columns, it is seen that out of 2608 intervals containing 10,097 particles, the number of times that 3  $\alpha$  particles were observed was 525. The number calculated from the equation was the same, viz. 525."

| $\alpha$              | 0  | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8  | 9  | 10 | 11 | 12 | 13 | 14 | Number<br>of<br>particles | Number<br>$\alpha$<br>of<br>particles | Number<br>of<br>inter-<br>vals | Average<br>number<br>of<br>inter-<br>vals |
|-----------------------|----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|----|----|---------------------------|---------------------------------------|--------------------------------|-------------------------------------------|
| I .....               | 15 | 56  | 106 | 152 | 170 | 122 | 88  | 50  | 17 | 12 | 3  | 0  | 0  | 1  | 0  | 3179                      | 792                                   | 4.01                           |                                           |
| II .....              | 17 | 39  | 88  | 116 | 120 | 98  | 63  | 37  | 4  | 9  | 4  | 1  | 0  | 0  | 0  | 2334                      | 596                                   | 3.92                           |                                           |
| III .....             | 15 | 56  | 97  | 139 | 118 | 96  | 60  | 26  | 18 | 3  | 3  | 1  | 0  | 0  | 0  | 2373                      | 632                                   | 3.75                           |                                           |
| IV .....              | 10 | 52  | 92  | 118 | 124 | 92  | 62  | 26  | 6  | 3  | 0  | 2  | 0  | 0  | 1  | 2211                      | 588                                   | 3.76                           |                                           |
| Sum ....              | 57 | 203 | 383 | 525 | 532 | 408 | 273 | 139 | 45 | 27 | 10 | 4  | 0  | 1  | 1  | 10097                     | 2608                                  | 3.87                           |                                           |
| Theoretical<br>values | 54 | 210 | 407 | 525 | 508 | 394 | 254 | 140 | 68 | 29 | 11 | 4  | 1  | 4  | 1  |                           |                                       |                                |                                           |

Table 5.1: Rutherford and Geiger's data

Finally, how did Rutherford and Geiger compare their actual and theoretical distributions? They did it with a plot, which we reproduce as Figure 5.4. Their conclusion:

"It will be seen that, on the whole, theory and experiment are in excellent accord. . . We may consequently conclude that the distribution of  $\alpha$  particles in time is in agreement with the laws of probability and that the  $\alpha$  particles are emitted at random. . . Apart from their bearing on radioactive problems, these results are of interest as an example of a method of testing the laws of probability by observing the variations in quantities involved in a spontaneous material process."

### Example 5.2 (neurobiology)

This example continues Example 2.6. We would like to know whether this neuron responds differently to different tastants and, if so, how. To that end, we'll see how often the neuron fires in a short period of time after receiving a tastant and we'll compare the results for different tastants. Specifically, we'll count the number of spikes in the 150 milliseconds ( $150 \text{ msec} = .15 \text{ s}$ ) immediately following the delivery of each tastant. ( $.15 \text{ msec}$  is about the rate at which rats can lick and is thought by neurobiologists to be about the right interval of time.) Let  $Y_{ij}$  be the number of spikes in the 150 msec following the  $j$ 'th delivery of tastant  $i$ . Because we're counting the number of events in a fixed period of time we'll adopt a Poisson model:

$$Y_{ij} \sim \text{Poi}(\lambda_i)$$

where  $\lambda_i$  is the average firing rate of this neuron to tastant  $i$ .

We begin by making a list to hold the data. There should be one element for each tastant. That element should be a vector whose length is the number of times that tastant was delivered. Here is the R code to do it. (Refer to Example 2.6 for reading in the data.)

```
nspikes <- list(
 MSG100 = rep (NA, length(tastants$MSG100)),
 MSG300 = rep (NA, length(tastants$MSG300)),
 NaCl100 = rep (NA, length(tastants$NaCl100)),
 NaCl300 = rep (NA, length(tastants$NaCl300)),
 water = rep (NA, length(tastants$water))
)
```

Now we fill in each element by counting the number of neuron firings in the time interval.

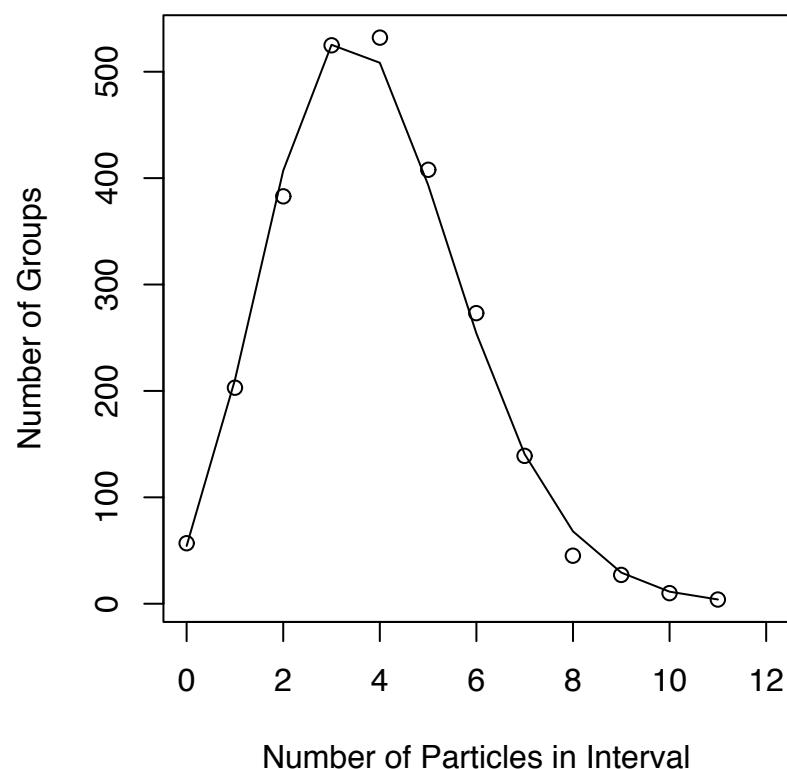


Figure 5.4: Rutherford and Geiger's Figure 1 comparing theoretical (solid line) to actual (open circles) distribution of  $\alpha$  particle counts.

```

for (i in seq(along=nspikes))
for (j in seq(along=nspikes[[i]]))
 nspikes[[i]][j] <- sum (spikes[[8]] > tastants[[i]][j]
 & spikes[[8]] <= tastants[[i]][j] + .15
)

```

Now we can see how many times the neuron fired after each delivery of, say, MSG100 by typing `nspikes$MSG100`.

Figure 5.5 compares the five tastants graphically. Panel **A** is a stripchart. It has five tick marks on the  $x$ -axis for the five tastants. Above each tick mark is a collection of circles. Each circle represents one delivery of the tastant and shows how many times the neuron fired in the 150 msec following that delivery. Panel **B** shows much the same information in a mosaic plot. The heights of the boxes show how often that tastant produced 0, 1, ..., 5 spikes. The width of each column shows how often that tastant was delivered. Panel **C** shows much the same information in yet a different way. It has one line for each tastant; that line shows how often the neuron responded with 0, 1, ..., 5 spikes. Panel **D** compares likelihood functions. The five curves are the likelihood functions for  $\lambda_1, \dots, \lambda_5$ .

There does not seem to be much difference in the response of this neuron to different tastants. Although we can compute the m.l.e.  $\hat{\lambda}_i$ 's with

```
lapply (nspikes, mean)
```

and find that they range from a low of  $\hat{\lambda}_3 \approx 0.08$  for .1 M NaCl to a high of  $\hat{\lambda}_1 \approx 0.4$  for .1M MSG, panel **D** suggests the plausibility of  $\lambda_1 = \dots = \lambda_5 \approx .2$ .

Figure 5.5 was produced with the following snippet.

```

spiketable <- matrix (NA, length(nspikes), 6,
 dimnames = list (tastant = 1:5,
 counts = 0:5)
)
for (i in seq(along=nspikes))
 spiketable[i,] <- hist (nspikes[[i]], seq(-.5,5.5,by=1),
 plot=F)$counts
freqtable <- apply (spiketable, 1, function(x)x/sum(x))

```

## A Neuron's Responses to 5 Tastants

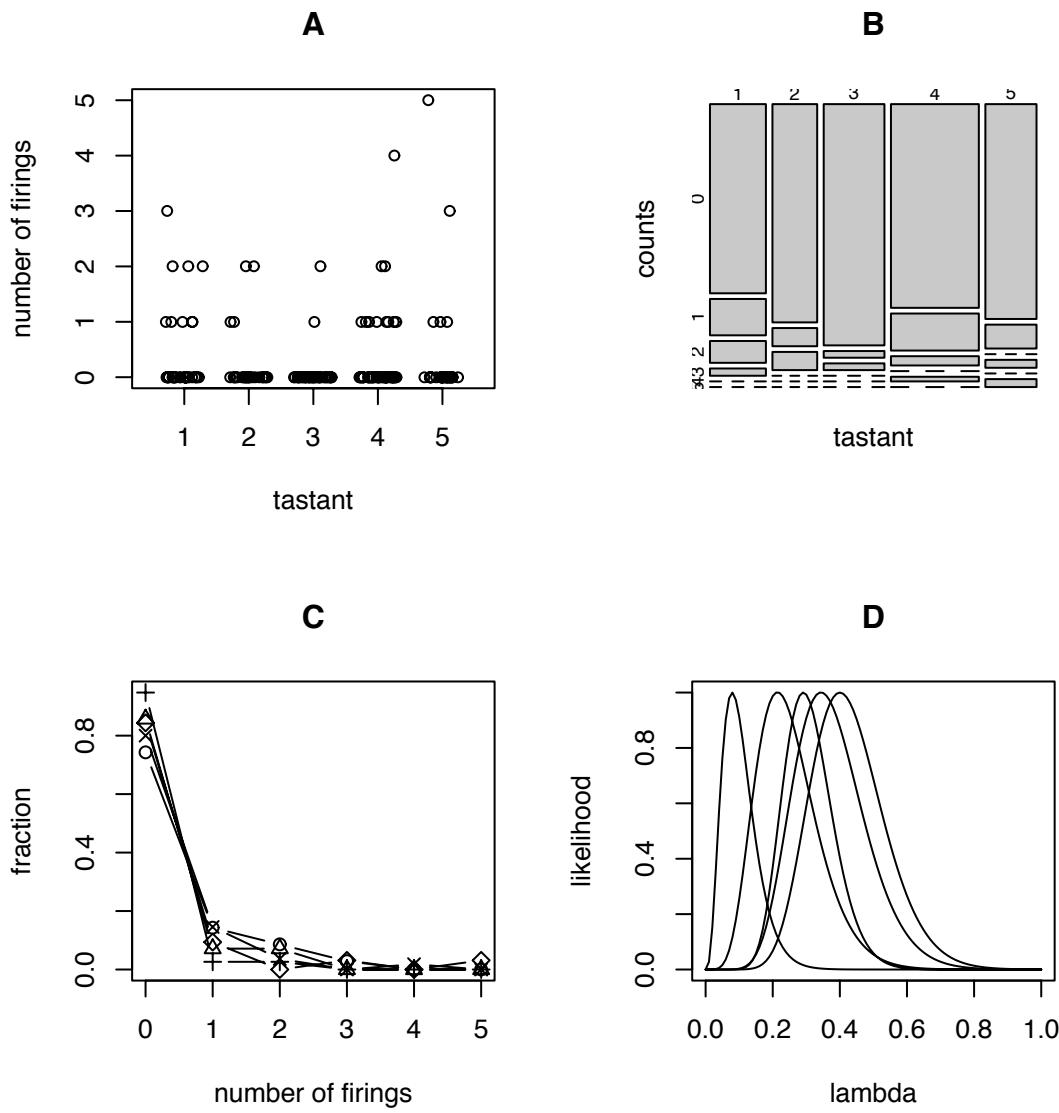


Figure 5.5: Numbers of firings of a neuron in 150 msec after five different tastants. **Tastants:** 1=MSG .1M; 2=MSG .3M; 3=NaCl .1M; 4=NaCl .3M; 5=water. **Panels:** A: A stripchart. Each circle represents one delivery of a tastant. B: A mosaic plot. C: Each line represents one tastant. D: Likelihood functions. Each line represents one tastant.

- The line `spiketable <- ...` creates a matrix to hold the data and illustrates the use of `dimnames` to name the dimensions. Some plotting commands use those names for labelling axes.
- The line `spiketable[i,] <- ...` shows an interesting use of the `hist` command. Instead of plotting a histogram it can simply return the counts.
- The line `freqtable <- ...` divides each row of the matrix by its sum, turning counts into proportions.

But let's investigate a little further. Do the data really follow a Poisson distribution? Figure 5.6 shows the  $\text{Poi}(.2)$  distribution while the circles show the actual fractions of firings. There is apparently good agreement. But numbers close to zero can be deceiving. The R command `dpois ( 0:5, .2 )` reveals that the probability of getting 5 spikes is less than 0.00001, assuming  $\lambda \approx 0.2$ . So either the  $\lambda_i$ 's are not all approximately .2, neuron spiking does not really follow a Poisson distribution, or we have witnessed a very unusual event.

Figure 5.6 was produced with the following snippet.

```
matplot (0:5, freqtable, pch=1, col=1,
 xlab="number of firings", ylab="fraction")
lines (0:5, dpois (0:5, 0.2))
```

Let us examine one aspect of Example 5.1 more closely. Rutherford and Geiger are counting  $\alpha$  particle emissions from a polonium source and find that the number of emissions in a fixed interval of time has a Poisson distribution. But they could have reasoned as follows: at the beginning of the experiment there is a fixed number of polonium atoms; each atom either decays or not; atoms decay independently of each other; therefore the number of decays, and hence the number of  $\alpha$  particles, has a Binomial distribution where  $n$  is the number of atoms and  $p$  is the probability that a given atom decays within the time interval.

Why did Rutherford and Geiger end up with the Poisson distribution and not the Binomial; are Rutherford and Geiger wrong? The answer is that a binomial distribution with large  $n$  and small  $p$  is extremely well approximated by a Poisson distribution with  $\lambda = np$ , so Rutherford and Geiger are correct, to a very high degree of accuracy. For a precise statement of this result we consider a sequence of random variables  $X_1 \sim \text{Bin}(n_1, p_1), X_2 \sim$

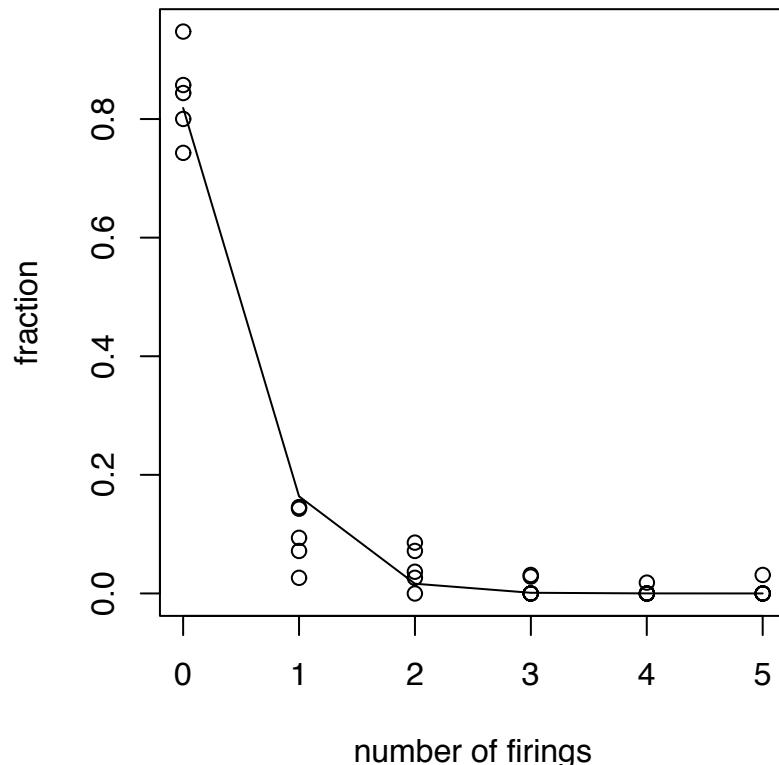


Figure 5.6: The line shows Poisson probabilities for  $\lambda = 0.2$ ; the circles show the fraction of times the neuron responded with 0, 1, ..., 5 spikes for each of the five tastants.

$\text{Bin}(n_2, p_2), \dots$  such that  $\lim_{i \rightarrow \infty} n_i = \infty$  and  $\lim_{i \rightarrow \infty} p_i = 0$ . But if the sequence were something like  $X_1 \sim \text{Bin}(100, .01)$ ,  $X_2 \sim \text{Bin}(1000, .005)$ ,  $X_3 \sim \text{Bin}(10000, .0025)$ ,  $\dots$  then the  $X_i$ 's would have expected values  $\mathbb{E}[X_1] = 1$ ,  $\mathbb{E}[X_2] = 5$ ,  $\mathbb{E}[X_3] = 25$ ,  $\dots$  and their distributions would not converge. To get convergence we need, at a minimum, for the expected values to converge. So let  $M$  be a fixed number and consider the sequence of random variables  $X_1 \sim \text{Bin}(n_1, p_1)$ ,  $X_2 \sim \text{Bin}(n_2, p_2)$ ,  $\dots$  where  $p_i = M/n_i$ . That way,  $\mathbb{E}[X_i] = M$ , for all  $i$ .

**Theorem 5.11.** *Let  $X_1, X_2, \dots$  be a sequence of random variables such that  $X_i \sim \text{Bin}(n_i, p_i)$  where for some constant  $M$ ,  $p_i = M/n_i$ . Then for any integer  $k \geq 0$*

$$\lim_{i \rightarrow \infty} P[X_i = k] = \frac{\exp^{-M} M^k}{k!}.$$

*I.e., the limit of the distributions of the  $X_i$ 's is the  $\text{Poi}(M)$  distribution.*

*Proof.* See Exercise xyz. □

## 5.4 The Uniform Distribution

**The Discrete Uniform Distribution** The discrete uniform distribution is the distribution that gives equal weight to each integer  $1, \dots, n$ . We write  $Y \sim \text{U}(1, n)$ . The pmf is

$$p(y) = 1/n \tag{5.2}$$

for  $y = 1, \dots, n$ . The discrete uniform distribution is used to model, for example, dice rolls, or any other experiment in which the outcomes are deemed equally likely. The only parameter is  $n$ . It is not an especially useful distribution in practical work but can be used to illustrate concepts in a simple setting. For an applied example see Exercise 23.

**The Continuous Uniform Distribution** The continuous uniform distribution is the distribution whose pdf is flat over the interval  $[a, b]$ . We write  $Y \sim \text{U}(a, b)$ . Although the notation might be confused with the discrete uniform, the context will indicate which is meant. The pdf is

$$p(y) = 1/(b - a)$$

for  $y \in [a, b]$ . The mean, variance, and moment generating function are left as Exercise 24.

Suppose we observe a random sample  $y_1, \dots, y_n$  from  $\text{U}(a, b)$ . What is the m.l.e.  $(\hat{a}, \hat{b})$ ? The joint density is

$$p(y_1, \dots, y_n) = \begin{cases} \left(\frac{1}{b-a}\right)^n & \text{if } a \leq y_{(1)} \text{ and } b \geq y_{(n)} \\ 0 & \text{otherwise} \end{cases}$$

which is maximized, as a function of  $(a, b)$ , if  $b - a$  is as small as possible without making the joint density 0. Thus,  $\hat{a} = y_{(1)}$  and  $\hat{b} = y_{(n)}$ .

## 5.5 The Gamma, Exponential, and Chi Square Distributions

“ $\Gamma$ ” is the upper case Greek letter Gamma. The *gamma function* is a special mathematical function defined on  $\mathbb{R}^+$  as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

Information about the gamma function can be found in mathematics texts and reference books. For our purposes, the key facts are:

$$\begin{aligned}\Gamma(\alpha + 1) &= \alpha\Gamma(\alpha) \quad \text{for } \alpha > 0 \\ \Gamma(n) &= (n - 1)! \quad \text{for positive integers } n \\ \Gamma(1/2) &= \sqrt{\pi}\end{aligned}$$

For any positive numbers  $\alpha$  and  $\beta$ , the  $\text{Gamma}(\alpha, \beta)$  distribution has pdf

$$p(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \quad \text{for } y \geq 0 \tag{5.3}$$

We write  $Y \sim \text{Gam}(\alpha, \beta)$ .

Figure 5.7 shows Gamma densities for four values of  $\alpha$  and four values of  $\beta$ .

- In each panel of Figure 5.7 the curves for different  $\alpha$ 's have different shapes. Sometimes  $\alpha$  is called the *shape* parameter of the Gamma distribution.
- The four panels look identical except for the axes. I.e., the four curves with  $\alpha = .5$ , one from each panel, have the same shape but different scales. The different scales correspond to different values of  $\beta$ . For this reason  $\beta$  is called a *scale* parameter. One can see directly from Equation 5.3 that  $\beta$  is a scale parameter because  $p(y)$  depends on  $y$  only through the ratio  $y/\beta$ . The idea of scale parameter is embodied in Theorem 5.12. See Section ?? for more on scale parameters.

Figure 5.7 was produced by the following snippet.

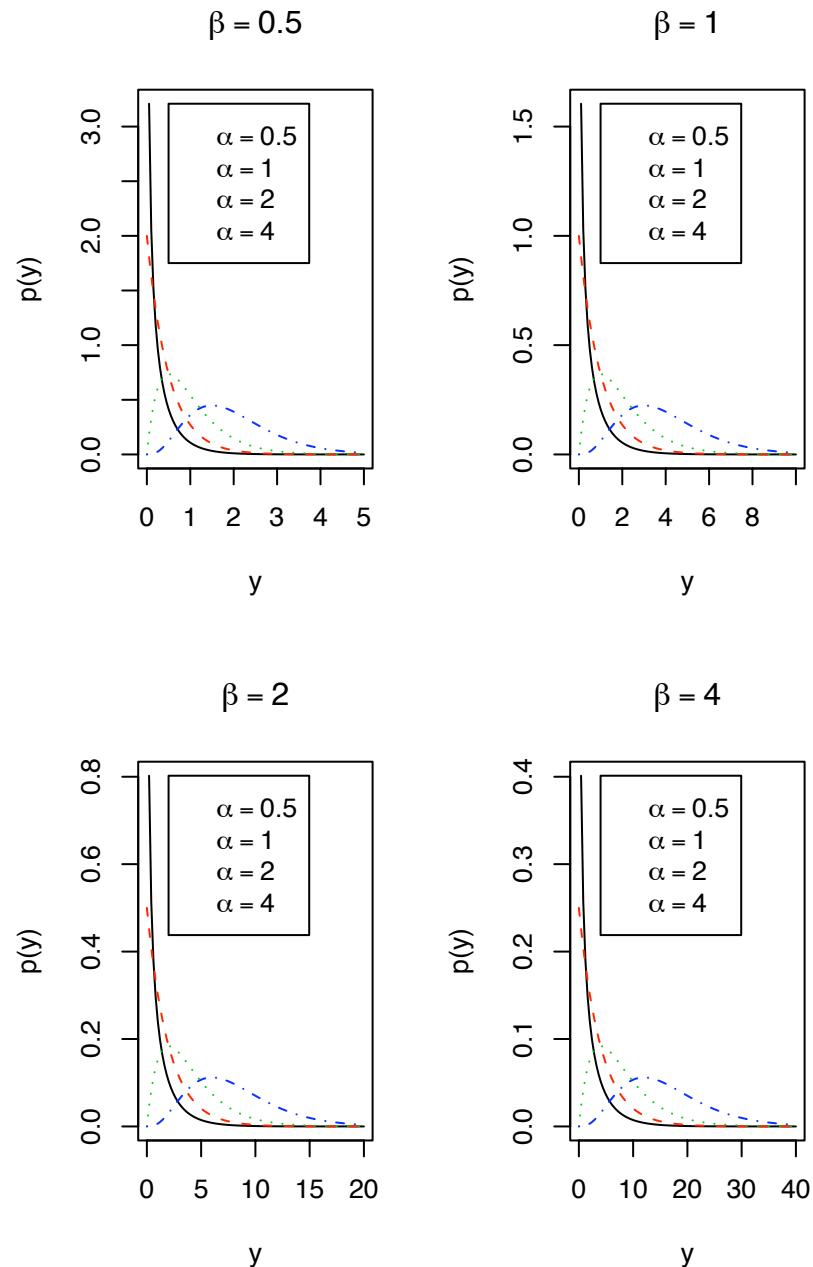


Figure 5.7: Gamma densities for various values of  $\alpha$  and  $\beta$ .

```

par (mfrow=c(2,2))

shape <- c (.5, 1, 2, 4)
scale <- c (.5, 1, 2, 4)
leg <- expression (alpha == .5, alpha == 1,
 alpha == 2, alpha == 4)

for (i in seq(along=scale)) {
 ymax <- scale[i]*max(shape) + 3*sqrt(max(shape))*scale[i]
 y <- seq (0, ymax, length=100)
 den <- NULL
 for (sh in shape)
 den <- cbind (den, dgamma(y,shape=sh,scale=scale[i]))
 matplot (y, den, type="l", main=letters[i], ylab="p(y)")
 legend (ymax*.1, max(den[den!=Inf]), legend = leg)
}

```

**Theorem 5.12.** Let  $X \sim \text{Gam}(\alpha, \beta)$  and let  $Y = cX$ . Then  $Y \sim \text{Gam}(\alpha, c\beta)$ .

*Proof.* Use Theorem 1.1.

$$p_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta},$$

Since  $Y = cX$ ,  $x = y/c$  and  $dx/dy = 1/c$ , so

$$\begin{aligned} p_Y(y) &= \frac{1}{c\Gamma(\alpha)\beta^\alpha} (y/c)^{\alpha-1} e^{-y/c\beta} \\ &= \frac{1}{\Gamma(\alpha)(c\beta)^\alpha} (y)^{\alpha-1} e^{-y/c\beta} \end{aligned}$$

which is the  $\text{Gam}(\alpha, c\beta)$  density. Also see Exercise 10.  $\square$

The mean, mgf, and variance are recorded in the next several theorems.

**Theorem 5.13.** Let  $Y \sim \text{Gam}(\alpha, \beta)$  Then  $\mathbb{E}[Y] = \alpha\beta$ .

*Proof.*

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^\infty y \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} dy \\ &= \frac{\Gamma(\alpha+1)\beta}{\Gamma(\alpha)} \int_0^\infty \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} y^\alpha e^{-y/\beta} dy \\ &= \alpha\beta.\end{aligned}$$

The last equality follows because (1)  $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ , and (2) the integrand is a Gamma density so the integral is 1. Also see Exercise 10.  $\square$

The last trick in the proof — recognizing an integrand as a density and concluding that the integral is 1 — is very useful. Here it is again.

**Theorem 5.14.** *Let  $Y \sim \text{Gam}(\alpha, \beta)$ . Then the moment generating function is  $M_Y(t) = (1 - t\beta)^{-\alpha}$  for  $t < 1/\beta$ .*

*Proof.*

$$\begin{aligned}M_Y(t) &= \int_0^\infty e^{ty} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} dy \\ &= \frac{\left(\frac{\beta}{1-t\beta}\right)^\alpha}{\beta^\alpha} \int_0^\infty \frac{1}{\Gamma(\alpha)\left(\frac{\beta}{1-t\beta}\right)^\alpha} y^{\alpha-1} e^{-y\frac{1-t\beta}{\beta}} dy \\ &= (1 - t\beta)^{-\alpha}\end{aligned}$$

$\square$

**Theorem 5.15.** *Let  $Y \sim \text{Gam}(\alpha, \beta)$ . Then*

$$\text{Var}(Y) = \alpha\beta^2$$

and

$$\text{SD}(Y) = \sqrt{\alpha}\beta.$$

*Proof.* See Exercise 11.  $\square$

**The Exponential Distribution** We often have to deal with situations such as

- the lifetime of an item
- the time until a specified event happens

The most fundamental probability distribution for such situations is the *exponential* distribution. Let  $Y$  be the time until the item dies or the event occurs. If  $Y$  has an exponential distribution then for some  $\lambda > 0$  the pdf of  $Y$  is

$$p_Y(y) = \lambda^{-1} e^{-y/\lambda} \quad \text{for } y \geq 0.$$

and we write  $Y \sim \text{Exp}(\lambda)$ . This density is pictured in Figure 5.8 (a repeat of Figure 1.7) for four values of  $\lambda$ . The exponential distribution is the special case of the Gamma distribution when  $\alpha = 1$ . The mean, SD, and mgf are given by Theorems 5.13 – 5.15.

Each exponential density has its maximum at  $y = 0$  and decreases monotonically. The value of  $\lambda$  determines the value  $p_Y(0|\lambda)$  and the rate of decrease. Usually  $\lambda$  is unknown. Small values of  $y$  are evidence for small values of  $\lambda$ ; large values of  $y$  are evidence for large values of  $\lambda$ .

### Example 5.3 (Radioactive Decay)

It is well known that some chemical elements are radioactive. Every atom of a radioactive element will eventually decay into smaller components. E.g., uranium-238 (by far the most abundant uranium isotope,  $^{238}\text{U}$ ) decays into thorium-234 and an  $\alpha$  particle while plutonium-239 (the isotope used in nuclear weapons,  $^{239}\text{Pu}$ ) decays into uranium-235 ( $^{235}\text{U}$ ) and an  $\alpha$  particle.

(See [HTTP://WWW.EPA.GOV/RADIATION/RADIONUCLIDES](http://www.epa.gov/RADIATION/RADIONUCLIDES) for more information.)

The time  $Y$  at which a particular atom decays is a random variable that has an exponential distribution. Each radioactive isotope has its own distinctive value of  $\lambda$ . A radioactive isotope is usually characterized by its median lifetime, or *half-life*, instead of  $\lambda$ . The half-life is the value  $m$  which satisfies  $P[Y \leq m] = P[Y \geq m] = 0.5$ . The half-life  $m$  can be found by solving

$$\int_0^m \lambda^{-1} e^{-y/\lambda} dy = 0.5.$$

The answer is  $m = \lambda \log 2$ . You will be asked to verify this claim in Exercise 31.

Uranium-238 has a half-life of 4.47 billion years. Thus its  $\lambda$  is about 6.45 billion. Plutonium-239 has a half-life of 24,100 years. Thus its  $\lambda$  is about 35,000.

Exponential distributions have an interesting and unique memoryless property. To demonstrate, we examine the  $\text{Exp}(\lambda)$  distribution as a model for  $T$ , the amount of time a computer Help line caller spends on hold. Suppose the caller has already spent  $t$  minutes on hold; i.e.,  $T \geq t$ . Let  $S$  be the remaining time on hold; i.e.,  $S = T - t$ . What is the

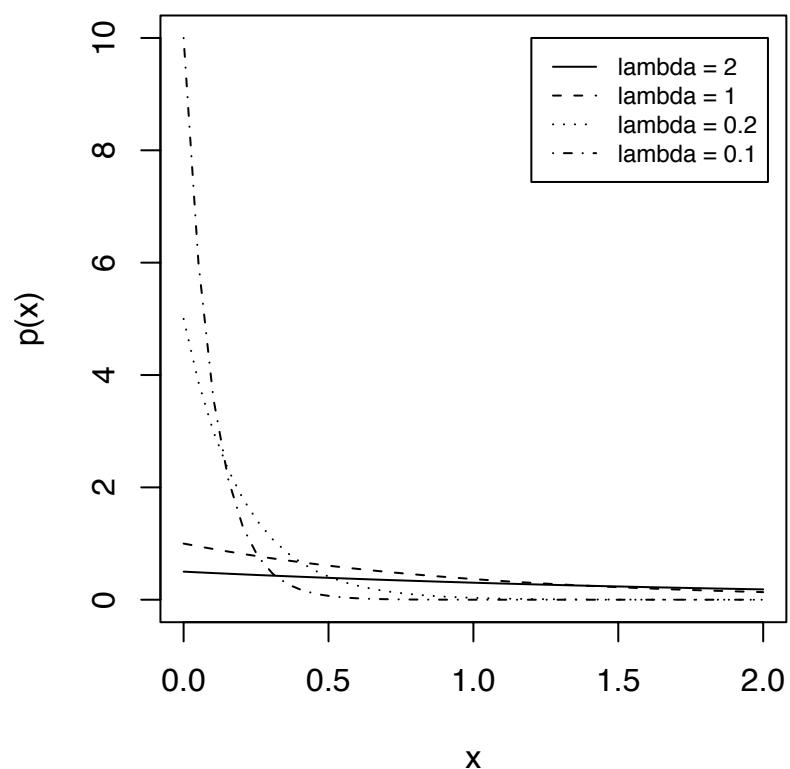


Figure 5.8: Exponential densities

distribution of  $S$  given  $T > t$ ? For any number  $r > 0$ ,

$$\begin{aligned} P[S > r | T \geq t] &= P[T \geq t + r | T \geq t] = \frac{P[T \geq t + r, T \geq t]}{P[T \geq t]} \\ &= \frac{P[T \geq t + r]}{P[T \geq t]} = \frac{e^{-(t+r)/\lambda}}{e^{-t/\lambda}} = e^{-r/\lambda}. \end{aligned}$$

In other words,  $S$  has an  $\text{Exp}(\lambda)$  distribution (Why?) that does not depend on the currently elapsed time  $t$  (Why?). This is a unique property of the Exponential distribution; no other continuous distribution has it. Whether it makes sense for the amount of time on hold is a question that could be verified by looking at data. If it's not sensible, then  $\text{Exp}(\lambda)$  is not an accurate model for  $T$ .

### Example 5.4

Some data here that don't look exponential

**The Poisson process** There is a close relationship between Exponential, Gamma, and Poisson distributions. For illustration consider a company's customer call center. Suppose that calls arrive according to a rate  $\lambda$  such that

1. in a time interval of length  $T$ , the number of calls is a random variable with distribution  $\text{Poi}(\lambda T)$  and
2. if  $I_1$  and  $I_2$  are disjoint time intervals then the number of calls in  $I_1$  is independent of the number of calls in  $I_2$ .

When calls arrive in this way we say the calls follow a *Poisson process*.

Suppose we start monitoring calls at time  $t_0$ . Let  $T_1$  be the time of the first call after  $t_0$  and  $Y_1 = T_1 - t_0$ , the time until the first call.  $T_1$  and  $Y_1$  are random variables. What is the distribution of  $Y_1$ ? For any positive number  $y$ ,

$$Pr[Y_1 > y] = Pr[\text{no calls in } [t_0, t_0 + y]] = e^{-\lambda y}$$

where the second equality follows by the Poisson assumption. But

$$Pr[Y_1 > y] = e^{-\lambda y} \Rightarrow Pr[Y \leq y] = 1 - e^{-\lambda y} \Rightarrow p_Y(y) = \lambda e^{-\lambda y} \Rightarrow Y_1 \sim \text{Exp}(1/\lambda)$$

What about the time to the second call? Let  $T_2$  be the time of the second call after  $t_0$  and  $Y_2 = T_2 - t_0$ . What is the distribution of  $Y_2$ ? For any  $y > 0$ ,

$$\begin{aligned} Pr[Y_2 > y] &= Pr[\text{fewer than 2 calls in } [t_0, y]] \\ &= Pr[0 \text{ calls in } [t_0, y]] + Pr[1 \text{ call in } [t_0, y]] \\ &= e^{-\lambda y} + y\lambda e^{-\lambda y} \end{aligned}$$

and therefore

$$p_{Y_2}(y) = \lambda e^{-\lambda y} - \lambda e^{-\lambda y} + y\lambda^2 e^{-\lambda y} = \frac{\lambda^2}{\Gamma(2)}ye^{-\lambda y}$$

so  $Y_2 \sim \text{Gam}(2, 1/\lambda)$ .

In general, the time  $Y_n$  until the  $n$ 'th call has the  $\text{Gam}(n, 1/\lambda)$  distribution. This fact is an example of the following theorem.

**Theorem 5.16.** *Let  $Y_1, \dots, Y_n$  be mutually independent and let  $Y_i \sim \text{Gam}(\alpha_i, \beta)$ . Then*

$$Y \equiv \sum Y_i \sim \text{Gam}(\alpha, \beta)$$

where  $\alpha \equiv \sum \alpha_i$ .

*Proof.* See Exercise 32. □

In Theorem 5.16 note that the  $Y_i$ 's must all have the same  $\beta$  even though they may have different  $\alpha_i$ 's.

### Poisson-Gamma conjugacy $\mathbf{F} = \mathbf{Gam}/\mathbf{Gam}$

**The Chi-squared Distribution** The Gamma distribution with  $\beta = 2$  and  $\alpha = p/2$  where  $p$  is a positive integer is called the *chi-squared distribution* with  $p$  degrees of freedom. We write  $Y \sim \chi_p^2$ .

**Theorem 5.17.** *Let  $Y_1, \dots, Y_n \sim \text{i.i.d. } N(0, 1)$ . Define  $X = \sum Y_i^2$ . Then  $X \sim \chi_n^2$ .*

*Proof.* This theorem will be proved in Section 5.7. □

## 5.6 The Beta Distribution

For positive numbers  $\alpha$  and  $\beta$ , the  $\text{Beta}(\alpha, \beta)$  distribution is a distribution for a random variable  $Y$  on the unit interval. The density is

$$p_Y(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1} \quad \text{for } y \in [0, 1]$$

The parameters are  $(\alpha, \beta)$ . We write  $Y \sim \text{Be}(\alpha, \beta)$ . The mean and variance are given by Theorem 5.18.

**Theorem 5.18.** *Let  $Y \sim \text{Be}(\alpha, \beta)$ . Then*

$$\begin{aligned} \mathbb{E}[Y] &= \frac{\alpha}{\alpha + \beta} \\ \text{Var}(Y) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{aligned}$$

*Proof.* See Exercise 28. □

Figure 5.9 shows some Beta densities. Each panel shows four densities having the same mean. It is evident from the Figure and the definition that the parameter  $\alpha$  ( $\beta$ ) controls whether the density rises or falls at the left (right). If both  $\alpha > 1$  and  $\beta > 1$  then  $p(y)$  is unimodal. The  $\text{Be}(1, 1)$  is the same as the  $\text{U}(0, 1)$  distribution.

The Beta distribution arises as the distribution of order statistics from the  $\text{U}(0, 1)$  distribution. Let  $x_1, \dots, x_n \sim \text{i.i.d. } \text{U}(0, 1)$ . What is the distribution of  $x_{(1)}$ , the first order statistic? Our strategy is first to find the cdf of  $x_{(1)}$ , then differentiate to get the pdf.

$$\begin{aligned} F_{X_{(1)}}(x) &= \text{P}[X_{(1)} \leq x] \\ &= 1 - \text{P}[\text{all } X_i \text{'s are greater than } x] \\ &= 1 - (1 - x)^n \end{aligned}$$

Therefore,

$$p_{X_{(1)}}(x) = \frac{d}{dx} F_{X_{(1)}}(x) = n(1 - x)^{n-1} = \frac{\Gamma(n+1)}{\Gamma(1)\Gamma(n)}(1 - x)^{n-1}$$

which is the  $\text{Be}(1, n)$  density. For the distribution of the largest order statistic see Exercise 29.

Figure 5.9 was produced by the following R snippet.

```
par (mfrow=c(3,1))
y <- seq (0, 1, length=100)
mean <- c (.2, .5, .9)
alpha <- c (.3, 1, 3, 10)
for (i in 1:3) {
 beta <- (alpha - mean[i]*alpha) / mean[i]
 den <- NULL
 for (j in 1:length(beta))
 den <- cbind (den, dbeta(y,alpha[j],beta[j]))
 matplot (y, den, type="l", main=letters[i], ylab="p(y)")
 if (i == 1)
 legend (.6, 8, paste ("(a,b) = (", round(alpha,2), ",",
 round(beta,2), ")", sep=""), lty=1:4)
 else if (i == 2)
 legend (.1, 4, paste ("(a,b) = (", round(alpha,2), ",",
 round(beta,2), ")", sep=""), lty=1:4)
 else if (i == 3)
```

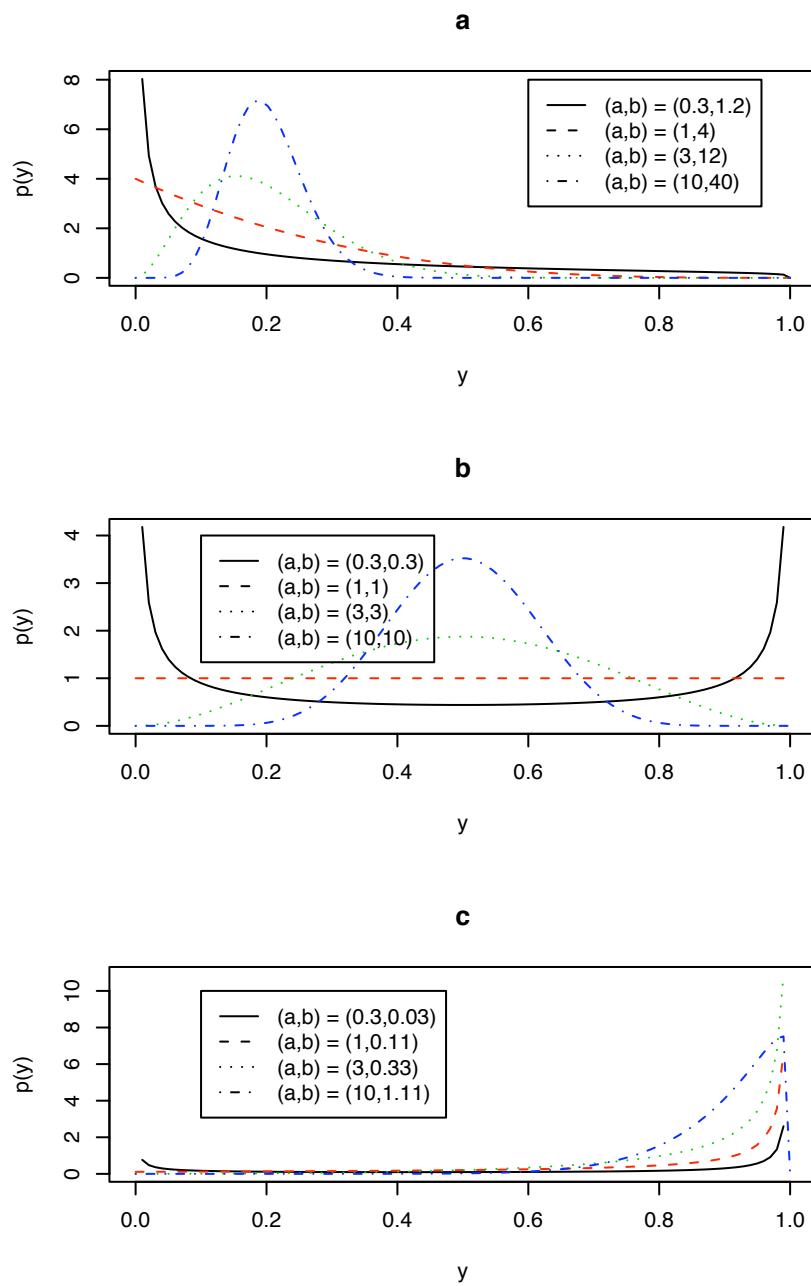


Figure 5.9: Beta densities — **a**: Beta densities with mean .2; **b**: Beta densities with mean .5; **c**: Beta densities with mean .9;

```
legend (.1, 10, paste ("(a,b) = (", round(alpha,2), ",",
round(beta,2), ")"), sep=""), lty=1:4)
```

The Beta density is closely related to the Gamma density by the following theorem.

**Theorem 5.19.** *Let  $X_1 \sim \text{Gam}(\alpha_1, \beta)$ ;  $X_2 \sim \text{Gam}(\alpha_2, \beta)$ ; and  $X_1 \perp X_2$ . Then*

$$Y \equiv \frac{X_1}{X_1 + X_2} \sim \text{Be}(\alpha_1, \alpha_2)$$

*Proof.* See Exercise 33. □

Note that Theorem 5.19 requires  $X_1$  and  $X_2$  both to have the same value of  $\beta$ , but the result doesn't depend on what that value is.

## 5.7 The Normal and Related Distributions

### 5.7.1 The Univariate Normal Distribution

The histograms in Figure 1.12 on page 31

- are approximately unimodal,
- are approximately symmetric,
- have different means, and
- have different standard deviations.

Data with these properties is ubiquitous in nature. Statisticians and other scientists often have to model similar looking data. One common probability density for modelling such data is the Normal density, also known as the *Gaussian density*. The Normal density is also important because of the Central Limit Theorem.

For some constants  $\mu \in \mathbb{R}$  and  $\sigma > 0$ , the Normal density is

$$p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \quad (5.4)$$

**Example 5.5** (Ocean temperatures, continued)

To see a Normal density in more detail, Figure 5.10 reproduces the top right histogram from Figure 1.12 redrawn with the Normal density overlaid, for the values  $\mu \approx 8.08$  and  $\sigma \approx 0.94$ . The vertical axis is drawn on the density scale. There are 112 temperature measurements that go into this histogram; they were recorded between 1949 and 1997; their latitudes are all between  $44^\circ$  and  $46^\circ$ ; their longitudes are all between  $-21^\circ$  and  $-19^\circ$ .

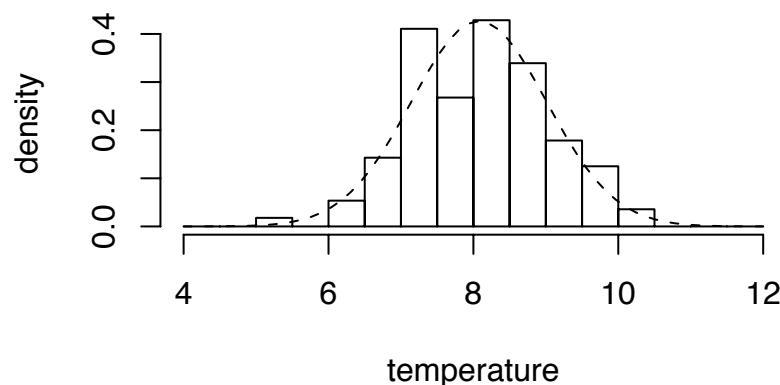


Figure 5.10: Water temperatures ( $^{\circ}\text{C}$ ) at 1000m depth,  $44 - 46^{\circ}\text{N}$  latitude and  $19 - 21^{\circ}\text{W}$  longitude. The dashed curve is the  $\text{N}(8.08, 0.94)$  density.

Figure 5.10 was produced by the following R snippet.

```
good <- abs (med.1000$lon - lons[3]) < 1 &
 abs (med.1000$lat - lats[1]) < 1
temp <- med.1000$temp[good]
hist (temp, xlim=c(4,12), breaks=seq(4,12,by=.5),
 freq=F, xlab="temperature", ylab="density", main = "")
mu <- mean (temp)
sig <- sqrt (var (temp))
x <- seq (4, 12, length=60)
```

```
lines (x, dnorm(x,mu,sig), lty=2)
```

Visually, the Normal density appears to fit the data well. Randomly choosing one of the 112 historical temperature measurements, or making a new measurement near 45°N and 20°W at a randomly chosen time are like drawing a random variable  $t$  from the  $N(8.08, 0.94)$  distribution.

Look at temperatures between 8.5° and 9.0°C. The  $N(8.08, 0.94)$  density says the probability that a randomly drawn temperature  $t$  is between 8.5° and 9.0°C is

$$P[t \in (8.5, 9.0)] = \int_{8.5}^{9.0} \frac{1}{\sqrt{2\pi} 0.94} e^{-\frac{1}{2}(\frac{t-8.08}{0.94})^2} dt \approx 0.16. \quad (5.5)$$

The integral in Equation 5.5 is best done on a computer, not by hand. In R it can be done with `pnorm(9.0, 8.08, .94) - pnorm(8.5, 8.08, .94)`. A fancier way to do it is `diff(pnorm(c(8.5, 9), 8.08, .94)))`.

- When  $x$  is a vector, `pnorm(x, mean, sd)` returns a vector of `pnorm`'s.
- When  $x$  is a vector, `diff(x)` returns the vector of differences  $x[2]-x[1]$ ,  $x[3]-x[2]$ , ...,  $x[n]-x[n-1]$ .

In fact, 19 of the 112 temperatures fell into that bin, and  $19/112 \approx 0.17$ , so the  $N(8.08, 0.94)$  density seems to fit very well.

However, the  $N(8.08, 0.94)$  density doesn't fit as well for temperatures between 7.5° and 8.0°C.

$$P[t \in (7.5, 8.0)] = \int_{7.5}^{8.0} \frac{1}{\sqrt{2\pi} 0.94} e^{-\frac{1}{2}(\frac{t-8.08}{0.94})^2} dt \approx 0.20.$$

In fact, 15 of the 112 temperatures fell into that bin; and  $15/112 \approx 0.13$ . Even so, the  $N(8.08, 0.94)$  density fits the data set very well.

**Theorem 5.20.** *Let  $Y \sim N(\mu, \sigma)$ . Then*

$$M_Y(t) = e^{\frac{\sigma^2 t^2}{2} + \mu t}.$$

*Proof.*

$$\begin{aligned}
M_Y(t) &= \int e^{ty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y^2 - (2\mu - 2\sigma^2 t)y + \mu^2)} dy \\
&= e^{-\frac{\mu^2}{2\sigma^2}} \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y - (\mu + \sigma^2 t))^2 + \frac{(\mu + \sigma^2 t)^2}{2\sigma^2}} dy \\
&= e^{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}} \\
&= e^{\frac{\sigma^2 t^2}{2} + \mu t}.
\end{aligned}$$

□

The technique used in the proof of Theorem 5.20 is worth remembering, so let's look at it more abstractly. Apart from multiplicative constants, the first integral in the proof is

$$\int e^{ty} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy = \int e^{-\frac{1}{2\sigma^2}(y-\mu)^2 + ty} dy$$

The exponent is quadratic in  $y$  and therefore, for some values of  $a, b, c, d, e$ , and  $f$  can be written

$$-\frac{1}{2\sigma^2}(y - \mu)^2 + ty = ay^2 + by + c = -\frac{1}{2} \left( \frac{y - d}{e} \right)^2 + f$$

This last expression has the form of a Normal distribution with mean  $d$  and SD  $e$ . So the integral can be evaluated by putting it in this form and manipulating the constants so it becomes the integral of a pdf and therefore equal to 1. It's a technique that is often useful when working with integrals arising from Normal distributions.

**Theorem 5.21.** Let  $Y \sim N(\mu, \sigma)$ . Then

$$\mathbb{E}[Y] = \mu \quad \text{and} \quad \text{Var}(Y) = \sigma^2.$$

*Proof.* For the mean,

$$\mathbb{E}[Y] = M'_Y(0) = (t\sigma^2 + \mu)e^{\frac{\sigma^2 t^2}{2} + \mu t} \Big|_{t=0} = \mu.$$

For the variance,

$$\mathbb{E}[Y^2] = M''_Y(0) = \sigma^2 e^{\frac{\sigma^2 t^2}{2} + \mu t} + (t\sigma^2 + \mu)^2 e^{\frac{\sigma^2 t^2}{2} + \mu t} \Big|_{t=0} = \sigma^2 + \mu^2.$$

So,

$$\text{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \sigma^2.$$

□

The  $N(0, 1)$  distribution is called the *standard Normal distribution*. As Theorem 5.22 shows, all Normal distributions are just shifted, rescaled versions of the standard Normal distribution. The mean is a location parameter; the standard deviation is a scale parameter. See Section ??.

**Theorem 5.22.**

1. If  $X \sim N(0, 1)$  and  $Y = \sigma X + \mu$  then  $Y \sim N(\mu, \sigma)$ .

2. If  $Y \sim N(\mu, \sigma)$  and  $X = (Y - \mu)/\sigma$  then  $X \sim N(0, 1)$ .

*Proof.* 1. Let  $X \sim N(0, 1)$  and  $Y = \sigma X + \mu$ . By Theorem 4.8

$$M_Y(t) = e^{\mu t} M_X(\sigma t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$$

2. Let  $Y \sim N(\mu, \sigma)$  and  $X = (Y - \mu)/\sigma$ . Then

$$M_X(t) = e^{-\mu t/\sigma} M_Y(t/\sigma) = e^{-\mu t/\sigma} e^{\frac{\sigma^2(t/\sigma)^2}{2} + \mu t/\sigma} = e^{\frac{t^2}{2}}$$

□

Section 5.5 introduced the  $\chi^2$  distribution, noting that it is a special case of the Gamma distribution. The  $\chi^2$  distribution arises in practice as a sum of squares of standard Normals. Here we restate Theorem 5.17, then prove it.

**Theorem 5.23.** Let  $Y_1, \dots, Y_n \sim i.i.d. N(0, 1)$ . Define  $X = \sum Y_i^2$ . Then  $X \sim \chi_n^2$ .

*Proof.* Start with the case  $n = 1$ .

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tY_1^2}] = \int e^{ty^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)y^2} dy \\ &= (1-2t)^{-1/2} \int \frac{\sqrt{1-2t}}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)y^2} dy \\ &= (1-2t)^{-1/2} \end{aligned}$$

So  $X \sim \text{Gam}(1/2, 2) = \chi_1^2$ .

If  $n > 1$  then by Corollary 4.10

$$M_X(t) = M_{Y_1^2 + \dots + Y_n^2}(t) = (1-2t)^{-n/2}$$

So  $X \sim \text{Gam}(n/2, 2) = \chi_n^2$ .

□

### 5.7.2 The Multivariate Normal Distribution

Let  $\vec{X}$  be an  $n$ -dimensional random vector with mean  $\mu_{\vec{X}}$  and covariance matrix  $\Sigma_{\vec{X}}$ . We say that  $\vec{X}$  has a multivariate Normal distribution if its joint density is

$$p_{\vec{X}}(\vec{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}-\mu_{\vec{X}})' \Sigma^{-1} (\vec{x}-\mu_{\vec{X}})} \quad (5.6)$$

where  $|\Sigma|$  refers to the determinant of the matrix  $\Sigma$ . We write  $\vec{X} \sim N(\mu, \Sigma)$ . Comparison of Equations 5.4 (page 311) and 5.6 shows that the latter is a generalization of the former. The multivariate version has the covariance matrix  $\Sigma$  in place of the scalar variance  $\sigma^2$ .

To become more familiar with the multivariate Normal distribution, we begin with the case where the covariance matrix is diagonal:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots \\ 0 & \sigma_2^2 & 0 & \cdots \\ 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \cdots & \sigma_n^2 \end{pmatrix}$$

In this case the joint density is

$$\begin{aligned} p_{\vec{X}}(\vec{x}) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}-\mu_{\vec{X}})' \Sigma^{-1} (\vec{x}-\mu_{\vec{X}})} \\ &= \left( \frac{1}{\sqrt{2\pi}} \right)^n \prod_{i=1}^n \left( \frac{1}{\sigma_i} \right) e^{-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2}} \\ &= \prod_{i=1}^n \left( \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2} \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2} \right), \end{aligned}$$

the product of  $n$  separate one dimensional Normal densities, one for each dimension. Therefore the  $X_i$ 's are independent and Normally distributed, with  $X_i \sim N(\mu_i, \sigma_i)$ . Also see Exercise 34.

When  $\sigma_1 = \cdots = \sigma_n = 1$ , then  $\Sigma$  is the  $n$ -dimensional identity matrix  $I_n$ . When, in addition,  $\mu_1 = \cdots = \mu_n = 0$ , then  $\vec{X} \sim N(0, I_n)$  and  $\vec{X}$  is said to have the *standard n-dimensional Normal distribution*.

Note: for two arbitrary random variables  $X_1$  and  $X_2$ ,  $X_1 \perp X_2$  implies  $\text{Cov}(X_1, X_2) = 0$ ; but  $\text{Cov}(X_1, X_2) = 0$  does not imply  $X_1 \perp X_2$ . However, if  $X_1$  and  $X_2$  are jointly Normally distributed then the implication is true. I.e. if  $(X_1, X_2) \sim N(\mu, \Sigma)$  and  $\text{Cov}(X_1, X_2) = 0$ , then  $X_1 \perp X_2$ . In fact, something stronger is true, as recorded in the next theorem.

**Theorem 5.24.** Let  $\vec{X} = (X_1, \dots, X_n) \sim N(\mu, \Sigma)$  where  $\Sigma$  has the so-called block-diagonal form

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0_{12} & \cdots & 0_{1m} \\ 0_{21} & \Sigma_{22} & \cdots & 0_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{m1} & \cdots & 0_{mm-1} & \Sigma_{mm} \end{pmatrix}$$

where  $\Sigma_{ii}$  is an  $n_i \times n_i$  matrix,  $0_{ij}$  is an  $n_i \times n_j$  matrix of 0's and  $\sum_i^m n_i = n$ . Partition  $\vec{X}$  to conform with  $\Sigma$  and define  $\vec{Y}_i$ 's:  $\vec{Y}_1 = (X_1, \dots, X_{n_1})$ ,  $\vec{Y}_2 = (X_{n_1+1}, \dots, X_{n_1+n_2})$ ,  $\dots$ ,  $\vec{Y}_m = (X_{n_1+\dots+n_{m-1}+1}, \dots, X_{n_m})$  and  $\nu_i$ 's:  $\nu_1 = (\mu_1, \dots, \mu_{n_1})$ ,  $\nu_2 = (\mu_{n_1+1}, \dots, \mu_{n_1+n_2})$ ,  $\dots$ ,  $\nu_m = (\mu_{n_1+\dots+n_{m-1}+1}, \dots, \mu_{n_m})$ . Then

1. The  $\vec{Y}_i$ 's are independent of each other, and
2.  $\vec{Y}_i \sim N(\nu_i, \Sigma_{ii})$

*Proof.* The transformation  $\vec{X} \rightarrow (\vec{Y}_1, \dots, \vec{Y}_m)$  is just the identity transformation, so

$$\begin{aligned} p_{\vec{Y}_1, \dots, \vec{Y}_m}(\vec{y}_1, \dots, \vec{y}_m) &= p_{\vec{X}}(\vec{y}_1, \dots, \vec{y}_m) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{y}-\mu)' \Sigma^{-1} (\vec{y}-\mu)} \\ &= \frac{1}{(2\pi)^{n/2} \prod_{i=1}^m |\Sigma_{ii}|^{\frac{1}{2}}} e^{-\frac{1}{2} \sum_{i=1}^m (\vec{y}_i - \nu_i)' \Sigma_{ii}^{-1} (\vec{y}_i - \nu_i)} \\ &= \prod_{i=1}^m \frac{1}{(2\pi)^{n_i/2} |\Sigma_{ii}|^{\frac{1}{2}}} e^{-\frac{1}{2} (\vec{y}_i - \nu_i)' \Sigma_{ii}^{-1} (\vec{y}_i - \nu_i)} \end{aligned}$$

□

To learn more about the multivariate Normal density, look at the curves on which  $p_{\vec{X}}$  is constant; i.e.,  $\{\vec{x} : p_{\vec{X}}(\vec{x}) = c\}$  for some constant  $c$ . The density depends on the  $x_i$ 's through the quadratic form  $(\vec{x} - \mu)' \Sigma^{-1} (\vec{x} - \mu)$ , so  $p_{\vec{X}}$  is constant where this quadratic form is constant. But when  $\Sigma$  is diagonal,  $(\vec{x} - \mu)' \Sigma^{-1} (\vec{x} - \mu) = \sum_1^n (x_i - \mu_i)^2 / \sigma_i^2$  so  $p_{\vec{X}}(\vec{x}) = c$  is the equation of an ellipsoid centered at  $\mu$  and with eccentricities determined by the ratios  $\sigma_i/\sigma_j$ .

What does this density look like? It is easiest to answer that question in two dimensions. Figure 5.11 shows three bivariate Normal densities. The left-hand column shows contour plots of the bivariate densities; the right-hand column shows samples from the joint distributions. In all cases,  $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$ . In the top row,  $\sigma_{X_1} = \sigma_{X_2} = 1$ ; in the second row,  $\sigma_{X_1} = 1; \sigma_{X_2} = 2$ ; in the third row,  $\sigma_{X_1} = 1/2; \sigma_{X_2} = 2$ . The standard

deviation is a scale parameter, so changing the SD just changes the scale of the random variable. That's what gives the second and third rows more vertical spread than the first, and makes the third row more horizontally squashed than the first and second.

Figure 5.11 was produced with the following R code.

```
par (mfrow=c(3,2)) # a 3 by 2 array of plots

x1 <- seq(-5,5,length=60)
x2 <- seq(-5,5,length=60)
den.1 <- dnorm (x1, 0, 1)
den.2 <- dnorm (x2, 0, 1)
den.jt <- den.1 %o% den.2
contour (x1, x2, den.jt, xlim=c(-5,5), ylim=c(-5,5), main="(a)",
 xlab=expression(x[1]), ylab=expression(x[2]))

samp.1 <- rnorm (300, 0, 1)
samp.2 <- rnorm (300, 0, 1)
plot (samp.1, samp.2, xlim=c(-5,5), ylim=c(-5,5), main="(b)",
 xlab=expression(x[1]), ylab=expression(x[2]), pch=".")

den.2 <- dnorm (x2, 0, 2)
den.jt <- den.1 %o% den.2
contour (x1, x2, den.jt, xlim=c(-5,5), ylim=c(-5,5), main="(c)",
 xlab=expression(x[1]), ylab=expression(x[2]),)

samp.2 <- rnorm (300, 0, 2)
plot (samp.1, samp.2, xlim=c(-5,5), ylim=c(-5,5), main="(d)",
 xlab=expression(x[1]), ylab=expression(x[2]), pch=".")

den.1 <- dnorm (x1, 0, .5)
den.jt <- den.1 %o% den.2
contour (x1, x2, den.jt, xlim=c(-5,5), ylim=c(-5,5), main="(e)",
 xlab=expression(x[1]), ylab=expression(x[2]))

samp.1 <- rnorm (300, 0, .5)
plot (samp.1, samp.2, xlim=c(-5,5), ylim=c(-5,5), main="(f)",
 xlab=expression(x[1]), ylab=expression(x[2]), pch=".")
```

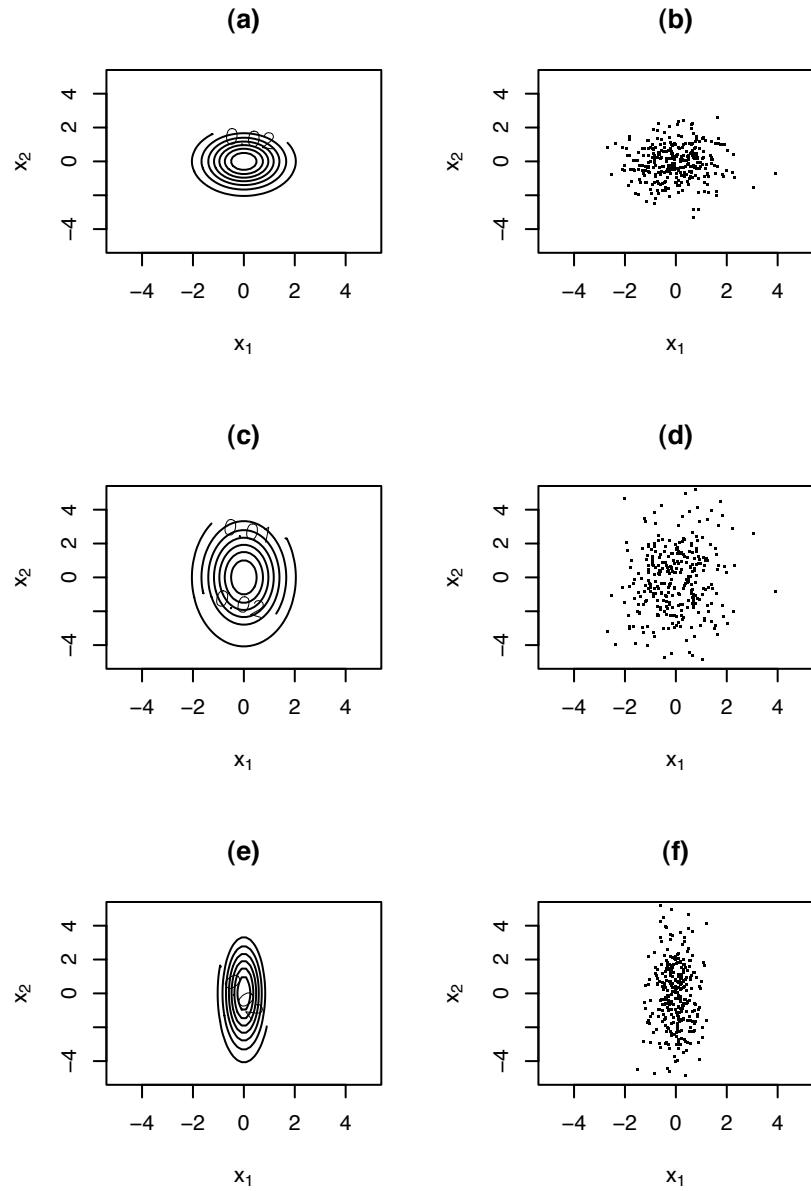


Figure 5.11: Bivariate Normal density.  $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$ .

**(a), (b):**  $\sigma_{X_1} = \sigma_{X_2} = 1$ .

**(c), (d):**  $\sigma_{X_1} = 1; \sigma_{X_2} = 2$ .

**(e), (f):**  $\sigma_{X_1} = 1/2; \sigma_{X_2} = 2$ .

**(a), (c), (e):** contours of the joint density.

**(b), (d), (f):** samples from the joint density.

- The code makes heavy use of the fact that  $X_1$  and  $X_2$  are independent for (a) calculating the joint density and (b) drawing random samples.
- `den.1 %% den.2` yields the *outer product* of `den.1` and `den.2`. It is a matrix whose  $ij$ 'th entry is `den.1[i] * den.2[j]`.

Now let's see what happens when  $\Sigma$  is not diagonal. Let  $\vec{Y} \sim N(\mu_{\vec{Y}}, \Sigma_{\vec{Y}})$ , so

$$p_{\vec{Y}}(\vec{y}) = \frac{1}{(2\pi)^{n/2} |\Sigma_{\vec{Y}}|^{\frac{1}{2}}} e^{-\frac{1}{2} (\vec{y} - \mu_{\vec{Y}})^T \Sigma_{\vec{Y}}^{-1} (\vec{y} - \mu_{\vec{Y}})},$$

and let  $\vec{X} \sim N(0, I_n)$ .  $\vec{X}$  is just a collection of independent  $N(0, 1)$  random variables. Its curves of constant density are just  $(n - 1)$ -dimensional spheres centered at the origin. Define  $\vec{Z} = \Sigma^{1/2} \vec{X} + \mu$ . We will show that  $p_{\vec{Z}} = p_{\vec{Y}}$ , therefore that  $\vec{Z}$  and  $\vec{Y}$  have the same distribution, and therefore that any multivariate Normal random vector has the same distribution as a linear transformation of a standard multivariate Normal random vector. To show  $p_{\vec{Z}} = p_{\vec{Y}}$  we apply Theorem 4.4. The Jacobian of the transformation from  $\vec{X}$  to  $\vec{Z}$  is  $|\Sigma|^{1/2}$ , the square root of the determinant of  $\Sigma$ . Therefore,

$$\begin{aligned} p_{\vec{Z}}(\vec{y}) &= p_{\vec{X}}\left(\Sigma^{-1/2}(\vec{y} - \mu)\right) |\Sigma|^{-1/2} \\ &= \frac{1}{\sqrt{2\pi^n}} e^{-1/2(\Sigma^{-1/2}(\vec{y} - \mu))^T (\Sigma^{-1/2}(\vec{y} - \mu))} |\Sigma|^{-1/2} \\ &= \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (\vec{y} - \mu)^T \Sigma^{-1} (\vec{y} - \mu)} \\ &= p_{\vec{Y}}(\vec{y}) \end{aligned}$$

The preceding result says that any multivariate Normal random variable,  $\vec{Y}$  in our notation above, has the same distribution as a linear transformation of a standard Normal random variable.

To see what multivariate Normal densities look like it is easiest to look at 2 dimensions. Figure 5.12 shows three bivariate Normal densities. The left-hand column shows contour plots of the bivariate densities; the right-hand column shows samples from the joint distributions. In all cases,  $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$  and  $\sigma_1 = \sigma_2 = 1$ . In the top row,  $\sigma_{1,2} = 0$ ; in the second row,  $\sigma_{1,2} = .5$ ; in the third row,  $\sigma_{1,2} = -.8$ .

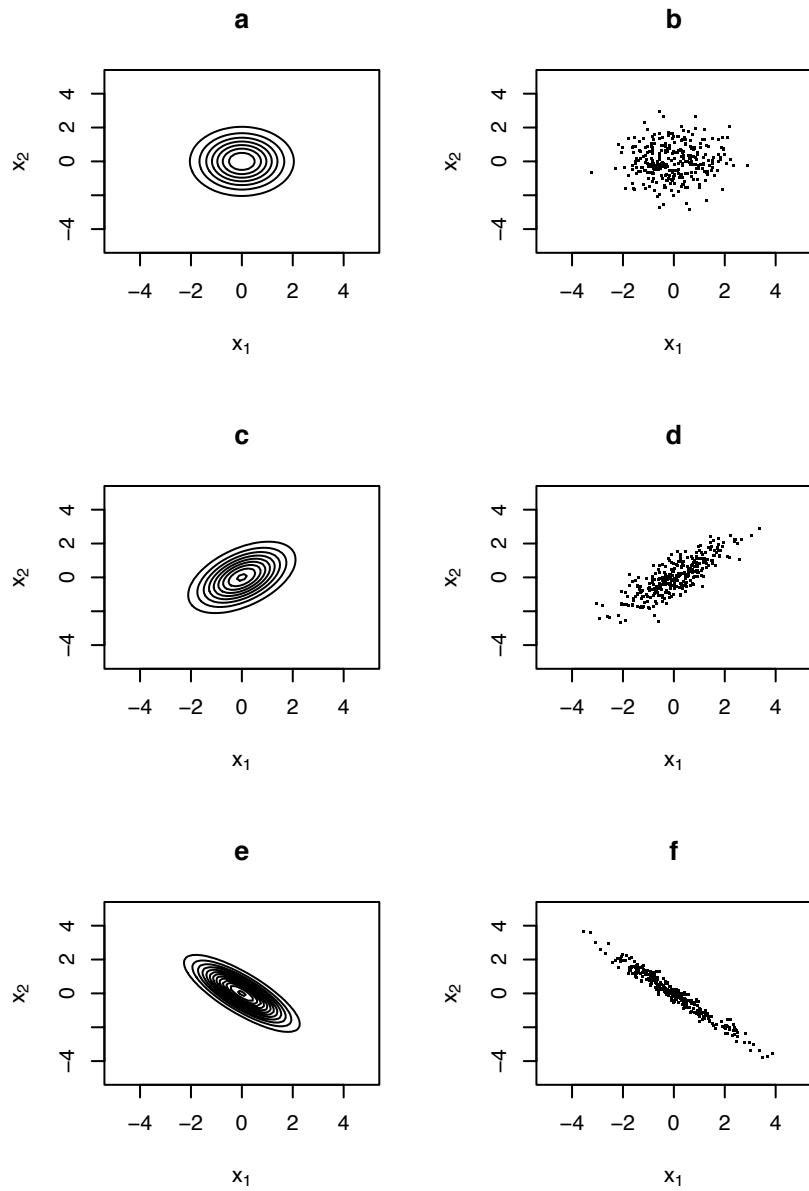


Figure 5.12: Bivariate Normal density.  $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$ ;  $\sigma_1 = \sigma_2 = 1$ .

**(a), (b):**  $\sigma_{1,2} = 0$ .

**(c), (d):**  $\sigma_{1,2} = .5$ .

**(e), (f):**  $\sigma_{1,2} = -.8$ .

**(a), (c), (e):** contours of the joint density.

**(b), (d), (f):** samples from the joint density.

Figure 5.12 was produced with the following R code.

```
par (mfrow=c(3,2)) # a 3 by 2 array of plots
npts <- 60
sampszie <- 300

x1 <- seq(-5,5,length=npts)
x2 <- seq(-5,5,length=npts)

Sigma <- array (NA, c(3,2,2))
Sigma[1,,] <- c(1,0,0,1)
Sigma[2,,] <- c(1,.5,.5,1)
Sigma[3,,] <- c(1,-.8,-.8,1)

den.jt <- matrix (NA, npts, npts)

for (i in 1:3)
 Sig <- Sigma[i,,]
 Siginv <- solve(Sig) # matrix inverse
 for (j in 1:npts)
 for (k in 1:npts)
 x <- c (x1[j], x2[k])
 den.jt[j,k] <- (1 / sqrt(2*pi*det(Sig))) *
 exp (-.5 * t(x) %*% Siginv %*% x)

contour (x1, x2, den.jt, xlim=c(-5,5), ylim=c(-5,5),
 drawlabels=F,
 xlab=expression(x[1]),
 ylab=expression(x[2]), main=letters[2*i-1])

samp <- matrix (rnorm (2*sampszie), 2, sampszie)
samp <- Sig %*% samp
plot (samp[1,], samp[2,], pch=".",
 xlim=c(-5,5), ylim=c(-5,5),
 xlab=expression(x[1]),
 ylab=expression(x[2]), main=letters[2*i])
```

### 5.7.3 Marginal, Conditional, and Related Distributions

We conclude this section with some theorems about Normal random variables that will prove useful later.

**Theorem 5.25.** *Let  $\vec{X} \sim N(\mu, \Sigma)$  be an  $n$ -dimensional Normal random variable; let  $A$  be a full rank  $n$  by  $n$  matrix; and let  $Y = AX$ . Then  $Y \sim N(A\mu, A\Sigma A^t)$ .*

*Proof.* By Theorem 4.4 (pg. 265),

$$\begin{aligned} p_{\vec{y}}(\vec{y}) &= p_{\vec{X}}(A^{-1}\vec{y})|A^{-1}| \\ &= \frac{1}{(2\pi)^{n/2}|A||\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(A^{-1}\vec{y}-\mu_{\vec{X}})^t \Sigma^{-1}(A^{-1}\vec{y}-\mu_{\vec{X}})} \\ &= \frac{1}{(2\pi)^{n/2}|A||\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(A^{-1}(\vec{y}-A\mu_{\vec{X}}))^t \Sigma^{-1}(A^{-1}(\vec{y}-A\mu_{\vec{X}}))} \\ &= \frac{1}{(2\pi)^{n/2}|A||\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{y}-A\mu_{\vec{X}})^t (A^{-1})^t \Sigma^{-1} A^{-1} (\vec{y}-A\mu_{\vec{X}})} \\ &= \frac{1}{(2\pi)^{n/2}|A\Sigma A^t|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{y}-A\mu_{\vec{X}})^t (A\Sigma A^t)^{-1} (\vec{y}-A\mu_{\vec{X}})} \end{aligned}$$

which we recognize as the  $N(A\mu, A\Sigma A^t)$  density.  $\square$

**Corollary 5.26.** *Let  $\vec{X} \sim N(\mu, \Sigma)$  be an  $n$ -dimensional Normal random variable; let  $A$  be a full rank  $n$  by  $n$  matrix; let  $b$  be a vector of length  $n$ ; and let  $Y = AX + b$ . Then  $Y \sim N(A\mu + b, A\Sigma A^t)$ .*

*Proof.* See Exercise 35.  $\square$

Let  $\vec{X} \sim N(\mu, \Sigma)$ . Divide  $\vec{X}$  into two parts:  $\vec{X} = \begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix}$  where  $\vec{X}_1$  consists of the first several components of  $\vec{X}$  and  $\vec{X}_2$  consists of the last several components. Then the mean and covariance matrix can be divided conformably:  $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ ;  $\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}$  where  $\Sigma_{2,1} = \Sigma_{1,2}^t$ . We often need to know the marginal distribution of  $\vec{X}_1$  or the conditional distribution of  $\vec{X}_2$  given  $\vec{X}_1$ . Those are given by Theorems 5.27 and 5.28. Our proofs follow ANDERSON [1984].

**Theorem 5.27.**  $\vec{X}_1 \sim N(\mu_1, \Sigma_{1,1})$ .

*Proof.* Let  $B = -\Sigma_{2,1}\Sigma_{1,1}^{-1}$  and define

$$\vec{Y} \equiv \begin{pmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{pmatrix} = \begin{pmatrix} I & 0 \\ B & I \end{pmatrix} \begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix}.$$

Theorem 5.25 says  $\vec{Y}$  has a Normal distribution. The mean is

$$\mathbb{E}(\vec{Y}) = \begin{pmatrix} I & 0 \\ B & I \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ B\mu_1 + \mu_2 \end{pmatrix}.$$

The covariance matrix is

$$\begin{aligned} \text{Cov}(\vec{Y}) &= \begin{pmatrix} I & 0 \\ B & I \end{pmatrix} \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix} \begin{pmatrix} I & 0 \\ B & I \end{pmatrix}^t \\ &= \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ B\Sigma_{1,1} + \Sigma_{2,1} & B\Sigma_{1,2} + \Sigma_{2,2} \end{pmatrix} \begin{pmatrix} I & B^t \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,1}B^t + \Sigma_{1,2} \\ B\Sigma_{1,1} + \Sigma_{2,1} & B\Sigma_{1,1}B^t + \Sigma_{2,1}B^t + B\Sigma_{1,2} + \Sigma_{2,2} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{1,1} & 0 \\ 0 & \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2} \end{pmatrix}. \end{aligned}$$

Then Theorem 5.24 says  $\vec{Y}_1 \sim N(\mu_1, \Sigma_{1,1})$ . It also says  $\vec{Y}_2 \sim N(B\mu_1 + \mu_2, \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})$ , a fact we shall use in the proof of Theorem 5.28  $\square$

**Theorem 5.28.**  $\vec{X}_2 | \vec{X}_1 \sim N(\mu_2 + \Sigma_{2,1}\Sigma_{1,1}^{-1}(\vec{X}_1 - \mu_1), \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})$ .

*Proof.* Let  $B = -\Sigma_{2,1}\Sigma_{1,1}^{-1}$  and define

$$\vec{Y} \equiv \begin{pmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{pmatrix} = \begin{pmatrix} I & 0 \\ B & I \end{pmatrix} \begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix}.$$

The distribution of  $\vec{Y}$  was worked out in the proof of Theorem 5.27. Here we work backwards and derive the joint density of  $\begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix}$ , then divide by the marginal density of  $\vec{X}_1$  to get the density of  $\vec{X}_2 | \vec{X}_1$ . This may seem strange, because we already know the joint density of  $\begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix}$ . However, in this proof we shall write the joint density in a different way.

Because the Jacobian of the transformation is 1, the joint density  $p_{\vec{X}_1, \vec{X}_2}(\vec{x}_1, \vec{x}_2)$  is found by writing the joint density  $p_{\vec{Y}_1, \vec{Y}_2}(\vec{y}_1, \vec{y}_2) = p_{\vec{Y}_1}(\vec{y}_1) \cdot p_{\vec{Y}_2 | \vec{Y}_1}(\vec{y}_2 | \vec{y}_1) = p_{\vec{Y}_1}(\vec{y}_1) \cdot p_{\vec{Y}_2}(\vec{y}_2)$  (because

$\vec{Y}_1 \perp \vec{Y}_2$ ), then substituting  $\vec{x}_1$  for  $\vec{y}_1$  and  $B\vec{x}_1 + \vec{x}_2$  for  $\vec{y}_2$ . Let  $n_2$  be the length of  $\vec{X}_2$  and  $\Sigma_{2,2|1} = \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}$ . Then,

$$\begin{aligned} p_{\vec{X}_2|\vec{X}_1}(\vec{x}_2 | \vec{x}_1) &= \frac{p_{\vec{X}_1, \vec{X}_2}(\vec{x}_1, \vec{x}_2)}{p_{\vec{X}_1}(\vec{x}_1)} \\ &= \frac{p_{\vec{Y}_1, \vec{Y}_2}(\vec{x}_1, B\vec{x}_1 + \vec{x}_2)}{p_{\vec{Y}_1}(\vec{x}_1)} \\ &= \frac{p_{\vec{Y}_1}(\vec{x}_1)p_{\vec{Y}_2}(B\vec{x}_1 + \vec{x}_2)}{p_{\vec{Y}_1}(\vec{x}_1)} \\ &= p_{\vec{Y}_2}(B\vec{x}_1 + \vec{x}_2) \\ &= \frac{1}{(2\pi)^{\frac{n_2}{2}} |\Sigma_{2,2|1}|^{\frac{1}{2}}} e^{-\frac{1}{2}[(B\vec{x}_1 + \vec{x}_2 - (B\mu_1 + \mu_2))^t \Sigma_{2,2|1}^{-1} (B\vec{x}_1 + \vec{x}_2 - (B\mu_1 + \mu_2))]} \\ &= \frac{1}{(2\pi)^{\frac{n_2}{2}} |\Sigma_{2,2|1}|^{\frac{1}{2}}} e^{-\frac{1}{2}[(\vec{x}_2 - (\mu_2 - B(\vec{x}_1 - \mu_1)))^t \Sigma_{2,2|1}^{-1} (\vec{x}_2 - (\mu_2 - B(\vec{x}_1 - \mu_1)))]}, \end{aligned}$$

which is the Normal density with mean  $\mathbb{E}[\vec{X}_2 | \vec{X}_1] = \mu_2 - B(\vec{x}_1 - \mu_1) = \mu_2 + \Sigma_{2,1}\Sigma_{1,1}^{-1}(\vec{x}_1 - \mu_1)$  and covariance matrix  $\text{Cov}(\vec{X}_2 | \vec{X}_1) = \Sigma_{2,2|1} = \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}$ .  $\square$

**Theorem 5.29.** Let  $X_1, \dots, X_n \sim i.i.d. N(\mu, \sigma)$ . Define  $S^2 \equiv \sum_{i=1}^n (X_i - \bar{X})^2$ . Then  $\bar{X} \perp S^2$ .

*Proof.* Define the random vector  $\vec{Y} = (Y_1, \dots, Y_n)^t$  by

$$\begin{aligned} Y_1 &= X_1 - \bar{X} \\ Y_2 &= X_2 - \bar{X} \\ &\vdots \\ Y_{n-1} &= X_{n-1} - \bar{X} \\ Y_n &= \bar{X} \end{aligned}$$

The proof follows these steps.

1.  $S^2$  is a function only of  $(Y_1, \dots, Y_{n-1})^t$ ; i.e. not a function of  $Y_n$ .
2.  $(Y_1, \dots, Y_{n-1})^t \perp Y_n$ .
3. Therefore  $S^2 \perp Y_n$ .

1.  $\sum_{i=1}^n (X_i - \bar{X}) = 0$ . Therefore,  $(X_n - \bar{X}) = -\sum_{i=1}^{n-1} (X_i - \bar{X})$ . And therefore

$$S^2 = \sum_{i=1}^{n-1} (X_i - \bar{X})^2 + \left( \sum_{i=1}^{n-1} (X_i - \bar{X}) \right)^2 = \sum_{i=1}^{n-1} Y_i^2 + \left( \sum_{i=1}^{n-1} Y_i \right)^2$$

is a function of  $(Y_1, \dots, Y_{n-1})^t$ .

2.

$$\vec{Y} = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1 - \frac{1}{n} & -\frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \vec{X} \equiv A \vec{X}$$

where the matrix  $A$  is defined by the preceding equation, so

$\vec{Y} \sim N(A\mu, \sigma^2 AA')$ . The first  $n - 1$  rows of  $A$  are each orthogonal to the last row. Therefore

$$AA' = \begin{pmatrix} \Sigma_{11} & \vec{0} \\ \vec{0}' & 1/n \end{pmatrix}$$

where  $\Sigma_{11}$  has dimension  $(n - 1) \times (n - 1)$  and  $\vec{0}$  is the  $(n - 1)$ -dimensional vector of 0's. Thus, by Theorem 5.24,  $(Y_1, \dots, Y_{n-1})^t \perp Y_n$ .

3. Follows immediately from 1 and 2.

□

## 5.8 The $t$ Distribution

The  $t$  distribution arises when making inference about the mean of a Normal distribution.

Let  $X_1, \dots, X_n \sim$  i.i.d.  $N(\mu, \sigma)$  where both  $\mu$  and  $\sigma$  are unknown, and suppose our goal is to estimate  $\mu$ .  $\hat{\mu} = \bar{X}$  is a sensible estimator. Its sampling distribution is  $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$  or, equivalently,

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

We would like to use this equation to tell us how accurately we can estimate  $\mu$ . Apparently we can estimate  $\mu$  to within about  $\pm 2\sigma/\sqrt{n}$  most of the time. But that's not an immediately

useful statement because we don't know  $\sigma$ . So we estimate  $\sigma$  by  $\hat{\sigma} = \left( n^{-1} \sum(X_i - \bar{X})^2 \right)^{1/2}$  and say

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim N(0, 1),$$

approximately. This section derives the exact distribution of  $(\bar{X} - \mu)/(\hat{\sigma}/\sqrt{n})$  and assesses how good the Normal approximation is. We already know from Corollary 5.29 that  $\bar{X} \perp \hat{\sigma}$ . Theorem 5.31 gives the distribution of  $S^2 = n\hat{\sigma}^2 = \sum(X_i - \bar{X})^2$ . First we need a lemma.

**Lemma 5.30.** *Let  $V = V_1 + V_2$  and  $W = W_1 + W_2$  where  $V_1 \perp V_2$  and  $W_1 \perp W_2$ . If  $V$  and  $W$  have the same distribution, and if  $V_1$  and  $W_1$  have the same distribution, then  $V_2$  and  $W_2$  have the same distribution.*

*Proof.* Using moment generating functions,

$$\begin{aligned} M_{V_2}(t) &= M_V(t)/M_{V_1}(t) \\ &= M_W(t)/M_{W_1}(t) \\ &= M_{W_2}(t) \end{aligned}$$

□

**Theorem 5.31.** *Let  $X_1, \dots, X_n \sim i.i.d. N(\mu, \sigma)$ . Define  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ . Then*

$$\frac{S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

*Proof.* Let

$$V = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2.$$

Then  $V \sim \chi_n^2$  and

$$\begin{aligned} V &= \sum_{i=1}^n \left( \frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 + 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X}) \\ &= \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \\ &\equiv \frac{S^2}{\sigma^2} + V_2 \end{aligned}$$

where  $S^2/\sigma^2 \perp V_2$  and  $V_2 \sim \chi_1^2$ . But also,

$$V = \sum_{i=1}^{n-1} \left( \frac{X_i - \mu}{\sigma} \right)^2 + \left( \frac{X_n - \mu}{\sigma} \right)^2 \equiv W_1 + W_2$$

where  $W_1 \perp W_2$ ,  $W_1 \sim \chi_{n-1}^2$  and  $W_2 \sim \chi_1^2$ . Now the conclusion follows by Lemma 5.30.  $\square$

Define

$$T \equiv \sqrt{\frac{n-1}{n}} \left( \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \right) = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{S^2/(n-1)\sigma^2}}.$$

Then by Corollary 5.29 and Theorem 5.31,  $T$  has the distribution of  $U/\sqrt{V/(n-1)}$  where  $U \sim N(0, 1)$ ,  $V \sim \chi_{n-1}^2$ , and  $U \perp V$ . This distribution is called the  $t$  distribution with  $n-1$  degrees of freedom. We write  $T \sim t_{n-1}$ . Theorem 5.32 derives its density.

**Theorem 5.32.** *Let  $U \sim N(0, 1)$ ,  $V \sim \chi_p^2$ , and  $U \perp V$ . Then  $T \equiv U/\sqrt{V/p}$  has density*

$$p_T(t) = \frac{\Gamma(\frac{p+1}{2}) p^{\frac{p}{2}}}{\Gamma(\frac{p}{2}) \sqrt{\pi}} (t^2 + p)^{-\frac{p+1}{2}} = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2}) \sqrt{p\pi}} \left( 1 + \frac{t^2}{p} \right)^{-\frac{p+1}{2}}.$$

*Proof.* Define

$$T = \frac{U}{\sqrt{V/p}} \quad \text{and} \quad Y = V$$

We make the transformation  $(U, V) \rightarrow (T, Y)$ , find the joint density of  $(T, Y)$ , and then the marginal density of  $T$ . The inverse transformation is

$$U = \frac{TY^{\frac{1}{2}}}{\sqrt{p}} \quad \text{and} \quad V = Y$$

The Jacobian is

$$\begin{vmatrix} \frac{dU}{dT} & \frac{dU}{dY} \\ \frac{dV}{dT} & \frac{dV}{dY} \end{vmatrix} = \begin{vmatrix} \frac{Y^{\frac{1}{2}}}{\sqrt{p}} & \frac{TY^{-\frac{1}{2}}}{2\sqrt{p}} \\ 0 & 1 \end{vmatrix} = \frac{Y^{\frac{1}{2}}}{\sqrt{p}}$$

The joint density of  $(U, V)$  is

$$p_{U,V}(u, v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \frac{1}{\Gamma(\frac{p}{2}) 2^{\frac{p}{2}}} v^{\frac{p}{2}-1} e^{-\frac{v}{2}}.$$

Therefore the joint density of  $(T, Y)$  is

$$p_{T,Y}(t, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2p}} \frac{1}{\Gamma(\frac{p}{2}) 2^{\frac{p}{2}}} y^{\frac{p}{2}-1} e^{-\frac{y}{2}} \frac{y^{\frac{1}{2}}}{\sqrt{p}}$$

and the marginal density of  $T$  is

$$\begin{aligned} p_T(t) &= \int p_{T,Y}(t, y) dy \\ &= \frac{1}{\sqrt{2\pi} \Gamma(\frac{p}{2}) 2^{\frac{p}{2}} \sqrt{p}} \int_0^\infty y^{\frac{p+1}{2}-1} e^{-\frac{y}{2}(\frac{t^2}{p}+1)} dy \\ &= \frac{1}{\sqrt{\pi} \Gamma(\frac{p}{2}) 2^{\frac{p+1}{2}} \sqrt{p}} \Gamma(\frac{p+1}{2}) \left( \frac{2p}{t^2+p} \right)^{\frac{p+1}{2}} \\ &\quad \times \int_0^\infty \frac{1}{\Gamma(\frac{p+1}{2}) \left( \frac{2p}{t^2+p} \right)^{\frac{p+1}{2}}} y^{\frac{p+1}{2}-1} e^{-\frac{y}{2p/(t^2+p)}} dy \\ &= \frac{\Gamma(\frac{p+1}{2}) p^{p/2}}{\Gamma(\frac{p}{2}) \sqrt{\pi}} \left( t^2 + p \right)^{-\frac{p+1}{2}} \\ &= \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2}) \sqrt{p\pi}} \left( 1 + \frac{t^2}{p} \right)^{-\frac{p+1}{2}}. \end{aligned}$$

□

Figure 5.8 shows the  $t$  density for 1, 4, 16, and 64 degrees of freedom, and the  $N(0, 1)$  density. The two points to note are

1. The  $t$  densities are unimodal and symmetric about 0, but have less mass in the middle and more mass in the tails than the  $N(0, 1)$  density.
2. In the limit, as  $p \rightarrow \infty$ , the  $t_p$  density appears to approach the  $N(0, 1)$  density. (The appearance is correct. See Exercise 36.)

Figure 5.8 was produced with the following snippet.

```
x <- seq(-5, 5, length=100)
dens <- cbind(dt(x, 1), dt(x, 4), dt(x, 16), dt(x, 64),
 dnorm(x))
```

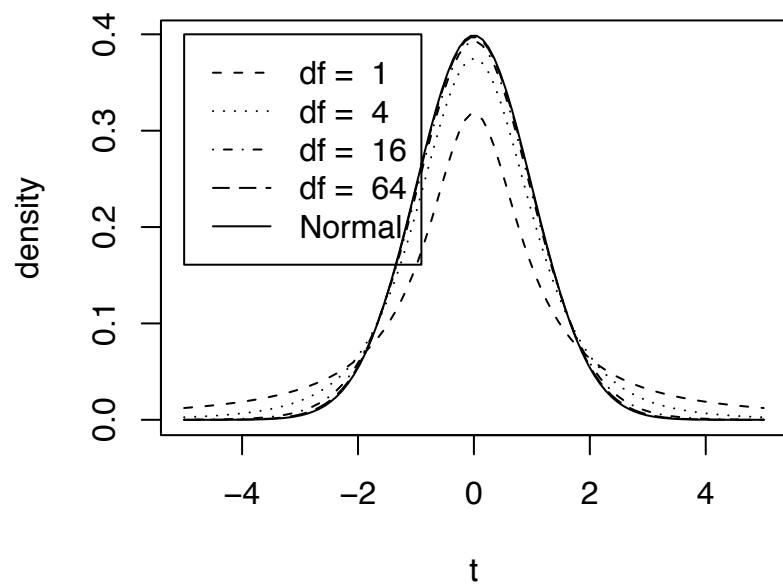


Figure 5.13:  $t$  densities for four degrees of freedom and the  $N(0, 1)$  density

```

matplot (x, dens, type="l", ylab="density", xlab="t",
 lty=c(2:5,1), col=1)
legend (x=-5, y=.4, lty=c(2:5,1),
 legend=c (paste("df = ", c(1,4,16,64)),
 "Normal"))

```

At the beginning of Section ?? we said the quantity  $\sqrt{n}(\bar{X} - \mu)/\hat{\sigma}$  had a  $N(0, 1)$  distribution, approximately. Theorem 5.32 derives the density of the related quantity  $\sqrt{n-1}(\bar{X} - \mu)/\hat{\sigma}$  which has a  $t_{n-1}$  distribution, exactly. Figure 5.8 shows how similar those distributions are. The  $t$  distribution has slightly more spread than the  $N(0, 1)$  distribution, reflecting the fact that  $\sigma$  has to be estimated. But when  $n$  is large, i.e. when  $\sigma$  is well estimated, then the two distributions are nearly identical.

If  $T \sim t_p$ , then

$$\mathbb{E}[T] = \int_{-\infty}^{\infty} t \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2}) \sqrt{p\pi}} \left(1 + \frac{t^2}{p}\right)^{-\frac{p+1}{2}} dt \quad (5.7)$$

In the limit as  $t \rightarrow \infty$ , the integrand behaves like  $t^{-p}$ ; hence 5.7 is integrable if and only if  $p > 1$ . Thus the  $t_1$  distribution, also known as the *Cauchy distribution*, has no mean. When  $p > 1$ ,  $\mathbb{E}[T] = 0$ , by symmetry. By a similar argument, the  $t_p$  distribution has a variance if and only if  $p > 2$ . When  $p > 2$ , then  $\text{Var}(T) = p/(p-2)$ . In general,  $T$  has a  $k$ -th moment ( $\mathbb{E}[T^k] < \infty$ ) if and only if  $p > k$ .

## 5.9 Exercises

1. Prove Theorem 5.4 by moment generating functions.
2. Refer to Theorem 5.8.
  - (a) What was the point of the next to last step?
  - (b) Justify the last step.
3. Assume that all players on a basketball team are 70% free throw shooters and that free throws are independent of each other.
  - (a) The team takes 40 free throws in a game. Write down a formula for the probability that they make exactly 37 of them. You do not need to evaluate the formula.

- (b) The team takes 20 free throws the next game. Write down a formula for the probability that they make exactly 9 of them.
- (c) Write down a formula for the probability that the team makes exactly 37 free throws in the first game and exactly 9 in the second game. That is, write a formula for the probability that they accomplish both feats.
4. Explain the role of `qnbinom(...)` in the R code for Figure 5.2.
5. Write down the distribution you would use to model each of the following random variables. Be as specific as you can. I.e., instead of answering “Poisson distribution”, answer “`Poi(3)`” or instead of answering “Binomial”, answer “`Bin(n, p)` where  $n = 13$  but  $p$  is unknown.”
- The temperature measured at a randomly selected point on the surface of Mars.
  - The number of car accidents in January at the corner of Broad Street and Main Street.
  - Out of 20 people in a post office, the number who, when exposed to anthrax spores, actually develop anthrax.
  - Out of 10,000 people given a smallpox vaccine, the number who develop smallpox.
  - The amount of Mercury in a fish caught in Lake Ontario.
6. A student types `dpois(3, 1.5)` into R. R responds with 0.1255107.
- Write down in words what the student just calculated.
  - Write down a mathematical formula for what the student just calculated.
7. Name the distribution. Your answers should be of the form  $\text{Poi}(\lambda)$  or  $\text{N}(3, 22)$ , etc. Use numbers when parameters are known, symbols when they’re not.
- You spend the evening at the roulette table in a casino. You bet on red 100 times. Each time the chance of winning is  $18/38$ . If you win, you win \$1; if you lose, you lose \$1. The average amount of time between bets is 90 seconds; the standard deviation is 5 seconds.
- the number of times you win
  - the number of times you lose
  - the number of losses until your third win (If you don’t win three times in 100 tries, you keep playing.)

- (d) the number of wins until your thirtieth loss (If you don't lose thirty times in 100 tries, you keep playing.)
- (e) the amount of time to play your first 40 bets
- (f) the additional amount of time to play your next 60 bets
- (g) the total amount of time to play your 100 bets
- (h) your net profit at the end of the evening
- (i) the amount of time until a stranger wearing a red carnation sits down next to you
- (j) the number of times you are accidentally jostled by the person standing behind you
8. A golfer plays the same golf course daily for a period of many years. You may assume that he does not get better or worse, that all holes are equally difficult and that the results on one hole do not influence the results on any other hole. On any one hole, he has probabilities .05, .5, and .45 of being under par, exactly par, and over par, respectively. Write down what distribution best models each of the following random variables. Be as specific as you can. I.e., instead of answering "Poisson distribution" answer "Poi(3)" or "Poi( $\lambda$ ) where  $\lambda$  is unknown." For some parts the correct answer might be "I don't know."
- (a) X, the number of holes over par on 17 September, 2002
- (b) W, the number of holes over par in September, 2002
- (c) Y, the number of rounds over par in September, 2002
- (d) Z, the number of times he is hit by lightning in this decade
- (e) H, the number of holes-in-one this decade
- (f) T, the time, in years, until his next hole-in-one
9. During a PET scan, a source (your brain) emits photons which are counted by a detector (the machine). The detector is mounted at the end of a long tube, so only photons that head straight down the tube are detected. In other words, though the source emits photons in all directions, the only ones detected are those that are emitted within the small range of angles that lead down the tube to the detector.
- Let X be the number of photons emitted by the source in 5 seconds. Suppose the detector captures only 1% of the photons emitted by the source. Let Y be the number of photons captured by the detector in those same 5 seconds.

- (a) What is a good model for the distribution of  $X$ ?
- (b) What is the conditional distribution of  $Y$  given  $X$ ?
- (c) What is the marginal distribution of  $Y$ ?

Try to answer these questions from first principles, without doing any calculations.

10. (a) Prove Theorem 5.12 using moment generating functions.  
(b) Prove Theorem 5.13 using moment generating functions.
11. (a) Prove Theorem 5.15 by finding  $\mathbb{E}[Y^2]$  using the trick that was used to prove Theorem 5.13.  
(b) Prove Theorem 5.15 by finding  $\mathbb{E}[Y^2]$  using moment generating functions.
12. Case Study 4.2.3 in Larsen and Marx [add reference] claims that the number of fumbles per team in a football game is well modelled by a  $\text{Poisson}(2.55)$  distribution. For this quiz, assume that claim is correct.
  - (a) What is the expected number of fumbles per team in a football game?
  - (b) What is the expected total number of fumbles by both teams?
  - (c) What is a good model for the total number of fumbles by both teams?
  - (d) In a game played in 2002, Duke fumbled 3 times and Navy fumbled 4 times. Write a formula (Don't evaluate it.) for the probability that Duke will fumble exactly 3 times in next week's game.
  - (e) Write a formula (Don't evaluate it.) for the probability that Duke will fumble exactly three times given that they fumble at least once.
13. Clemson University, trying to maintain its superiority over Duke in ACC football, recently added a new practice field by reclaiming a few acres of swampland surrounding the campus. However, the coaches and players refused to practice there in the evenings because of the overwhelming number of mosquitos.  
To solve the problem the Athletic Department installed 10 bug zappers around the field. Each bug zapper, each hour, zaps a random number of mosquitos that has a  $\text{Poisson}(25)$  distribution.
  - (a) What is the exact distribution of the number of mosquitos zapped by 10 zappers in an hour? What are its expected value and variance?
  - (b) What is a good approximation to the distribution of the number of mosquitos zapped by 10 zappers during the course of a 4 hour practice?

- (c) Starting from your answer to the previous part, find a random variable relevant to this problem that has approximately a  $N(0,1)$  distribution.
14. Bob is a high school senior applying to Duke and wants something that will make his application stand out from all the others. He figures his best chance to impress the admissions office is to enter the Guinness Book of World Records for the longest amount of time spent continuously brushing one's teeth with an electric toothbrush. (Time out for changing batteries is permissible.) Batteries for Bob's toothbrush last an average of 100 minutes each, with a variance of 100. To prepare for his assault on the world record, Bob lays in a supply of 100 batteries.
- The television cameras arrive along with representatives of the Guinness company and the American Dental Association and Bob begins the quest that he hopes will be the defining moment of his young life. Unfortunately for Bob his quest ends in humiliation as his batteries run out before he can reach the record which currently stands at 10,200 minutes.
- Justice is well served however because, although Bob did take AP Statistics in high school, he was not a very good student. Had he been a good statistics student he would have calculated in advance the chance that his batteries would run out in less than 10,200 minutes.
- Calculate, approximately, that chance for Bob.
15. An article on statistical fraud detection (BOLTON AND HAND [1992]), when talking about records in a database, says:
- "One of the difficulties with fraud detection is that typically there are many legitimate records for each fraudulent one. A detection method which correctly identifies 99% of the legitimate records as legitimate and 99% of the fraudulent records as fraudulent might be regarded as a highly effective system. However, if only 1 in 1000 records is fraudulent, then, on average, in every 100 that the system flags as fraudulent, only about 9 will in fact be so."
- QUESTION: Can you justify the "about 9"?
16. In 1988 men averaged around 500 on the math SAT, the SD was around 100 and the histogram followed the normal curve.
- Estimate the percentage of men getting over 600 on this test in 1988.
  - One of the men who took the test in 1988 will be picked at random, and you have to guess his test score. You will be given a dollar if you guess it right to within 50 points.

- i. What should you guess?
- ii. What is your chance of winning?

This question was inspired by FREEDMAN ET AL. [1998].

17. Multiple choice.

- (a)  $X \sim \text{Poi}(\lambda)$ .  $\Pr[X \leq 7] =$ 
  - i.  $\sum_{x=-\infty}^7 e^{-\lambda} \lambda^x / x!$
  - ii.  $\sum_{x=0}^7 e^{-\lambda} \lambda^x / x!$
  - iii.  $\sum_{\lambda=0}^7 e^{-\lambda} \lambda^x / x!$
- (b)  $X$  and  $Y$  are distributed uniformly on the unit square.  
 $\Pr[X \leq .5 | Y \leq .25] =$ 
  - i. .5
  - ii. .25
  - iii. can't tell from the information given.
- (c)  $X \sim \text{Normal}(\mu, \sigma)$ .  $\Pr[X > \mu + \sigma]$ 
  - i. is more than .5
  - ii. is less than .5
  - iii. can't tell from the information given.
- (d)  $X_1, \dots, X_{100} \sim N(0, 1)$ .  $\bar{X} \equiv (X_1 + \dots + X_{100})/100$ .  $Y \equiv (X_1 + \dots + X_{100})$ . Calculate
  - i.  $\Pr[-.2 \leq \bar{X} \leq .2]$
  - ii.  $\Pr[-.2 \leq X_i \leq .2]$
  - iii.  $\Pr[-.2 \leq Y \leq .2]$
  - iv.  $\Pr[-2 \leq \bar{X} \leq 2]$
  - v.  $\Pr[-2 \leq X_i \leq 2]$
  - vi.  $\Pr[-2 \leq Y \leq 2]$
  - vii.  $\Pr[-20 \leq \bar{X} \leq 20]$
  - viii.  $\Pr[-20 \leq X_i \leq 20]$
  - ix.  $\Pr[-20 \leq Y \leq 20]$
- (e)  $X \sim \text{Bin}(100, \theta)$ .  $\sum_{\theta=0}^{100} f(x|\theta) =$ 
  - i. 1
  - ii. the question doesn't make sense

- iii. can't tell from the information given.
- (f)  $X$  and  $Y$  have joint density  $f(x, y)$  on the unit square.  $f(x) =$
- $\int_0^1 f(x, y) dx$
  - $\int_0^1 f(x, y) dy$
  - $\int_0^x f(x, y) dy$
- (g)  $X_1, \dots, X_n \sim \text{Gamma}(r, 1/\lambda)$  and are mutually independent.  
 $f(x_1, \dots, x_n) =$
- $[\lambda^r / (r-1)!] (\prod x_i)^{r-1} e^{-\lambda \sum x_i}$
  - $[\lambda^{nr} / ((r-1)!)^n] (\prod x_i)^{r-1} e^{-\lambda \prod x_i}$
  - $[\lambda^{nr} / ((r-1)!)^n] (\prod x_i)^{r-1} e^{-\lambda \sum x_i}$
18. In Figure 5.2, the plots look increasingly Normal as we go down each column. Why?  
Hint: a well-known theorem is involved.
19. Prove Theorem 5.7.
20. Prove a version of Equation 5.1 on page 286. Let  $k = 2$ . Start from the joint pmf of  $Y_1$  and  $Y_2$  (Use Theorem 5.7.), derive the marginal pmf of  $Y_1$ , and identify it.
21. Ecologists who study forests have a concept of *seed rain*. The seed rain in an area is the number of seeds that fall on that area. Seed rain is a useful concept in studying how forests rejuvenate and grow. After falling to earth, some seeds germinate and become seedlings; others do not. For a particular square-meter quadrat, let  $Y_1$  be the number of seeds that fall to earth and germinate and  $Y_2$  be the number of seeds that fall to earth but do not germinate. Let  $Y = Y_1 + Y_2$ . Adopt the statistical model  $Y_1 \sim \text{Poi}(\lambda_1)$ ;  $Y_2 \sim \text{Poi}(\lambda_2)$  and  $Y_1 \perp Y_2$ . Theorem 5.10 says  $Y \sim \text{Poi}(\lambda)$  where  $\lambda = \lambda_1 + \lambda_2$ . Find the distribution of  $Y_1$  given  $Y$ . I.e., find  $P[Y_1 = y_1 | Y = y]$ .
22. Prove Theorem 5.11. Hint: use moment generating functions and the fact that  $\lim_{n \rightarrow \infty} (1 + 1/n)^n = e$ .
23. (a) Let  $Y \sim \text{U}(1, n)$  where the parameter  $n$  is an unknown positive integer. Suppose we observe  $Y = 6$ . Find the m.l.e.  $\hat{n}$ . Hint: *Equation 5.2 defines the pmf for  $y \in \{1, 2, \dots, n\}$ . What is  $p(y)$  when  $y \notin \{1, 2, \dots, n\}$ ?*
- (b) In World War II, when German tanks came from the factory they had serial numbers labelled consecutively from 1. I.e., the numbers were 1, 2, .... The Allies wanted to estimate  $T$ , the total number of German tanks and had, as data,

the serial numbers of the tanks they had captured. Assuming that tanks were captured independently of each other and that all tanks were equally likely to be captured find the m.l.e.  $\hat{T}$ .

24. Let  $Y$  be a continuous random variable,  $Y \sim U(a, b)$ .
- Find  $E[Y]$ .
  - Find  $\text{Var}(Y)$ .
  - Find  $M_Y(t)$ .
25. (a) Is there a discrete distribution that is uniform on the positive integers? Why or why not? If there is such a distribution then we might call it  $U(1, \infty)$ .  
(b) Is there a continuous distribution that is uniform on the real line? Why or why not? If there is, then we might call it  $U(-\infty, \infty)$ .
26. Ecologists are studying pitcher plants in a bog. The ecologists want to estimate the total number of pitcher plants in the bog and make an inference about  $\lambda$ . They adopt the following sampling plan. First, they randomly choose some sites  $s_1, \dots, s_n$  in the bog. Then they go to each site, find the nearest pitcher plant and record its location. The data are these locations  $L_1, \dots, L_n$ . Some sites may share a nearest plant, so some of the  $L_i$ 's may be referring to the same plant. Let  $D_i$  be the distance from  $s_i$  to  $L_i$  and  $D_{i,j}$  be the distance from  $s_i$  to  $L_j$ .
- Let  $D_1$  be the distance from  $s_1$  to  $L_1$ . Find the density  $p(d_1|\lambda)$ . You may assume that pitcher plants arise according to a homogeneous Poisson process with rate  $\lambda$ . Hint: use the relationship between a Poisson process and the exponential distribution.
  - When the ecologists go to  $s_2$  they discover that the nearest plant is the same plant they already found nearest to  $s_1$ . In other words, they discover that  $D_2 = D_{2,1}$ . Find  $\Pr[D_2 = D_{2,1}|\lambda]$ .
  - Find the likelihood function  $\ell(\lambda) \equiv p[L_1, \dots, L_n|\lambda]$ .
  - Find the m.l.e.  $\hat{\lambda} \equiv \text{argmax}_\lambda \ell(\lambda)$ .
  - Suppose the prior distribution for  $\lambda$  is  $\text{Gam}(\alpha, \beta)$ . Find the posterior distribution for  $\lambda$ .
27. Let  $x \sim \text{Gam}(\alpha, \beta)$  and let  $y = 1/x$ . Find the pdf of  $y$ . We say that  $y$  has an *inverse Gamma* distribution with parameters  $\alpha$  and  $\beta$  and write  $y \sim \text{invGam}(\alpha, \beta)$ .

28. Prove Theorem 5.18. **Hint:** Use the method of Theorem 5.13.
29. Let  $x_1, \dots, x_n \sim$  i.i.d.  $U(0, 1)$ . Find the distribution of  $x_{(n)}$ , the largest order statistic.
30. In the R code to create Figure 2.19, explain how to use `dgamma(...)` instead of `dpois(...)`.
31. Prove the claim on page 305 that the half-life of a radioactive isotope is  $m = \lambda \log 2$ .
32. Prove Theorem 5.16.
33. Prove Theorem 5.19.
34. Page 316 shows that the  $n$ -dimensional Normal density with a diagonal covariance matrix is the product of  $n$  separate univariate Normal densities. In this problem you are to work in the opposite direction. Let  $X_1, \dots, X_n$  be independent Normally distributed random variables with means  $\mu_1, \dots, \mu_n$  and SD's  $\sigma_1, \dots, \sigma_n$ .
  - Write down the density of each  $X_i$ .
  - Write down the joint density of  $X_1, \dots, X_n$ .
  - Show that the joint density can be written in the form of Equation 5.6.
  - Derive the mean vector and covariance matrix of  $X_1, \dots, X_n$ .
35. Prove Corollary 5.26.
36. Show, for every  $x \in \mathbb{R}$ ,
- $$\lim_{p \rightarrow \infty} p_{t_p}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$
- where  $p_{t_p}(x)$  is the  $t$  density with  $p$  degrees of freedom, evaluated at  $x$ . Hint: use Sterling's formula. (This problem is Exercise 5.18(c) in **Statistical Inference, 2nd ed.** by Casella and Berger.)
37. It turns out that the  $t$  distribution with  $p$  degrees of freedom can be written as a mixture of Normal distributions, a fact that is sometimes useful in statistical calculations. Let  $\tau \sim \text{Gam}(p/2, 2)$  and, conditional on  $\tau$ ,  $y \sim N(0, \sqrt{p/\tau})$ . Show that the marginal distribution of  $y$  is the  $t$  distribution with  $p$  degrees of freedom.
38. In Section 3.2.2 we saw that the m.l.e. of the coefficients in a Normal linear model, under the assumption  $Y \sim N(XB, \sigma^2 \mathbf{I})$  is  $\hat{B} = (X'X)^{-1}X'Y$ . Since  $Y$  is a random variable, and  $\hat{B}$  is a transformation of  $Y$ ,  $\hat{B}$  is also a random variable. Find its distribution.

39. This exercise will be useful in the Bayesian analysis of Normal distributions. It is a generalization of Exercise 28. Let  $\tau \sim \text{Gam}(\alpha, \beta)$  and, conditional on  $\tau$ , let  $\mu$  have a Normal distribution with mean  $m$  and variance  $v/\tau$ . All of  $\alpha, \beta, m$  and  $v$  are known.

- (a) Find the joint density of  $(\mu, \tau)$ .
- (b) Find the marginal distribution of  $\mu$ .

Given  $(\mu, \tau)$ , let  $Y_1, \dots, Y_n$  be a random sample from the Normal distribution with mean  $\mu$  and precision  $\tau$ . Our goal is to find the posterior distribution of  $(\mu, \tau)$ .

- (c) Write an expression that is proportional to  $p(\mu, \tau | y_1, \dots, y_n)$ .
- (d) Show that the conditional posterior  $p(\mu | \tau, y_1, \dots, y_n)$  is Normal. Find its mean and variance.
- (e) Show that the marginal posterior  $p(\tau | y_1, \dots, y_n)$  is Gamma. Find its parameters.
- (f) Show that the marginal posterior  $p(\mu | y_1, \dots, y_n)$  is  $t$ . Find its parameters.

## CHAPTER 6

# BAYESIAN STATISTICS

## 6.1 Multidimensional Bayesian Analysis

This chapter takes up Bayesian statistics. Modern Bayesian statistics relies heavily on computers, computation, programming, and algorithms, so that will be the major focus of this chapter. We cannot give a complete treatment here, but there are several good books that cover these topics in more depth. See, for example, GELMAN ET AL. [2004], LIU [2004], MARIN AND ROBERT [2007], or ROBERT AND CASELLA [1997].

Recall the framework of Bayesian inference from Section 2.5.

- We posit a parametric family of distributions  $\{p(y|\theta)\}$ .
- We express our old knowledge of  $\theta$  through a prior probability density  $p(\theta)$ .
- The previous two items combine to yield  $p(y, \theta)$  and, ultimately,  $p(\theta|y)$ .
- The posterior density  $p(\theta|y)$  represents our new state of knowledge about  $\theta$ .

The posterior density is

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta) d\theta} \propto p(\theta)p(y|\theta). \quad (6.1)$$

So far, so good. But in many interesting applications,  $\theta$  is multi-dimensional and problems arise when we want to examine the posterior. Equation 6.1 tells us how to evaluate the posterior at any value of  $\theta$ , but that's not always sufficient for getting a sense of which values of  $\theta$  are most likely, somewhat likely, unlikely, etc. One way to develop a feeling for a multidimensional posterior is to examine a marginal posterior density, say

$$p(\theta_1|y) = \int \cdots \int p(\theta_1, \dots, \theta_k|y) d\theta_2 \dots d\theta_k. \quad (6.2)$$

Unfortunately, the integral in Equation 6.2 is often not analytically tractable and must be integrated numerically. Standard numerical integration techniques such as quadrature may work well in low dimensions, but in Bayesian statistics Equation 6.2 is often sufficiently high dimensional that standard techniques are unreliable. Therefore, new numerical integration techniques are needed. The most important of these is called Markov chain Monte Carlo integration, or MCMC. Other techniques can be found in the references at the beginning of the chapter. For the purposes of this book, we investigate MCMC. But first, to get a feel for Bayesian analysis, we explore posteriors in low dimensional, numerically tractable examples.

The general situation is that there are multiple parameters  $\theta_1, \dots, \theta_k$ , and data  $y_1, \dots, y_n$ . We may be interested in marginal, conditional, or joint distributions of the parameters either *a priori* or *a posteriori*. Some examples:

- $p(\theta_1, \dots, \theta_k)$ , the joint prior
- $p(\theta_1, \dots, \theta_k | y_1, \dots, y_n)$ , the joint posterior
- $p(\theta_1 | y_1, \dots, y_n) = \int \cdots \int p(\theta_1, \dots, \theta_k | y_1, \dots, y_n) d\theta_2 \cdots d\theta_k$ , the marginal posterior of  $\theta_1$
- 

$$\begin{aligned} p(\theta_2, \dots, \theta_k | \theta_1, y_1, \dots, y_n) &= p(\theta_1, \dots, \theta_k | y_1, \dots, y_n) / p(\theta_1 | y_1, \dots, y_n) \\ &\propto p(\theta_1, \dots, \theta_k | y_1, \dots, y_n), \end{aligned}$$

the conditional joint posterior density of  $(\theta_2, \dots, \theta_k)$  given  $\theta_1$ , where the “ $\propto$ ” means that we substitute  $\theta_1$  into the numerator and treat the denominator as a constant.

The examples in this section illustrate the ideas.

### **Example 6.1** (Ice Cream Consumption, cont.)

This example continues Example 3.5 in which weekly ice cream consumption is modelled as a function of temperature and possibly other variables. We begin with the model in Equation 3.10:

$$\text{consumption} = \beta_0 + \beta_1 \text{temperature} + \text{error}$$

or

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where  $\epsilon_1, \dots, \epsilon_{30} \sim \text{i.i.d.} N(0, \sigma)$ . For now, let us suppose that  $\sigma$  is known. Later we'll drop that assumption. Because  $\sigma$  is known, there are only two parameters:  $\beta_0$  and  $\beta_1$ .

For a Bayesian analysis we need a prior distribution for them; then we can compute the posterior distribution. For now we adopt the following prior without comment. Later we will see why we chose this prior and examine its consequences.

$$\beta_0 \sim N(\mu_0, \sigma_0)$$

$$\beta_1 \sim N(\mu_1, \sigma_1)$$

$$\beta_0 \perp \beta_1$$

i. e.

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}\right)$$

for some choice of  $(\mu_0, \mu_1, \sigma_0, \sigma_1)$ . The likelihood function is

$$\begin{aligned} \ell(\beta_0, \beta_1) &= p(y_1, \dots, y_{30} | \beta_0, \beta_1) \\ &= \prod_{i=1}^{30} p(y_i | \beta_0, \beta_1) \\ &= \prod_{i=1}^{30} \left( \frac{1}{\sqrt{2\pi}\sigma} \right) e^{-\frac{1}{2} \left( \frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma} \right)^2} \end{aligned}$$

To find the posterior density we will use matrix notation. Let  $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ ,  $\mu = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}$ ,  $\Sigma = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}$ ,  $\vec{Y} = (Y_1, \dots, Y_n)^t$  and

$$X = \begin{pmatrix} 1 & \text{temperature}_1 \\ 1 & \text{temperature}_2 \\ \vdots & \vdots \\ 1 & \text{temperature}_{30} \end{pmatrix}.$$

Conditional on  $\beta$ ,  $\vec{Y}$  has a 30-dimensional Normal distribution with mean  $X\beta$  and covariance matrix  $\sigma^2 I_{30}$ . The posterior density is proportional to the prior times the likelihood.

$$p(\beta | \vec{Y}) \propto e^{-\frac{1}{2}(\beta - \mu)^t \Sigma^{-1}(\beta - \mu) - \frac{1}{2}(\vec{Y} - X\beta)^t (\vec{Y} - X\beta) / \sigma^2} \quad (6.3)$$

At this point we observe that the exponent is a quadratic form in  $\beta$ . Therefore the posterior density will be a two-dimensional Normal distribution for  $\beta$  and we just have

to complete the square to find the mean vector and covariance matrix. The exponent is, apart from a factor of  $-\frac{1}{2}$  and some irrelevant constants involving the  $y_i$ 's,

$$\beta'[\Sigma^{-1} + X'X/\sigma^2]\beta - 2\beta'[\Sigma^{-1}\mu + X'\vec{Y}/\sigma^2] + \dots = (\beta - \mu^*)^t(\Sigma^*)^{-1}(\beta - \mu^*) + \dots$$

where  $\Sigma^* = (\Sigma^{-1} + X'X/\sigma^2)^{-1}$  and  $\mu^* = \Sigma^*(\Sigma^{-1}\mu + X'\vec{Y}/\sigma^2)$ . Therefore, the posterior distribution of  $\beta$  given  $\vec{Y}$  is Normal with mean  $\mu^*$  and covariance matrix  $\Sigma^*$ . It is worth noting (1) that the posterior precision matrix  $(\Sigma^*)^{-1}$  is the sum of the prior precision matrix  $\Sigma^{-1}$  and a part that comes from the data,  $X'X/\sigma^2$  and (2) that the posterior mean is  $\Sigma^*(\Sigma^{-1}\mu + X'\vec{Y}/\sigma^2) = \Sigma^*(\Sigma^{-1}\mu + (X'X/\sigma^2)(X'X)^{-1}X'\vec{Y})$ , a matrix weighted average of the prior mean  $\mu$  and the least-squares estimate  $(X'X)^{-1}X'\vec{Y}$  where the weights are the two precisions  $\Sigma^{-1}$  and  $X'X/\sigma^2$ .

The derivation of the posterior distribution does not depend on any particular choice of  $(\mu, \Sigma)$ , but it does depend on the fact that the prior distribution was Normal because that's what gives us the quadratic form in the exponent. That's one reason we took the prior distribution for  $\beta$  to be Normal: it made the calculations easy.

Now let's look at the posterior more closely, see what it implies for  $(\beta_0, \beta_1)$ , and see how sensitive the conclusions are to the choice of  $(\mu, \Sigma)$ . We're also treating  $\sigma$  as known, so we'll need a value. Let's start with the choice

$$\mu = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \quad \Sigma = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 10^6 & 0 \\ 0 & 10^6 \end{pmatrix}; \quad \text{and} \quad \sigma = 0.05. \quad (6.4)$$

The large diagonal entries in  $\Sigma$  say that we have very little *a priori* knowledge of  $\beta$ . We can use R to calculate the posterior mean and covariance.

```
ic <- read.table ("data/ice_cream.txt", header=T)

mu <- c (0, 0)
Sig <- diag (rep (10^6, 2))
sig <- 0.05
x <- cbind (1, ic$temp)
y <- ic$IC

Sigstar <- solve (solve(Sig) + t(x) %*% x / (sig^2))
mustar <- Sigstar %*% (solve(Sig) %*% mu + t(x) %*% y / (sig^2))
```

The result is

$$\mu^* = \begin{pmatrix} .207 \\ .003 \end{pmatrix} \quad \text{and} \quad \Sigma^* = \begin{pmatrix} 8.54 \times 10^{-4} & -1.570 \times 10^{-5} \\ -1.570 \times 10^{-5} & 3.20 \times 10^{-7} \end{pmatrix}. \quad (6.5)$$

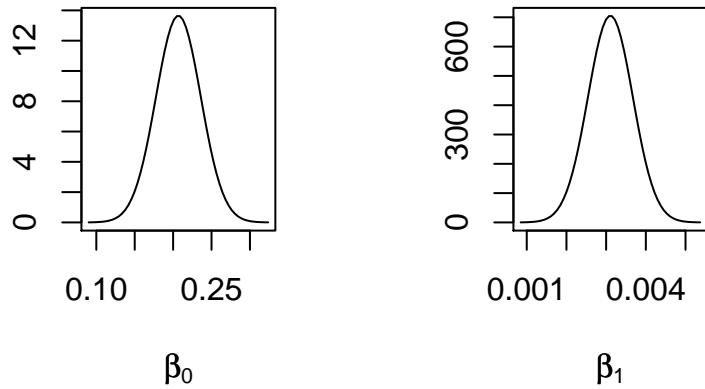


Figure 6.1: Posterior densities of  $\beta_0$  and  $\beta_1$  in the ice cream example using the prior from Equation 6.4.

Note:

- `solve` performs matrix inversions and solves systems of linear equations.
- `t(x)` is the transpose of `x`.
- `%*%` performs matrix multiplication.
- The matrix product `t(x) %*% y` is so common it has a special shortcut, `crossprod( x, y )`. We could have used `crossprod` to find  $\Sigma^*$  and  $\mu^*$ .

Figure 6.1 shows the posterior densities. The posterior density of  $\beta_0$  is not very meaningful because it pertains to ice cream consumption when the temperature is 0. Since our data was collected at temperatures between about 25 and 75, extrapolating to temperatures around 0 would be dangerous. And because  $\beta_0$  is not meaningful, neither is the joint density of  $(\beta_0, \beta_1)$ . On the other hand, our inference for  $\beta_1$  is meaningful. It says that ice cream consumption goes up about .003 ( $\pm .001$  or so) pints per person for every degree increase in temperature. You can verify whether that's sensible by looking at Figure 3.8.

Figure 6.1 was produced by the following snippet.

```
opar <- par (mfrow=c(1,2))
m <- mustar[1]
sd <- sqrt (Sigstar[1,1])
x <- seq (m-4*sd, m+4*sd, length=60)
plot (x, dnorm (x, m, sd), type="l",
 xlab=expression(beta[0]), ylab="")
m <- mustar[2]
sd <- sqrt (Sigstar[2,2])
x <- seq (m-4*sd, m+4*sd, length=60)
plot (x, dnorm (x, m, sd), type="l",
 xlab=expression(beta[1]), ylab="")
```

Now we'd like to investigate the role of the prior density. We notice that the prior SD of  $\beta_1$  was  $10^3$  while the posterior SD is  $\sqrt{3.2 \times 10^{-7}} \approx 5.7 \times 10^{-4}$ . In other words, the data has reduced the uncertainty by a huge amount. Because there's so much information in the data, we expect the prior to have little influence. We can verify that by considering priors with different SD's and comparing their posteriors. To that end, consider the prior

$$\mu = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \quad \Sigma = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \quad \text{and} \quad \sigma = 0.05. \quad (6.6)$$

With prior 6.6, the posterior would be

$$\mu^* = \begin{pmatrix} .207 \\ .003 \end{pmatrix} \quad \text{and} \quad \Sigma^* = \begin{pmatrix} 8.53 \times 10^{-4} & -1.568 \times 10^{-5} \\ -1.568 \times 10^{-5} & 3.19 \times 10^{-7} \end{pmatrix},$$

nearly identical to posterior 6.5. That's because the data contain much more information than the prior, so the prior plays a negligible role in determining the posterior distribution. The posterior precision matrix (inverse covariance matrix) is  $\Sigma^{-1} + X'X/\sigma^2$ . In our case,  $X'X/\sigma^2 = \begin{pmatrix} 30 & 1473 \\ 1473 & 80145 \end{pmatrix}$ , which has entries much bigger than those in  $\Sigma^{-1}$ , whichever prior we use. Even if we did a careful job of assessing our prior, it would be influential only if our prior precision matrix had entries of the same order of magnitude as  $X'X/\sigma^2$ . Since that's unlikely — our true *a priori* variances are probably not as small as  $1/30$  — there's little to be gained by choosing the prior carefully and much effort to be saved by using an arbitrary prior, as long as it has reasonably large variances.

| ring | ID    | xcoor | ycoor | spec | dbh  | 1998 | 1999 | 2000 |
|------|-------|-------|-------|------|------|------|------|------|
| 1    | 11003 | 0.71  | 0.53  | pita | 19.4 | 0    | 0    | 0    |
| 1    | 11004 | 1.26  | 2.36  | pita | 14.1 | 0    | 0    | 4    |
| 1    | 11011 | 1.44  | 6.16  | pita | 19.4 | 0    | 6    | 0    |
| 1    | 11013 | 3.56  | 5.84  | pita | 21.6 | 0    | 0    | 0    |
| 1    | 11017 | 3.75  | 8.08  | pita | 10.8 | 0    | 0    | 0    |
| ⋮    |       |       |       |      |      |      |      |      |
| 6    | 68053 | 0.82  | 10.73 | pita | 14.4 | 0    | 0    | 0    |
| 6    | 68055 | -2.24 | 13.34 | pita | 11   | 0    | 0    | 0    |
| 6    | 68057 | -0.78 | 14.21 | pita | 8    | 0    | 0    | 0    |
| 6    | 68058 | 0.76  | 14.55 | pita | 10.6 | 0    | 0    | 0    |
| 6    | 68059 | 1.48  | 13    | pita | 21.2 | 0    | 5    | 10   |

Table 6.1: The numbers of pine cones on trees in the FACE experiment, 1998–2000.

**Example 6.2** (Pine Cones)

One possible result of increased CO<sub>2</sub> in the atmosphere is that plants will use some of the excess carbon for reproduction, instead of growth. They may, for example produce more seeds, produce bigger seeds, produce seeds earlier in life, or produce seeds when they, the plants, are smaller. To investigate this possibility in the Duke FACE experiment (See Example 1.12 and its sequels.) a graduate student went to the FACE site each year and counted the number of pine cones on pine trees in the control and treatment plots (LADEAU AND CLARK [2001] and Example 3.8).

The data are in Table 6.1. The first column is *ring*. Rings 1, 5, and 6 were control; 2, 3, and 4 were treatment. The next column, *ID*, identifies each tree uniquely; *xcoor* and *ycoor* give the location of the tree. The next column, *spec*, gives the species; *pita* stands for *pinus taeda*, or loblolly pine, the dominant canopy tree in the FACE experiment. The column *dbh* gives diameter at breast height, a common way for foresters and ecologists to measure the size of a tree. The final three columns show the number of pine cones in 1998, 1999, and 2000. We investigate the relationship between *dbh* and the number of pine cones, and whether that relationship is the same in the control and treatment plots.

Figures 6.2, 6.3, and 6.4 plot the numbers of pine cones as a function of *dbh* in the years 1998–2000. In 1998, very few trees had pine cones and those that did had very few. But by 1999, many more trees had cones and had them in greater number. There does not appear to be a substantial difference between 1999 and 2000. As a quick check of our visual impression we can count the fraction of pine trees having

pine cones each year, by ring. The following R code does the job.

```
for (i in 1:6) {
 good <- cones$ring == i
 print (c (sum (cones$X1998[good] > 0) / sum(good) ,
 sum (cones$X1999[good] > 0) / sum(good) ,
 sum (cones$X2000[good] > 0) / sum(good)))
}
[1] 0.0000000 0.1562500 0.2083333
[1] 0.05633803 0.36619718 0.32394366
[1] 0.01834862 0.21100917 0.27522936
[1] 0.05982906 0.39316239 0.37606838
[1] 0.01923077 0.10576923 0.22115385
[1] 0.04081633 0.19727891 0.18367347
```

Since there's not much action in 1998 we will ignore the data from that year. The data show a greater contrast between treatment (rings 2, 3, 4) and control (rings 1, 5, 6) in 1999 than in 2000. So for the purpose of this example we'll use the data from 1999. A good scientific investigation, though, would use data from all years.

We're looking for a model with two features: (1) the probability of cones is an increasing function of dbh and of the treatment and (2) given that a tree has cones, the number of cones is an increasing function of dbh and treatment. Here we describe a simple model with these features. The idea is (1) a logistic regression with covariates dbh and treatment for the probability that a tree is sexually mature and (2) a Poisson regression with covariates dbh and treatment for the number of cones given that a tree is sexually mature. Let  $Y_i$  be the number of cones on the  $i$ 'th tree. Our model is

$$\begin{aligned} x_i &= \begin{cases} 1 & \text{if the } i\text{'th tree had extra CO}_2 \\ 0 & \text{otherwise} \end{cases} \\ \theta_i &= \begin{cases} 1 & \text{if the } i\text{'th tree is sexually mature} \\ 0 & \text{otherwise} \end{cases} \\ \pi_i &= P[\theta_i = 1] = \frac{\exp(\beta_0 + \beta_1 \text{dbh}_i + \beta_2 x_i)}{1 + \exp(\beta_0 + \beta_1 \text{dbh}_i + \beta_2 x_i)} \\ \phi_i &= \exp(\gamma_0 + \gamma_1 \text{dbh}_i + \gamma_2 x_i) \\ Y_i &\sim \text{Poi}(\theta_i \phi_i) \end{aligned} \tag{6.7}$$

This model is called a *zero-inflated Poisson model*. There are six unknown parameters:  $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2$ . We must assign prior distributions and compute posterior

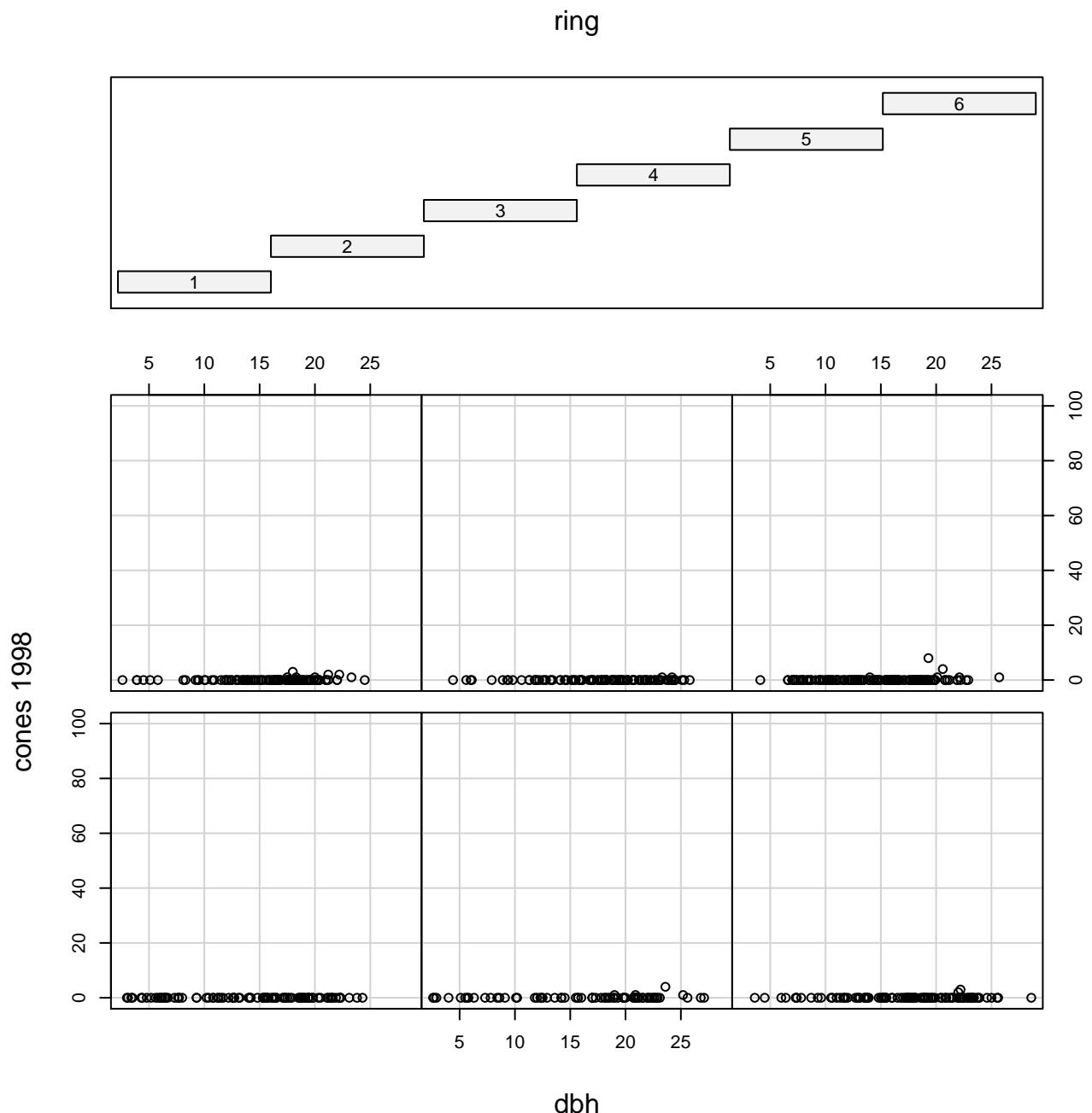


Figure 6.2: Numbers of pine cones in 1998 as a function of dbh

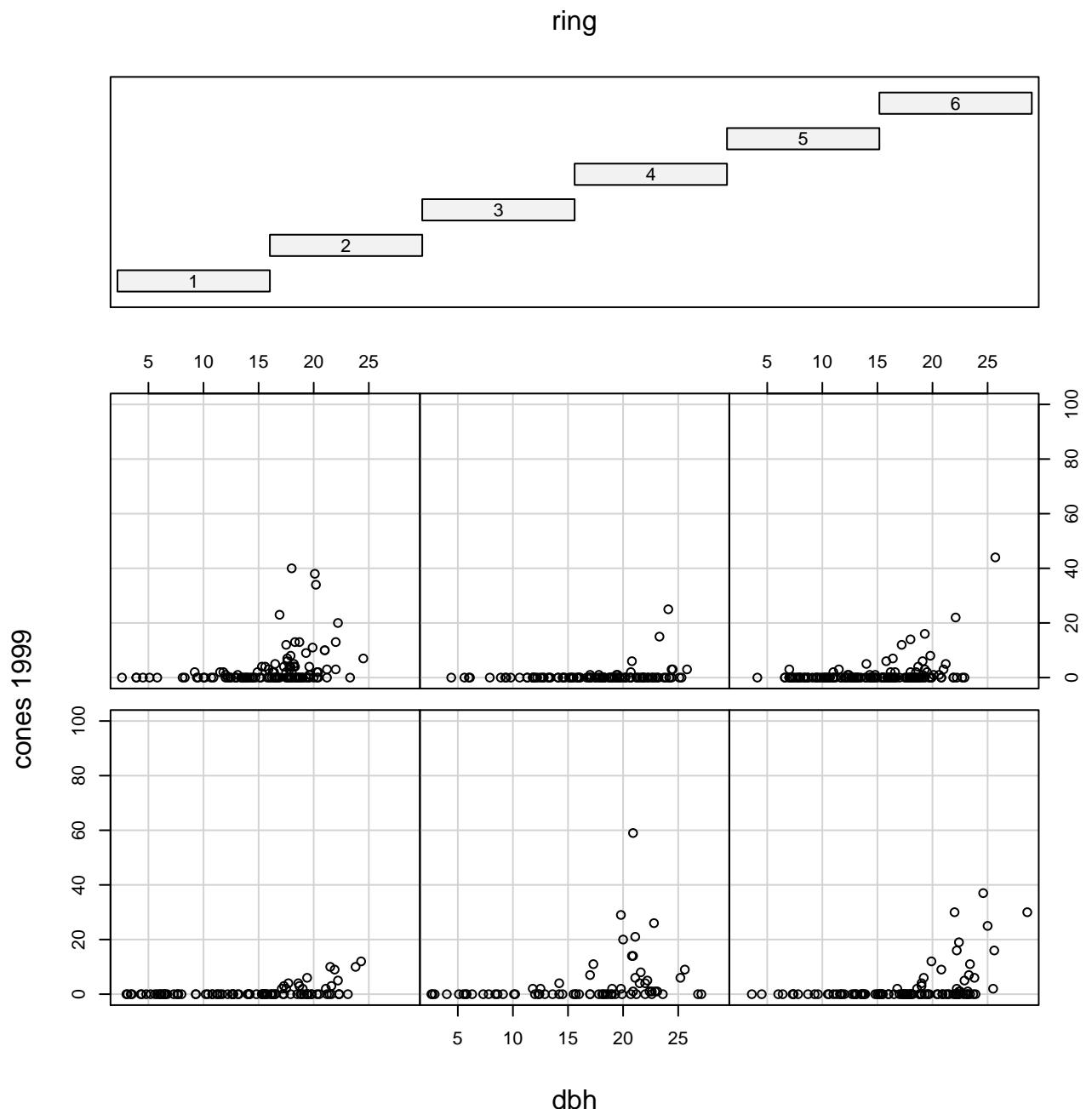


Figure 6.3: Numbers of pine cones in 1999 as a function of dbh

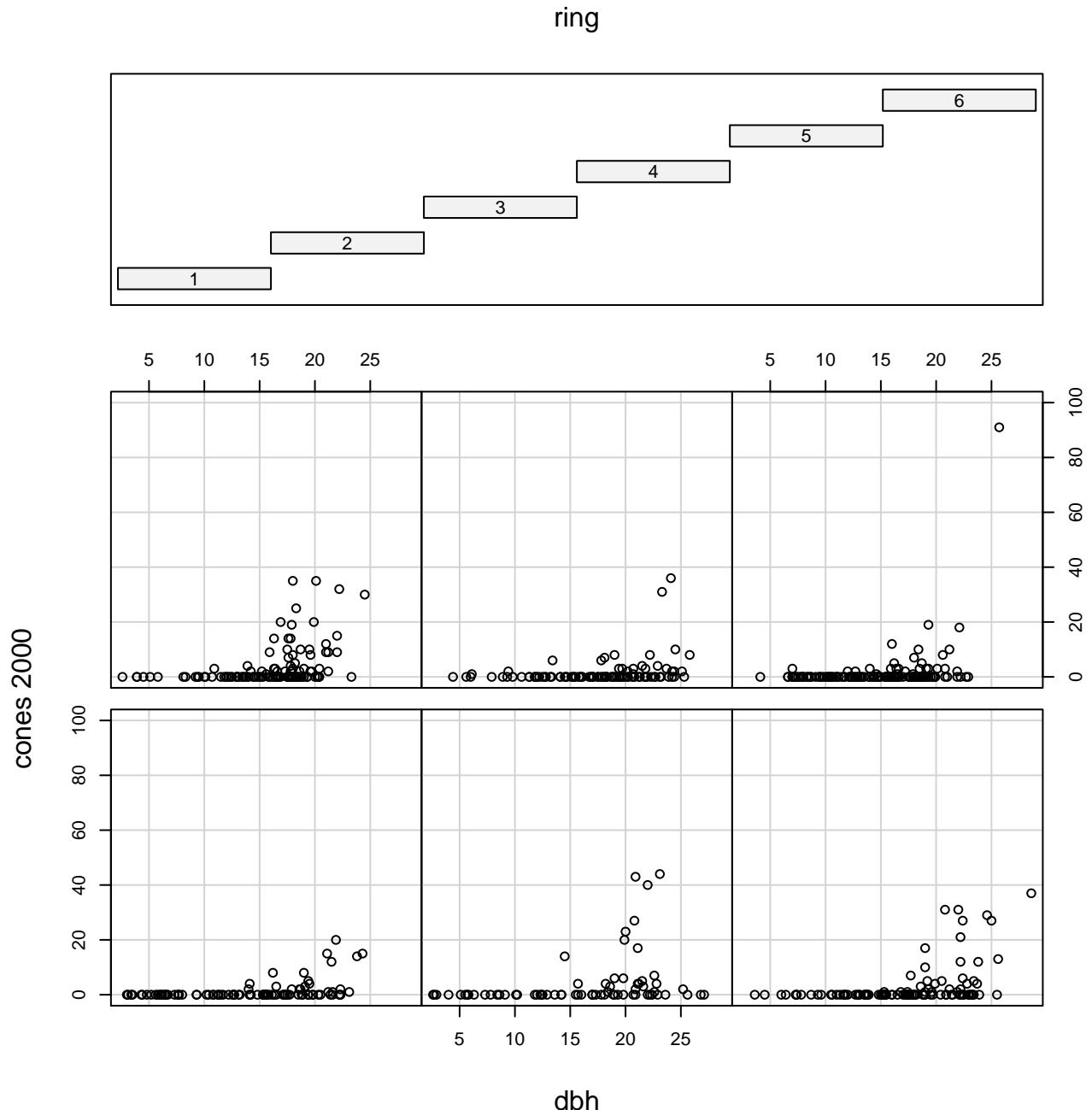


Figure 6.4: Numbers of pine cones in 2000 as a function of dbh

distributions of these parameters. In addition, each tree has an indicator  $\theta_i$  and we will be able to calculate the posterior probabilities  $P[\theta_i = 1 | y_1, \dots, y_n]$  for  $i = 1, \dots, n$ .

We start with the priors  $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2 \sim \text{i.i.d.U}(-100, 100)$ . This prior distribution is not, obviously, based on any substantive prior knowledge. Instead of arguing that this is a sensible prior, we will later check the robustness of conclusions to specification of the prior. If the conclusions are robust, then we will argue that almost any sensible prior would lead to roughly the same conclusions.

To begin the analysis we write down the joint distribution of parameters and data.

$$\begin{aligned}
p(y_1, \dots, y_n, \beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2) &= p(\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2) \times p(y_1, \dots, y_n | \beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2) \\
&= \left( \frac{1}{200} \right)^6 \mathbf{1}_{(-100, 100)}(\beta_0) \times \mathbf{1}_{(-100, 100)}(\beta_1) \times \mathbf{1}_{(-100, 100)}(\beta_2) \times \mathbf{1}_{(-100, 100)}(\gamma_0) \\
&\quad \times \mathbf{1}_{(-100, 100)}(\gamma_1) \times \mathbf{1}_{(-100, 100)}(\gamma_2) \\
&\times \prod_{i:y_i>0} \left( \frac{\exp(\beta_0 + \beta_1 \text{dbh}_i + \beta_2 x_i)}{1 + \exp(\beta_0 + \beta_1 \text{dbh}_i + \beta_2 x_i)} \right. \\
&\quad \times \left. \frac{\exp(-\exp(\gamma_0 + \gamma_1 \text{dbh}_i + \gamma_2 x_i)) \exp(\gamma_0 + \gamma_1 \text{dbh}_i + \gamma_2 x_i)^{y_i}}{y_i!} \right) \\
&\times \prod_{i:y_i=0} \left( \frac{1}{1 + \exp(\beta_0 + \beta_1 \text{dbh}_i + \beta_2 x_i)} + \right. \\
&\quad \left. \frac{\exp(\beta_0 + \beta_1 \text{dbh}_i + \beta_2 x_i)}{1 + \exp(\beta_0 + \beta_1 \text{dbh}_i + \beta_2 x_i)} \exp(-\exp(\gamma_0 + \gamma_1 \text{dbh}_i + \gamma_2 x_i)) \right) \quad (6.8)
\end{aligned}$$

In Equation 6.8 each term in the product  $\prod_{i:y_i>0}$  is

$$P[i\text{'th tree is sexually mature}] \times p(y_i | i\text{'th tree is sexually mature})$$

while each term in  $\prod_{i:y_i=0}$  is

$$P[i\text{'th tree is immature}] + P[i\text{'th tree is mature but produces no cones}].$$

The posterior  $p(\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2 | y_1, \dots, y_n)$  is proportional, as a function of  $(\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2)$ , to Equation 6.8. Similarly, conditional posteriors such as  $p(\beta_0 | \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2, y_1, \dots, y_n)$  are proportional, as a function of  $\beta_0$ , to Equation 6.8. But that doesn't allow for much simplification; it allows us to ignore only the factorials in the denominator.

To learn about the posterior in, say, Equation 6.8 it is easy to write an R function that accepts  $(\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2)$  as input and returns 6.8 as output. But that's quite a complicated function of  $(\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2)$  and it's not obvious how to use the function or what it says about any of the four parameters. Therefore, in Section 6.2 we present an algorithm that is very powerful for evaluating the integrals that often arise in multivariate Bayesian analyses.

## 6.2 The Metropolis, Metropolis-Hastings, and Gibbs Sampling Algorithms

In “Markov chain Monte Carlo”, the term “Monte Carlo” refers to evaluating an integral by using many random draws from a distribution. To fix ideas, suppose we want to evaluate Equation 6.2. Let  $\vec{\theta} = (\theta_1, \dots, \theta_k)$ . If we could generate many samples  $\vec{\theta}_1, \dots, \vec{\theta}_M$  of  $\vec{\theta}$  (where  $\vec{\theta}_i = (\theta_{i,1}, \dots, \theta_{i,k})$ ) from its posterior distribution then we could approximate Equation 6.2 by

1. discarding  $\theta_{i,2}, \dots, \theta_{i,k}$  from each iteration,
2. retaining  $\theta_{1,1}, \dots, \theta_{M,1}$ ,
3. using  $\theta_{1,1}, \dots, \theta_{M,1}$  and standard density estimation techniques (page 103) to estimate  $p(\theta_1 | y)$ , or
4. for any set  $A$ , using

$$\frac{\text{number of } \theta_{i,1}\text{'s in } A}{M}$$

as an estimate of  $P[\theta_1 \in A | y]$ .

That's the idea behind Monte Carlo integration.

The term “Markov chain” refers to how the samples  $\vec{\theta}_1, \dots, \vec{\theta}_M$  are produced. In a Markov chain there is a *transition density* or *transition kernel*  $k(\vec{\theta}_i | \vec{\theta}_{i-1})$  which is a density for generating  $\vec{\theta}_i$  given  $\vec{\theta}_{i-1}$ . We first choose  $\vec{\theta}_1$  almost arbitrarily, then generate  $(\vec{\theta}_2 | \vec{\theta}_1)$ ,  $(\vec{\theta}_3 | \vec{\theta}_2)$ , and so on, in succession, for as many steps as we like. Each  $\vec{\theta}_i$  has a density  $p_i \equiv p(\vec{\theta}_i)$  which depends on  $\vec{\theta}_1$  and the transition kernel. But,

1. under some fairly benign conditions (See the references at the beginning of the chapter for details.) the sequence  $p_1, p_2, \dots$  converges to a limit  $p$ , the *stationary distribution*, that does not depend on  $\vec{\theta}_1$ ;

2. the transition density  $k(\vec{\theta}_i | \vec{\theta}_{i-1})$  can be chosen so that the stationary distribution  $p$  is equal to  $p(\vec{\theta} | y)$ ;
3. we can find an  $m$  such that  $i > m \Rightarrow p_i \approx p = p(\vec{\theta} | y)$ ;
4. then  $\vec{\theta}_{m+1}, \dots, \vec{\theta}_M$  are, approximately, a sample from  $p(\vec{\theta} | y)$ .

The Metropolis-Hastings algorithm [METROPOLIS ET AL., 1953, HASTINGS, 1970] is one way to construct an MCMC algorithm whose stationary distribution is  $p(\vec{\theta} | y)$ . It works according to the following steps.

1. Choose a proposal density  $g(\vec{\theta}^* | \vec{\theta})$ .
2. Choose  $\vec{\theta}_1$ .
3. For  $i = 2, 3, \dots$ 
  - Generate a proposal  $\vec{\theta}^*$  from  $g(\vec{\theta}^* | \vec{\theta}_{i-1})$ .
  - Set
$$r \equiv \min \left\{ 1, \frac{p(\vec{\theta}^* | y)g(\vec{\theta}_{i-1} | \vec{\theta}^*)}{p(\vec{\theta}_{i-1} | y)g(\vec{\theta}^* | \vec{\theta}_{i-1})} \right\}. \quad (6.9)$$
  - Set
$$\vec{\theta}_i = \begin{cases} \vec{\theta}^* & \text{with probability } r, \\ \vec{\theta}_{i-1} & \text{with probability } 1 - r. \end{cases}$$

Step 3 define the transition kernel  $k$ . In many MCMC chains, the acceptance probability  $r$  may be strictly less than one, so the kernel  $k$  is a mixture of two parts: one that generates a new value of  $\vec{\theta}_{i+1} \neq \vec{\theta}_i$  and one that sets  $\vec{\theta}_{i+1} = \vec{\theta}_i$ .

To illustrate MCMC, suppose we want to generate a sample  $\theta_1, \dots, \theta_{10,000}$  from the  $\text{Be}(5, 2)$  distribution. We arbitrarily choose a proposal density  $g(\theta^* | \theta) = \text{U}(\theta - .1, \theta + .1)$  and arbitrarily choose  $\theta_1 = 0.5$ . The following R code draws the sample.

```
samp <- rep (NA, 10000)
samp[1] <- 0.5
for (i in 2:10000) {
 prev <- samp[i-1]
 thetastar <- runif (1, prev - .1, prev + .1)
 r <- min (1, dbeta(thetastar,5,2) / dbeta(prev,5,2))
 if (rbinom (1, 1, r) == 1)
 new <- thetastar
 samp[i] <- new}
```

```

 else
 new <- prev
 samp[i] <- new
 }
}

```

The top panel of Figure 6.5 shows the result. The solid curve is the  $\text{Be}(5, 2)$  density and the histogram is made from the Metropolis-Hastings samples. They match closely, showing that the algorithm performed well.

Figure 6.5 was produced by

```

par (mfrow=c(3,1))
hist (samp[-(1:1000)], prob=TRUE, xlab=expression(theta),
 ylab="", main="")
x <- seq(0,1,length=100)
lines (x, dbeta(x,5,2))
plot (samp, pch=". ", ylab=expression(theta))
plot (dbeta(samp,5,2), pch=". ", ylab=expression(p(theta)))

```

The code `samp[-(1:1000)]` discards the first 1000 draws in the hope that the sampler will have converged to its stationary distribution after 1000 iterations.

Assuming that convergence conditions have been met and that the algorithm is well-constructed, MCMC chains are guaranteed eventually to converge and deliver samples from the desired distribution. But the guarantee is asymptotic and in practice the output from the chain should be checked to diagnose potential problems that might arise in finite samples.

The main thing to check is *mixing*. An MCMC algorithm operates in the space of  $\vec{\theta}$ . At each iteration of the chain, i.e., for each value of  $i$ , there is a current location  $\vec{\theta}_i$ . At the next iteration the chain moves to a new location  $\vec{\theta}_i'$ . In this way the chain explores the  $\vec{\theta}$  space. While it is exploring it also evaluates  $p(\vec{\theta}_i')$ . In theory, the chain should spend many iterations at values of  $\vec{\theta}$  where  $p(\vec{\theta})$  is large — and hence deliver many samples of  $\vec{\theta}$ 's with large posterior density — and few iterations at values where  $p(\vec{\theta})$  is small. For the chain to do its job it must find the mode or modes of  $p(\vec{\theta})$ , it must move around in their vicinity, and it must move between them. The process of moving from one part of the space to another is called *mixing*.

The middle and bottom panels of Figure 6.5 illustrate mixing. The middle panel plots  $\theta_i$  vs.  $i$ . It shows that the chain spends most of its iterations in values of  $\theta$  between about 0.6

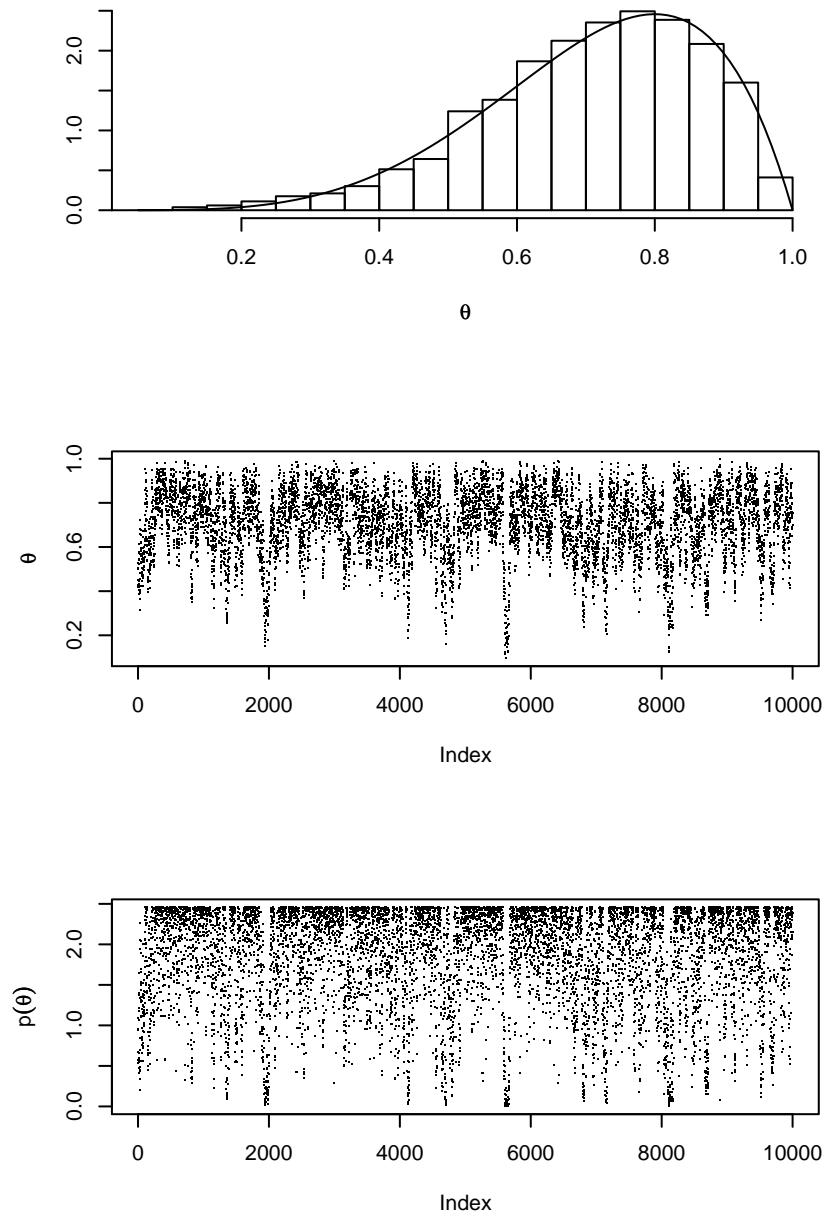


Figure 6.5: 10,000 MCMC samples of the  $\text{Be}(5, 2)$  density. **Top panel:** histogram of samples from the Metropolis-Hastings algorithm and the  $\text{Be}(5, 2)$  density. **Middle panel:**  $\theta_i$  plotted against  $i$ . **Bottom panel:**  $p(\theta_i)$  plotted against  $i$ .

and 0.9 but makes occasional excursions down to 0.4 or 0.2 or so. After each excursion it comes back to the mode around 0.8. The chain has taken many excursions, so it has explored the space well. The bottom panel plots  $p(\theta_i)$  vs.  $i$ . It shows that the chain spent most of its time near the mode where  $p(\theta) \approx 2.4$  but made multiple excursions down to places where  $p(\theta)$  is around 0.5, or even less. This chain mixed well.

To illustrate poor mixing we'll use the same MCMC algorithm but with different proposal kernels. First we'll use  $(\theta^* | \theta) = U(\theta - 100, \theta + 100)$  and change the corresponding line of code to

`thetastar <- runif ( 1, prev - 100, prev + 100 ).` Then we'll use  $(\theta^* | \theta) = U(\theta - .00001, \theta + .00001)$  and change the corresponding line of code to  
`thetastar <- runif ( 1, prev - .00001, prev + .00001 ).` Figure 6.6 shows the result. The left-hand side of the figure is for  $(\theta^* | \theta) = U(\theta - 100, \theta + 100)$ . The top panel shows a very much rougher histogram than Figure 6.5; the middle and bottom panels show why. The proposal radius is so large that most proposals are rejected; therefore,  $\theta_{i+1} = \theta_i$  for many iterations; therefore we get the flat spots in the middle and bottom panels. The plots reveal that the sampler explored fewer than 30 separate values of  $\theta$ . That's too few; the sampler has not mixed well. In contrast, the right-hand side of the figure — for  $(\theta^* | \theta) = U(\theta - .00001, \theta + .00001)$  — shows that  $\theta$  has drifted steadily downward, but over a very small range. There are no flat spots, so the sampler is accepting most proposals, but the proposal radius is so small that the sampler hasn't yet explored most of the space. It too has not mixed well.

Plots such as the middle and bottom plots of Figure 6.6 are called *trace* plots because they trace the path of the sampler.

In this problem, good mixing depends on getting the proposal radius not too large and not too small, but just right (HASSALL [1909]). To be sure, if we run the MCMC chain long enough, all three samplers would yield good samples from  $Be(5, 2)$ . But the first sampler mixed well with only 10,000 iterations while the others would require many more iterations to yield a good sample. In practice, one must examine the output of one's MCMC chain to diagnose mixing problems. No diagnostics are fool proof, but not diagnosing is foolhardy.

Several special cases of the Metropolis-Hastings algorithm deserve separate mention.

**Metropolis algorithm** It is often convenient to choose the proposal density  $g(\vec{\theta}^* | \vec{\theta})$  to be symmetric; i.e., so that  $g(\vec{\theta}^* | \vec{\theta}) = g(\vec{\theta} | \vec{\theta}^*)$ . In this case the Metropolis ratio  $p(\vec{\theta}^* | y)g(\vec{\theta}_{i-1} | \vec{\theta}^*)/p(\vec{\theta}_{i-1} | y)g(\vec{\theta}^* | \vec{\theta}_{i-1})$  simplifies to  $p(\vec{\theta}^* | y)/p(\vec{\theta}_{i-1} | y)$ . That's what happened in the  $Be(5, 2)$  illustration and why the line

`r <- min ( 1, dbeta(thetastar,5,2) / dbeta(prev,5,2) )` doesn't involve  $g$ .

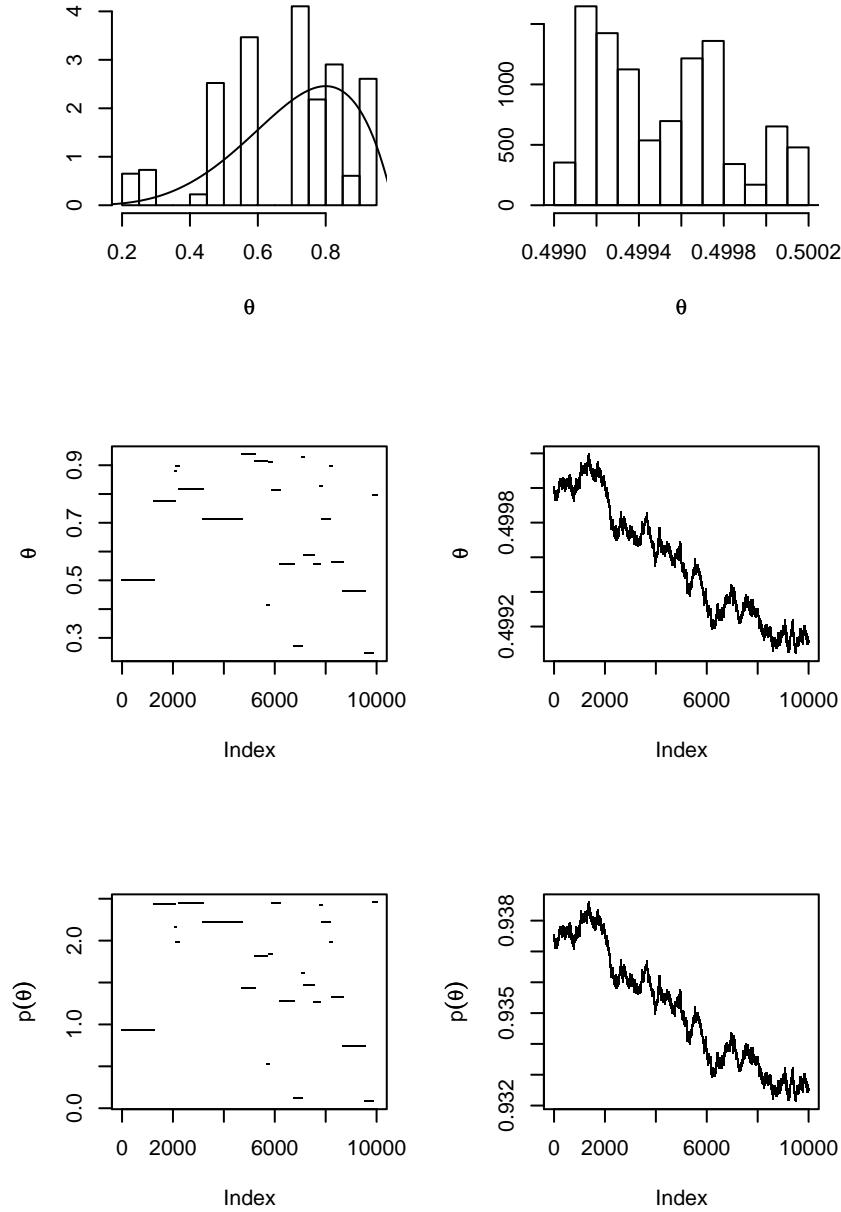


Figure 6.6: 10,000 MCMC samples of the  $\text{Be}(5, 2)$  density. **Left column:**  $(\theta^* | \theta) = U(\theta - 100, \theta + 100)$ ; **Right column:**  $(\theta^* | \theta) = U(\theta - .00001, \theta + .00001)$ . **Top:** histogram of samples from the Metropolis-Hastings algorithm and the  $\text{Be}(5, 2)$  density. **Middle:**  $\theta_i$  plotted against  $i$ . **Bottom:**  $p(\theta_i)$  plotted against  $i$ .

**Independence sampler** It may be convenient to choose  $g(\vec{\theta}^* | \vec{\theta}) = g(\vec{\theta}^*)$  not dependent on  $\vec{\theta}$ . For example, we could have used `thetastar <- runif(1)` in the `Be(5, 2)` illustration.

**Multiple transition kernels** We may construct multiple transition kernels, say  $g_1, \dots, g_m$ . Then for each iteration of the MCMC chain we can randomly choose  $j \in \{1, \dots, m\}$  and make a proposal according to  $g_j$ . We would do this either for convenience or to improve the convergence rate and mixing properties of the chain.

**Gibbs sampler** [GEMAN AND GEMAN, 1984] In many practical examples, the so-called *full conditionals* or *complete conditionals*  $p(\theta_j | \vec{\theta}_{(-j)}, y)$  are known and easy to sample for all  $j$ , where  $\theta_{(-j)} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_k)$ . In this case we may sample  $\theta_{i,j}$  from

$p(\theta_j | \theta_{i,1}, \dots, \theta_{i,j-1}, \theta_{i-1,j+1}, \dots, \theta_{i-1,k})$  for  $j = 1, \dots, k$  and set  $\vec{\theta}_i = (\theta_{i,1}, \dots, \theta_{i,k})$ . We would do this for convenience.

The next example illustrates several MCMC algorithms on the pine cone data of Example 6.2.

### Example 6.3 (Pine Cones, cont)

In this example we try several MCMC algorithms to evaluate and display the posterior distribution in Equation 6.8. Throughout this example, we shall, for compactness, refer to the posterior density as  $p(\vec{\theta})$  instead of  $p(\vec{\theta} | y_1, \dots, y_n)$ .

First we need functions to return the prior density and the likelihood function.

```
dprior <- function (params, log=FALSE) {
 logprior <- (dunif (params["b0"], -100, 100, log=TRUE)
 + dunif (params["b1"], -100, 100, log=TRUE)
 + dunif (params["b2"], -100, 100, log=TRUE)
 + dunif (params["g0"], -100, 100, log=TRUE)
 + dunif (params["g1"], -100, 100, log=TRUE)
 + dunif (params["g2"], -100, 100, log=TRUE)
)
 if (log) return (logprior)
 else return (exp(logprior))
}

lik <- function (params, n.cones=cones$X2000, dbh=cones$dbh,
 trt=cones$trt, log=FALSE) {
 zero <- n.cones == 0
```

```

tmp1 <- params["b0"] + params["b1"] * dbh + params["b2"] * trt
tmp2 <- params["g0"] + params["g1"] * dbh + params["g2"] * trt
etmp1 <- exp(tmp1)
etmp2 <- exp(tmp2)

loglik <- (sum (tmp1[!zero])
 - sum (etmp2[!zero])
 + sum (n.cones[!zero] * tmp2[!zero])
 + sum (log (1 + etmp1[zero] * exp (-etmp2[zero])))
 - sum (log (1 + etmp1))
)
if (log) return (loglik)
else return (exp(loglik))
}

```

Now we write a proposal function. This ones makes  $(\vec{\theta}^* | \vec{\theta}) \sim N(\vec{\theta}, .1\mathbf{I}_6)$ , where  $\mathbf{I}_6$  is the  $6 \times 6$  identity matrix.

```

g.all <- function (params) {
 sig <- c(.1,.1,.1,.1,.1)
 proposed <- mvrnorm (1, mu=params, Sigma=diag(sig))
 return (list (proposed=proposed, ratio=1))
}

```

Finally we write the main part of the code. Try to understand it; you may have to write something similar. Notice an interesting feature of R: assigning names to the components of `params` allows us to refer to the components by name in the `lik` function.

```

initial values
params <- c ("b0"=0, "b1"=0, "b2"=0, "g0"=0, "g1"=0, "g2"=0)

number of iterations
mc <- 10000

storage for output
mcmc.out <- matrix (NA, mc, length(params)+1)

```

```

the main loop
for (i in 1:mc) {
 prop <- g.all (params)
 new <- prop$proposed
 log.accept.ratio <- (dprior (new, log=TRUE)
 - dprior (params, log=TRUE)
 + lik (new, log=TRUE)
 - lik (params, log=TRUE)
 - log (prop$ratio)
)
 accept.ratio <- min (1, exp(log.accept.ratio))

 if (as.logical (rbinom(1,1,accept.ratio)))
 params <- new

 mcmc.out[i,] <- c (params, lik (params, log=TRUE))
}

```

Figure 6.7 shows trace plots of the output. The plots show that the sampler did not move very often; it did not mix well and did not explore the space effectively.

Figure 6.7 was produced by the following snippet.

```

par (mfrow=c(4,2), mar=c(4,4,1,1)+.1)
for (i in 1:6)
 plot (mcmc.out[,i], ylab=names(params)[i], pch=".")
plot (mcmc.out[,7], ylab=expression(p(theta)), pch=".")

```

When samplers get stuck, sometimes it's because the proposal radius is too large. So next we try a smaller radius: `sig <- rep(.01,6)`. Figure 6.8 shows the result. The sampler is still not mixing well. The parameter  $\beta_0$  travelled from its starting point of  $\beta_0 = 0$  to about  $\beta_0 \approx -1.4$  or so, then seemed to get stuck; other parameters behaved similarly. Let's try running the chain for more iterations: `mc <- 100000`. Figure 6.9 shows the result. Again, the sampler does not appear to have mixed well. Parameters  $\beta_0$  and  $\beta_1$ , for example, have not yet settled into any sort of steady-state behavior and  $p(\vec{\theta})$  seems to be steadily increasing, indicating that the sampler may not yet have found the posterior mode.

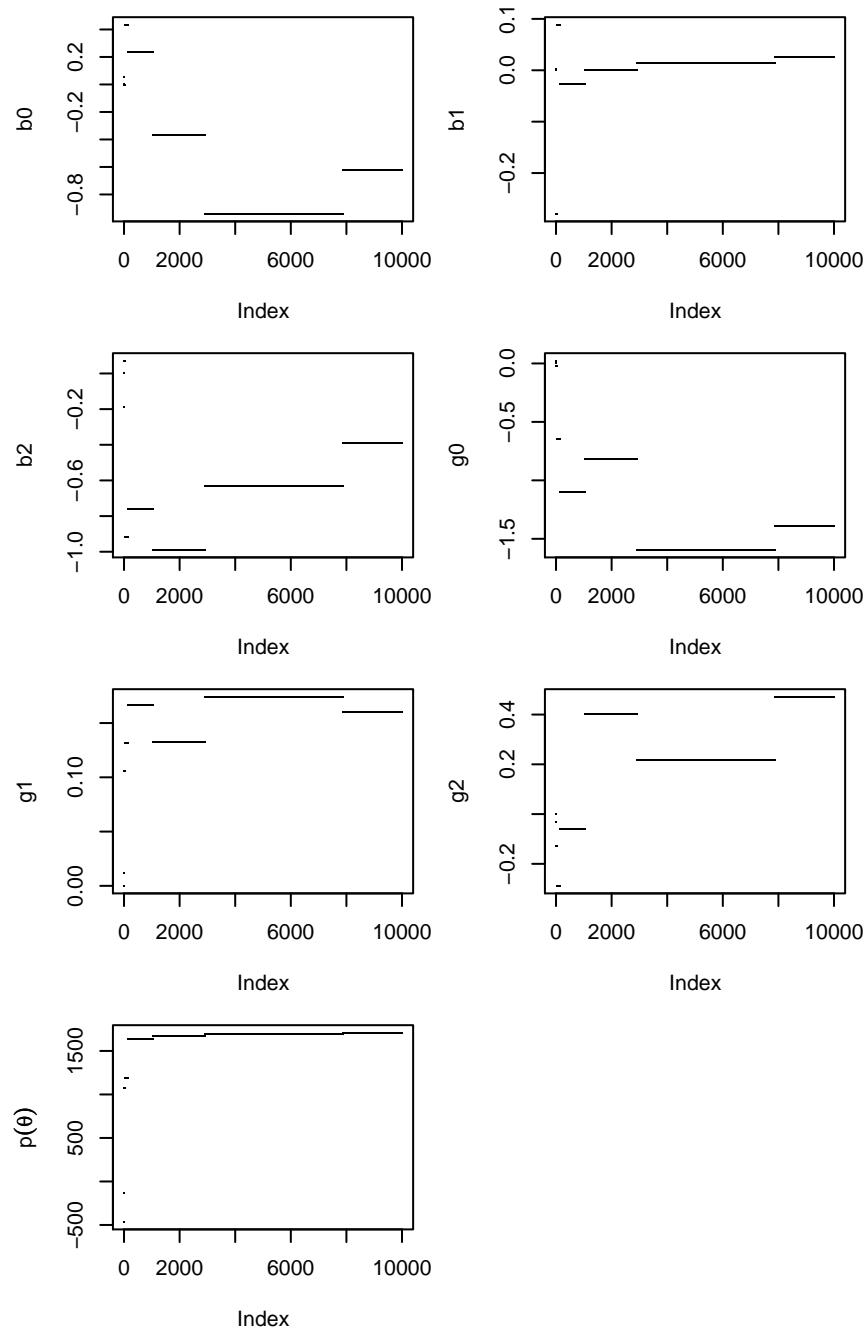


Figure 6.7: Trace plots of MCMC output from the pine cone code on page 360.

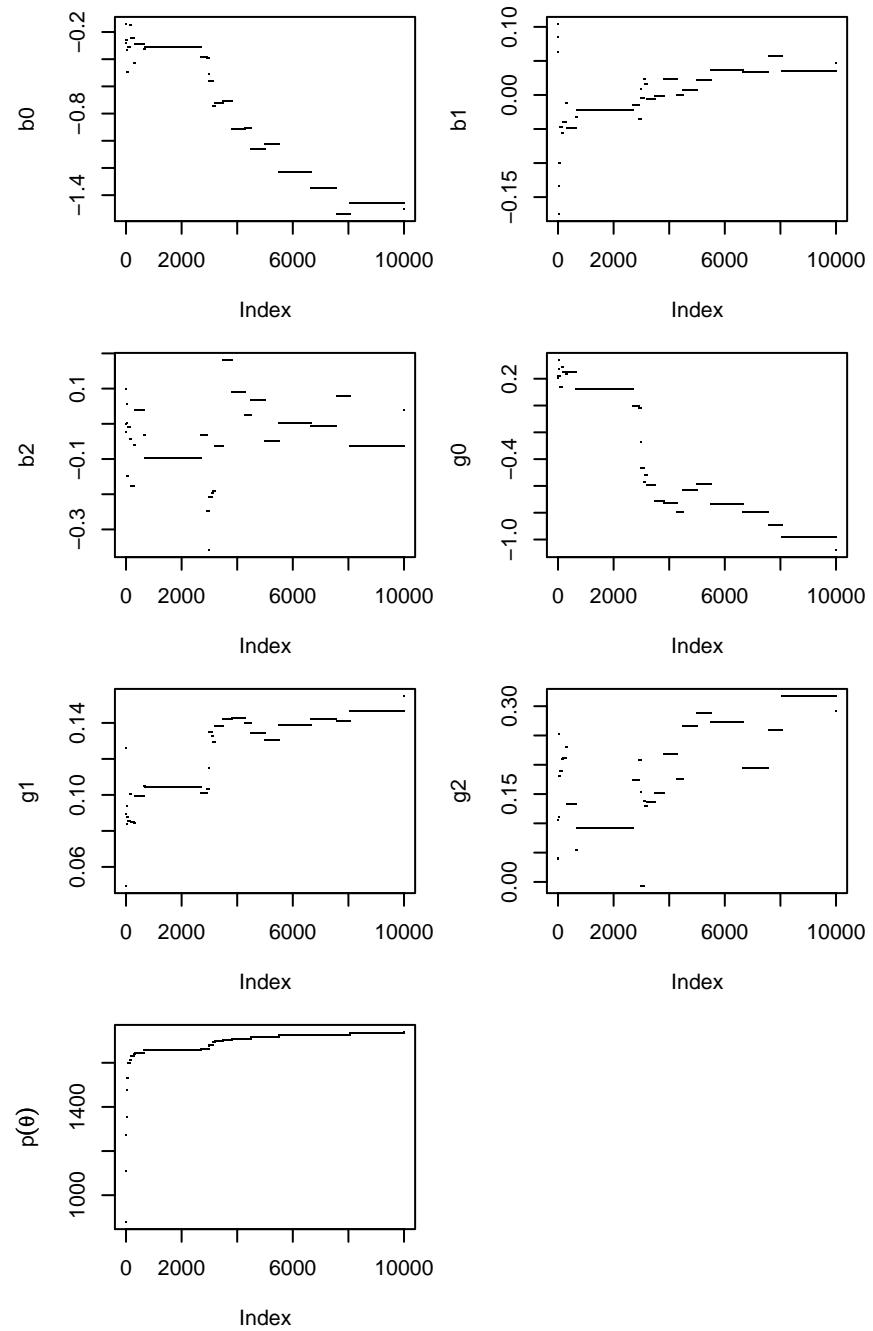


Figure 6.8: Trace plots of MCMC output from the pine cone code with a smaller proposal radius.

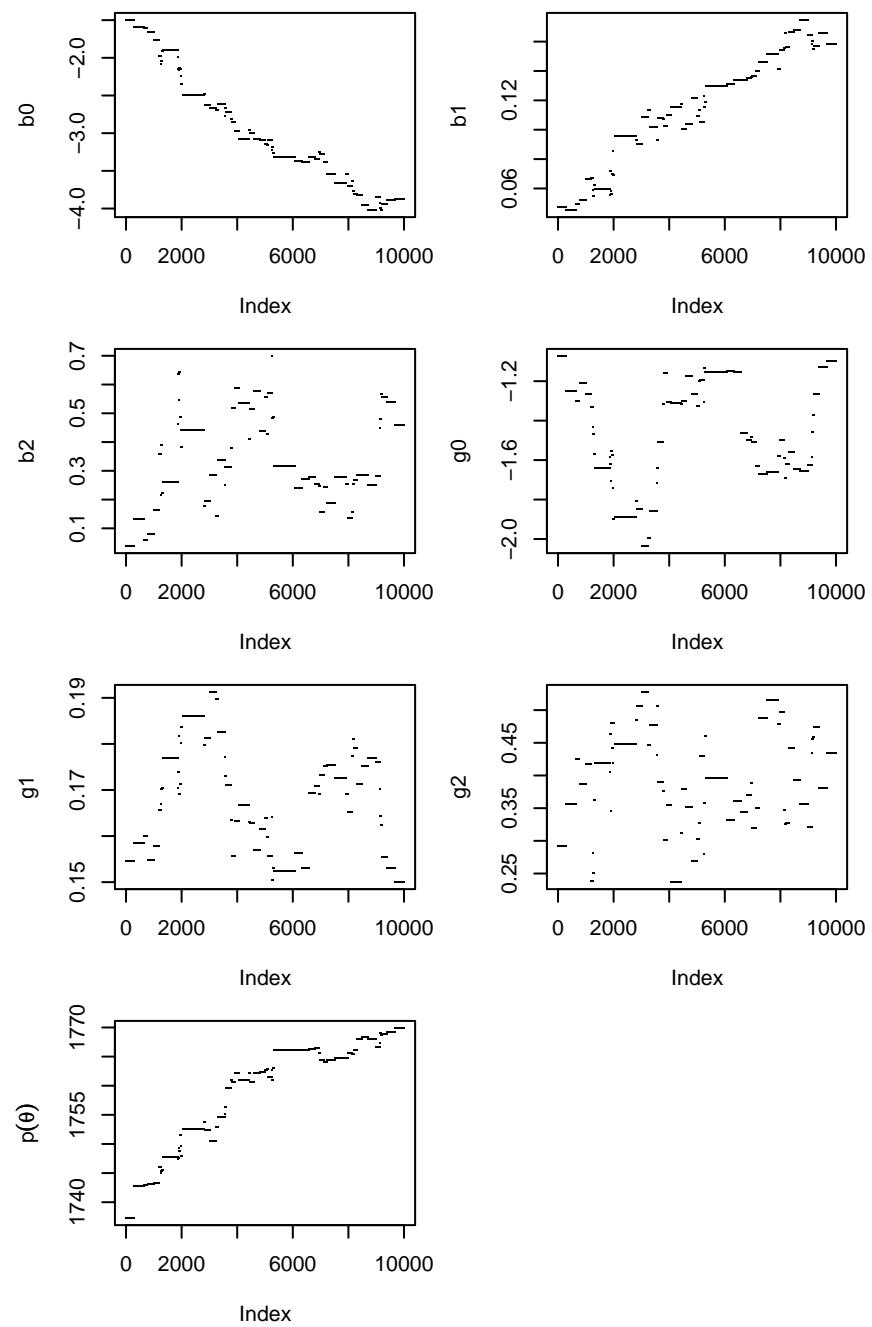


Figure 6.9: Trace plots of MCMC output from the pine cone code with a smaller proposal radius and 100,000 iterations. The plots show every 10'th iteration.

It is not always necessary to plot every iteration of an MCMC sampler. Figure 6.9 plots every 10'th iteration; plots of every iteration look similar. The figure was produced by the following snippet.

```
par (mfrow=c(4,2), mar=c(4,4,1,1)+.1)
plotem <- seq (1, 100000, by=10)
for (i in 1:6)
 plot (mcmc.out[plotem,i], ylab=names(params)[i], pch=".")
 plot (mcmc.out[plotem,7], ylab=expression(p(theta)), pch=".")
```

The sampler isn't mixing well. To write a better one we should try to understand why this one is failing. It could be that proposing a change in all parameters simultaneously is too dramatic, that once the sampler reaches a location where  $p(\vec{\theta})$  is large, changing all the parameters at once is likely to result in a location where  $p(\vec{\theta})$  is small, therefore the acceptance ratio will be small, and the proposal will likely be rejected. To ameliorate the problem we'll try proposing a change to only one parameter at a time. The new proposal function is

```
g.one <- function (params) {
 sig <- c ("b0"=.1, "b1"=.1, "b2"=.1, "g0"=.1, "g1"=.1, "g2"=.1)
 which <- sample (names(params), 1)
 proposed <- params
 proposed[which] <- rnorm (1, mean=params[which], sd=sig[which])
 return (list (proposed=proposed, ratio=1))
}
```

which randomly chooses one of the six parameters and proposes to update that parameter only. Naturally, we edit the main loop to use `g.one` instead of `g.all`. Figure 6.10 shows the result. This is starting to look better. Parameters  $\beta_2$  and  $\gamma_2$  are exhibiting steady-state behavior; so are  $\beta_0$  and  $\beta_1$ , after iteration 10,000 or so ( $x = 1000$  in the plots). Still,  $\gamma_0$  and  $\gamma_1$  do not look like they have converged.

Figure 6.11 illuminates some of the problems. In particular,  $\beta_0$  and  $\beta_1$  seem to be linearly related, as do  $\gamma_0$  and  $\gamma_1$ . This is often the case in regression problems; and we have seen it before for the pine cones in Figure 3.15. In the current setting it means that  $p(\vec{\theta}|y_1, \dots, y_n)$  has ridges: one along a line in the  $(\beta_0, \beta_1)$  plane and another along a line in the  $(\gamma_0, \gamma_1)$  plane.

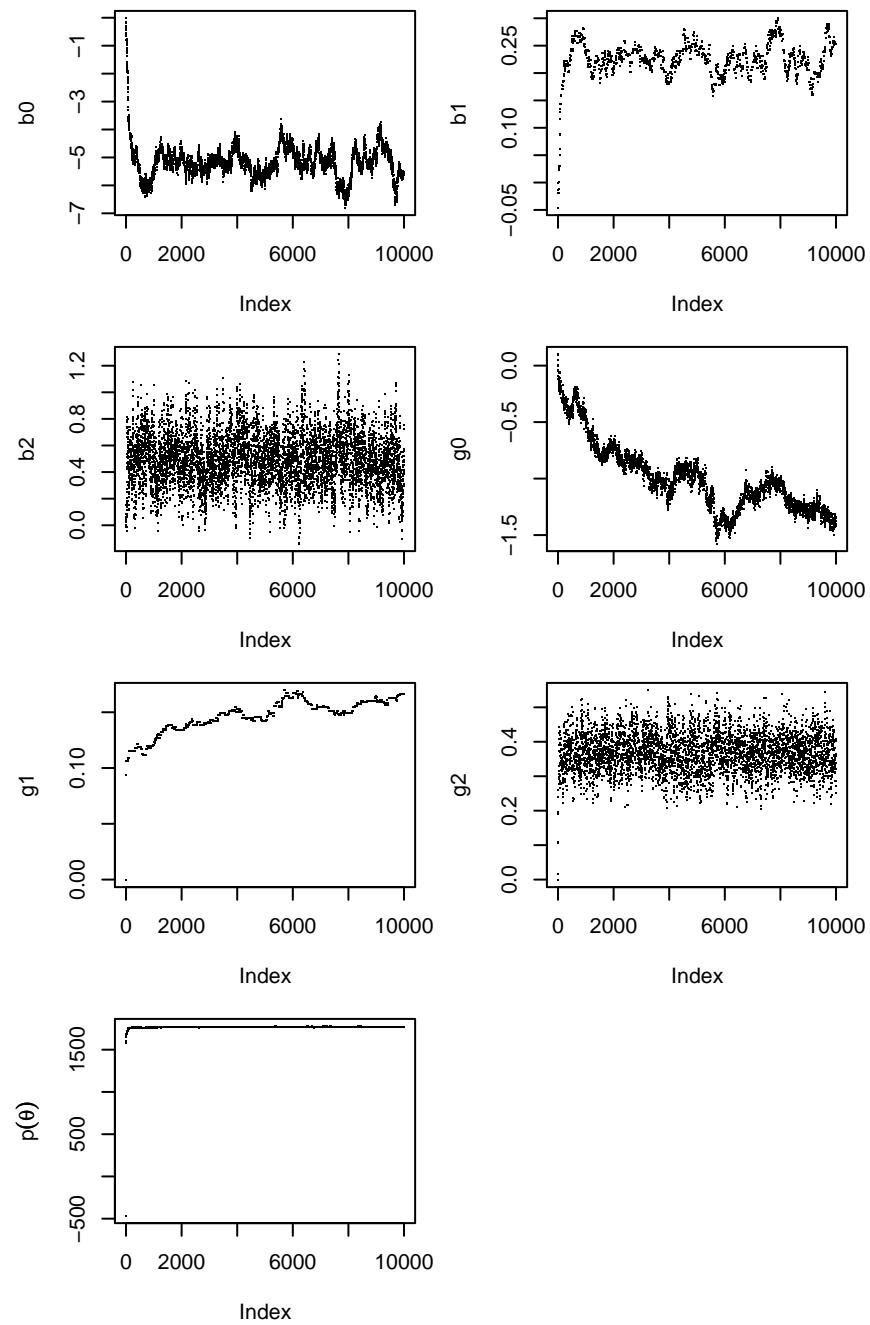


Figure 6.10: Trace plots of MCMC output from the pine cone code with proposal function `g.one` and 100,000 iterations. The plots show every 10'th iteration.

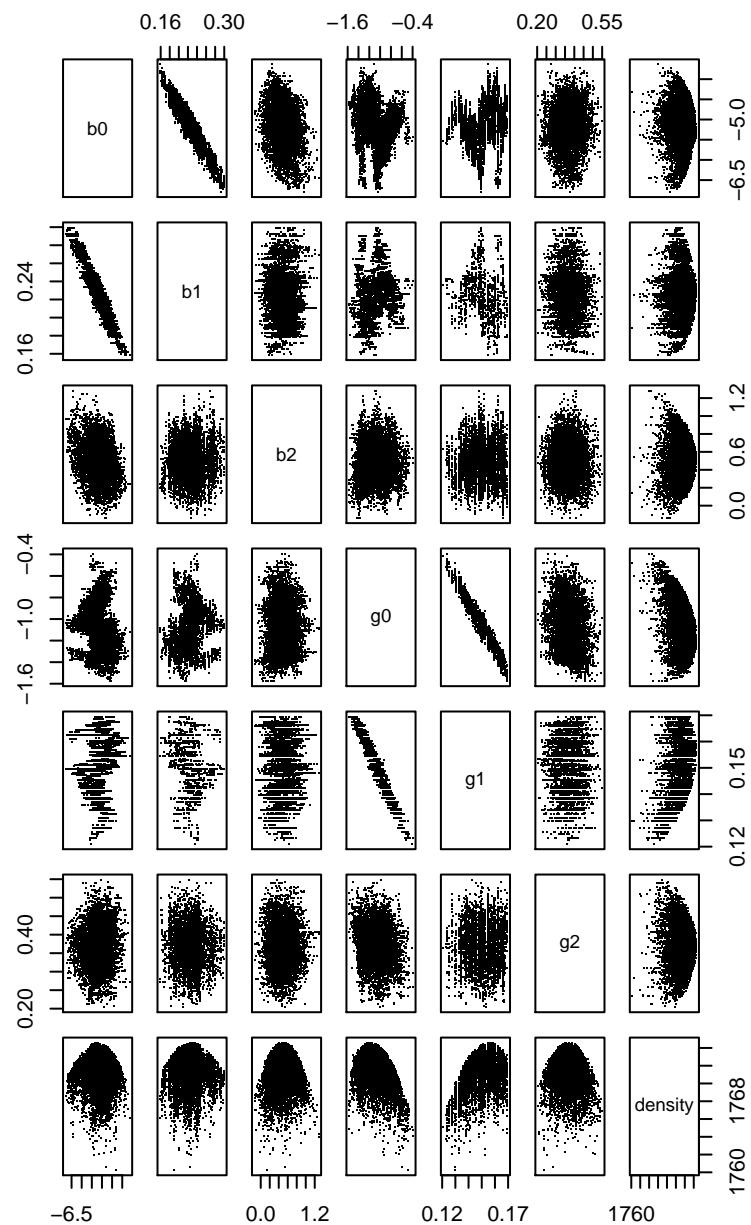


Figure 6.11: Pairs plots of MCMC output from the pine cones example.

Figure 6.11 was produced by the following snippet.

```
plotem <- seq (10000, 100000, by=10)
pairs (mcmc.out[plotem,], pch=".",
 labels=c(names(params), "density"))
```

As Figure 6.10 shows, it took the first 10,000 iterations or so for  $\beta_0$  and  $\beta_1$  to reach a roughly steady state and for  $p(\vec{\theta})$  to climb to a reasonably large value. If those iterations were included in Figure 6.11, the points after iteration 10,000 would be squashed together in a small region. Therefore we made `plotem <- seq ( 10,000, 100000, by=10 )` to drop the first 9999 iterations from the plots.

If our MCMC algorithm proposes a move along the ridge, the proposal is likely to be accepted. But if the algorithm proposes a move that takes us off the ridge, the proposal is likely to be rejected because  $p$  would be small and therefore the acceptance ratio would be small. But that's not happening here: our MCMC algorithm seems not to be stuck, so we surmise that it is proposing moves that are small compared to the widths of the ridges. However, because the proposals are small, the chain does not explore the space quickly. That's why  $\gamma_0$  and  $\gamma_1$  appear not to have reached a steady state. We could improve the algorithm by proposing moves that are roughly parallel to the ridges. And we can do that by making multivariate Normal proposals with a covariance matrix that approximates the posterior covariance of the parameters. We'll do that by finding the covariance of the samples we've generated and using it as the covariance matrix of our proposal distribution. The R code is

```
Sig <- cov (mcmc.out[10000:100000,-7])
g.group <- function (params) {
 proposed <- mvrnorm (1, mu=params, Sigma=Sig)
 return (list (proposed=proposed, ratio=1))
}
```

We drop the first 9999 iterations because they seem not to reflect  $p(\vec{\theta})$  accurately. Then we calculate the covariance matrix of the samples from the previous MCMC sampler. That covariance matrix is used in the proposal function. The results are shown in Figures 6.12 and 6.13. Figure 6.12 shows that the sampler seems to have converged after the first several thousand iterations. The posterior density has risen to a high level and is hovering there; all six variables appear to be mixing well. Figure 6.13

confirms our earlier impression that the posterior density seems to be approximately Normal — at least, it has Normal-looking two dimensional marginals — with  $\beta_0$  and  $\beta_1$  highly correlated with each other,  $\gamma_1$  and  $\gamma_2$  highly correlated with each other, and no other large correlations. The sampler seems to have found one mode and to be exploring it well.

Figures 6.12 and 6.13 were produced with the following snippet.

```
plotem <- seq (1, 100000, by=10)
par (mfrow=c(4,2), mar=c(4,4,1,1)+.1)
for (i in 1:6)
 plot (mcmc.out[plotem,i], ylab=names(params)[i], pch=".")
plot (mcmc.out[plotem,7], ylab=expression(p(theta)), pch=".")

plotem <- seq (1000, 100000, by=10)
pairs (mcmc.out[plotem,], pch=".",
 labels=c(names(params),"density"))
```

Now that we have a good set of samples from the posterior, we can use it to answer substantive questions. For instance, we might want to know whether the extra atmospheric CO<sub>2</sub> has allowed pine trees to reach sexual maturity at an earlier age or to produce more pine cones. This is a question of whether  $\beta_2$  and  $\gamma_2$  are positive, negative, or approximately zero. Figure 6.14 shows the answer by plotting the posterior densities of  $\beta_2$  and  $\gamma_2$ . Both densities put almost all their mass on positive values, indicating that  $P[\beta_2 > 0]$  and  $P[\gamma_2 > 0]$  are both very large, and therefore that pine trees with excess CO<sub>2</sub> mature earlier and produce more cones than pine trees grown under normal conditions.

Figure 6.14 was produced by the following snippet.

```
par (mfrow=c(1,2))
plot (density (mcmc.out[10000:100000,"b2"]),
 xlab=expression(beta[2]),
 ylab=expression(p(beta[2])), main="")
plot (density (mcmc.out[10000:100000,"g2"]),
 xlab=expression(gamma[2]),
 ylab=expression(p(gamma[2])), main="")
```

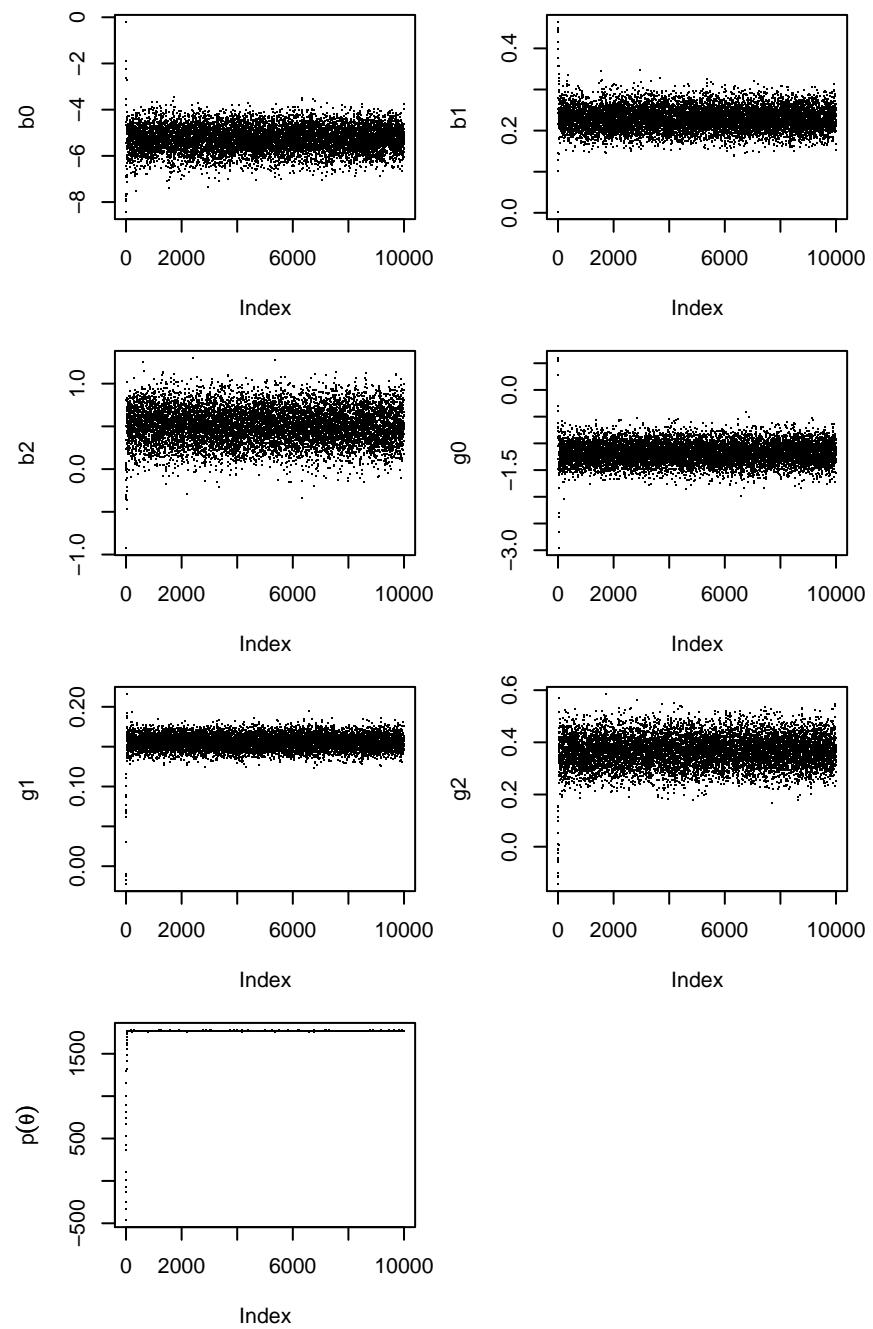


Figure 6.12: Trace plots of MCMC output from the pine cone code with proposal function `g.group` and 100,000 iterations. The plots show every 10'th iteration.

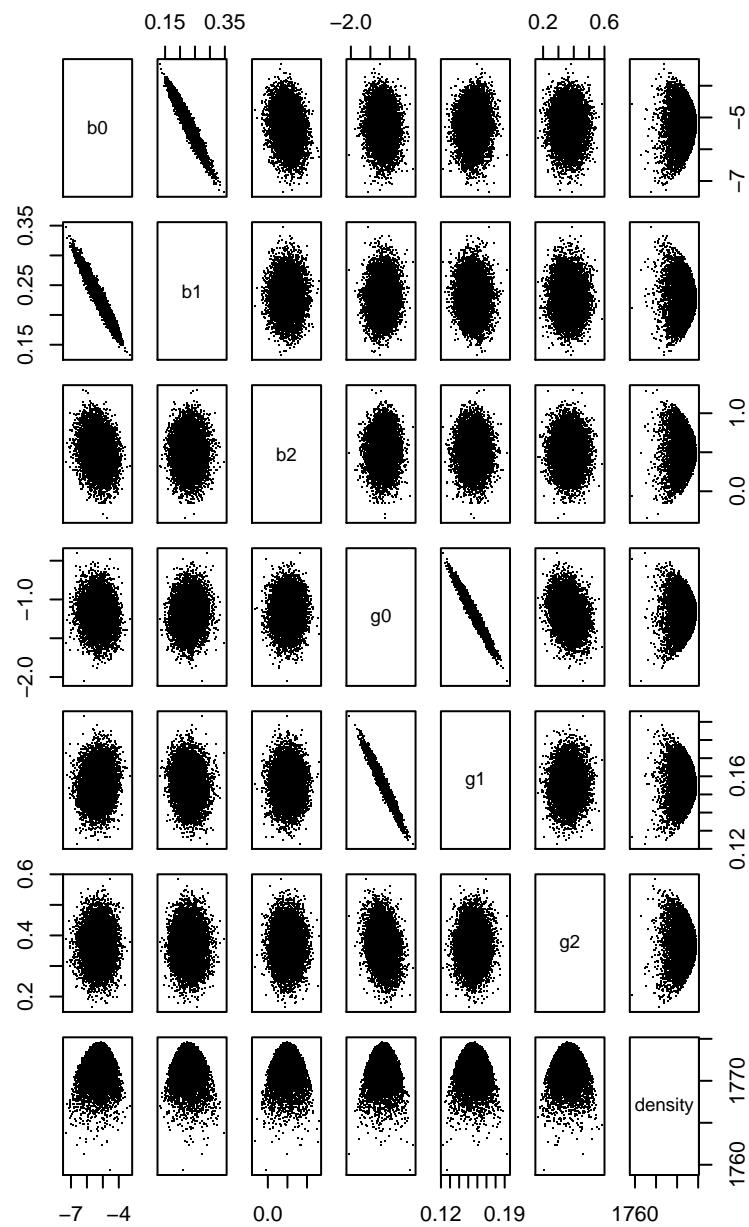


Figure 6.13: Pairs plots of MCMC output from the pine cones example with proposal `g.group`.

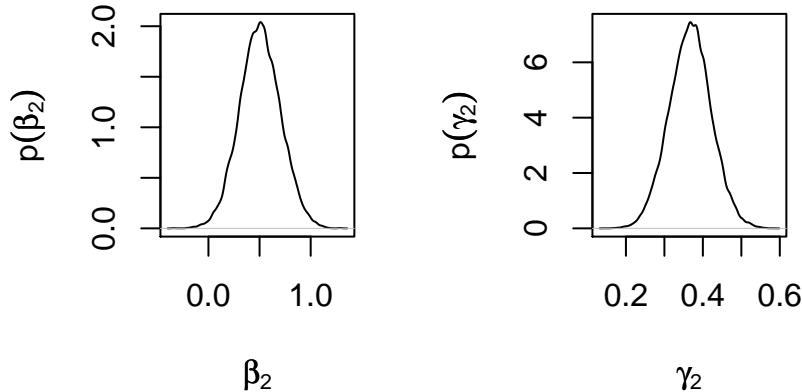


Figure 6.14: Posterior density of  $\beta_2$  and  $\gamma_2$  from Example 6.3.

## 6.3 Exercises

1. This exercise follows from Example 6.1.
  - (a) Find the posterior density for  $C_{50}$ , the expected amount of ice cream consumed when the temperature is 50 degrees, by writing  $C_{50}$  as a linear function of  $(\beta_0, \beta_1)$  and using the posterior from Example 6.1.
  - (b) Find the posterior density for  $C_{50}$  by reparameterizing. Instead of working with parameters  $(\beta_0, \beta_1)$ , work with parameters  $(C_{50}, \beta_1)$ . Write the equation for  $Y_i$  as a linear function of  $(C_{50}, \beta_1)$  and find the new  $X$  matrix. Use that new matrix and a convenient prior to calculate the posterior density of  $(C_{50}, \beta_1)$ .
  - (c) Do parts (a) and (b) agree about the posterior distribution of  $C_{50}$ . Does part (b) agree with Example 6.1 about the posterior distribution of  $\beta_1$ ? *Should* they agree?
2. This exercise asks you to enhance the code for the  $\text{Be}(5, 2)$  example on page 354.
  - (a) How many samples is enough? Instead of 10,000, try different numbers. How few samples can you get away with and still have an adequate approximation to the  $\text{Be}(5, 2)$  distribution? You must decide what "adequate" means; you can

- use either a firm or fuzzy definition. Illustrate your results with figures similar to 6.5.
- (b) Try an independence sampler in the  $\text{Be}(5, 2)$  example on page 354. Replace the proposal kernel with  $\theta^* \sim U(0, 1)$ . Run the sampler, make a figure similar to Figure 6.5 and describe the result.
  - (c) Does the proposal distribution matter? Instead of proposing with a radius of 0.1, try different numbers. How much does the proposal radius matter? Does the proposal radius change your answer to part 2A? Illustrate your results with figures similar to 6.5.
  - (d) Try a non-symmetric proposal. For example, you might try a proposal distribution of  $\text{Be}(5, 1)$ , or a distribution that puts 2/3 of its mass on  $(x_{i-1} - .1, x_{i-1})$  and 1/3 of its mass on  $(x_{i-1}, x_{i-1} + .1)$ . Illustrate your results with figures similar to 6.5.
  - (e) What would happen if your proposal distribution were  $\text{Be}(5, 2)$ ? How would the algorithm simplify?
3. (a) Some researchers are interested in  $\theta$ , the proportion of students who ever cheat on college exams. They randomly sample 100 students and ask “Have you ever cheated on a college exam?” Naturally, some students lie. Let  $\phi_1$  be the proportion of non-cheaters who lie and  $\phi_2$  be the proportion of cheaters who lie. Let  $X$  be the number of students who answer “yes” and suppose  $X = 40$ .
- i. Create a prior distribution for  $\theta$ ,  $\phi_1$ , and  $\phi_2$ . Use your knowledge guided by experience. Write a formula for your prior and plot the marginal prior density of each parameter.
  - ii. Write a formula for the likelihood function  $\ell(\theta, \phi_1, \phi_2)$ .
  - iii. Find the m.l.e..
  - iv. Write a formula for the joint posterior density  $p(\theta, \phi_1, \phi_2 | X = 40)$ .
  - v. Write a formula for the marginal posterior density  $p(\theta | X = 40)$ .
  - vi. Write an MCMC sampler to sample from the joint posterior.
  - vii. Use the sampler to find  $p(\theta | X = 40)$ . Summarize your results. Include information on how you assessed mixing and on what you learned about  $p(\theta | X = 40)$ .
  - viii. Assess the sensitivity of your posterior,  $p(\theta | X = 40)$ , to your prior for  $\phi_1$  and  $\phi_2$ .
- (b) **Randomized response** This part of the exercise uses ideas from Exercises 36 in Chapter 1 and 6 in Chapter 2. As explained there, researchers will sometimes instruct subjects as follows.

Toss a coin, but don't show it to me. If it lands Heads, answer question **(a)**. If it lands tails, answer question **(b)**. Just answer 'yes' or 'no'. Do not tell me which question you are answering.

- (a)** Does your telephone number end in an even digit?
- (b)** Have you ever cheated on an exam in college?

The idea of the randomization is, of course, to reduce the incentive to lie. Nonetheless, students may still lie.

- i. If about 40 students answered 'yes' in part (a), about how many do you think will answer 'yes' under the conditions of part (b)?
  - ii. Repeat part (a) under the conditions of part (b) and with your best guess about what  $X$  will be under these conditions.
  - iii. Assess whether researchers who are interested in  $\theta$  are better off using the conditions of part (a) or part (b).
4. Figures 6.12 and 6.13 suggest that the MCMC sampler has found one mode of the posterior density. Might there be others? Use the `lik` function and R's `optim` function to find out. Either design or randomly generate some starting values (You must decide on good choices for either the design or the randomization.) and use `optim` to find a mode of the likelihood function. Summarize and report your results.
  5. Example 6.3 shows that  $\beta_2$  and  $\gamma_2$  are very likely positive, and therefore that pine trees with extra CO<sub>2</sub> mature earlier and produce more cones. But how much earlier and how many more?
    - (a) Find the posterior means  $\mathbb{E}[\beta_2 | y_1, \dots, y_n]$  and  $\mathbb{E}[\gamma_2 | y_1, \dots, y_n]$  approximately, from the Figures in the text.
    - (b) Suppose there are three trees in the control plots that have probabilities 0.1, 0.5, and 0.9 of being sexually mature. Plugging in  $\mathbb{E}[\beta_2 | y_1, \dots, y_n]$  from the previous question, estimate their probabilities of being mature if they had grown with excess CO<sub>2</sub>.
    - (c) Is the plug-in estimate from the previous question correct? I.e., does it correctly calculate the probability that those trees would be sexually mature? Explain why or why not. If it's not correct, explain how to calculate the probabilities correctly.
  6. In the context of Equation 6.7 in Example 6.3 we might want to investigate whether the coefficient of dbh should be the same for control trees and for treated trees.

- (a) Write down a model enhancing that on page 348 to allow for the possibility of different coefficients for different treatments.
- (b) What parts of the R code have to change?
- (c) Write the new code.
- (d) Run it.
- (e) Summarize and report results. Report any difficulties with modifying and running the code. Say how many iterations you ran and how you checked mixing. Also report conclusions: does it look like different treatments need different coefficients? How can you tell?

## CHAPTER 7

# MORE MODELS

This chapter takes up a wide variety of statistical models. It is beyond the scope of this book to give a full treatment of any one of them. But we hope to introduce each model enough so the reader can see in what situations it might be useful, what its primary characteristics are, and how a simple analysis might be carried out in R. A more thorough treatment of many of these models can be found in VENABLES AND RIPLEY [2002].

Specialized models call for specialized methods. Because not all methods are built in to R, many people have contributed packages of specialized methods. Their packages can be downloaded from [HTTP://PROBABILITY.CA/CRAN/](http://PROBABILITY.CA/CRAN/). To use a package, you must first install it. Then, in each session of R in which you want to use it, you must load it. For example, to use the `survival` package for survival analysis you must first type `install.packages("survival")`. R should respond either by installing the package or by asking you to choose a mirror, then installing the package. Next, whenever you want to use the package, you must type `library("survival")`.

## 7.1 Fixed Effects, Random Effects, Hierarchical Models

To explain the ideas of fixed effects, random effects, and hierarchical models, it is best to look at an example. The `nlme` package comes with a dataset `Orthodont`. As the help file explains,

“Investigators at the University of North Carolina Dental School followed the growth of 27 children (16 males, 11 females) from age 8 until age 14. Every two years they measured the distance between the pituitary and the pterygomaxillary fissure, two points that are easily identified on x-ray exposures of the side of the head. . . .”

“[The] data frame contains the following columns:

**distance** a numeric vector of distances from the pituitary to the pterygomaxillary fissure (mm). These distances are measured on x-ray images of the skull.

**age** a numeric vector of ages of the subject (yr).

**Subject** an ordered factor indicating the subject on which the measurement was made. The levels are labelled M01 to M16 for the males and F01 to F13 for the females. The ordering is by increasing average distance within sex.

**Sex** a factor with levels Male and Female”

Each child was measured four times, so there are 108 lines in the data set. The first several lines of data are

```
Grouped Data: distance ~ age | Subject
 distance age Subject Sex
 1 26.0 8 M01 Male
 2 25.0 10 M01 Male
 3 29.0 12 M01 Male
 4 31.0 14 M01 Male
 5 21.5 8 M02 Male
 6 22.5 10 M02 Male
```

Figure 7.1 shows the data and was produced by `xyplot ( distance ~ age | Sex, data = Orthodont, groups = Subject, type="o", col=1 )`. The function `xyplot` is part of the `lattice` package. As illustrated in the figure, it produces plots of  $Y$  versus  $X$  in which points can be grouped by one variable and separated into different panels by another variable.

The figure shows that for most subjects, distance is an approximately linear function of age, that males tend to be larger than females, and that different subjects have different intercepts. Therefore, to fit this data we need a model with the following features: (1) a parameter for the average intercept of males (or females); (2) a parameter for the average difference in intercepts for males and females; (3) one parameter per subject for how that subject’s intercept differs from its group average; and (4) one parameter for the slope. Model 7.1 has those features.

$$Y_{i,t} = \text{distance for subject } i \text{ at time } t$$

$$Y_{i,t} \sim N(\mu_{i,t}, \sigma) \quad (7.1)$$

$$\mu_{i,t} = \beta_0 + \beta_1 \text{Sex}_i + \delta_i + \beta_2 t$$

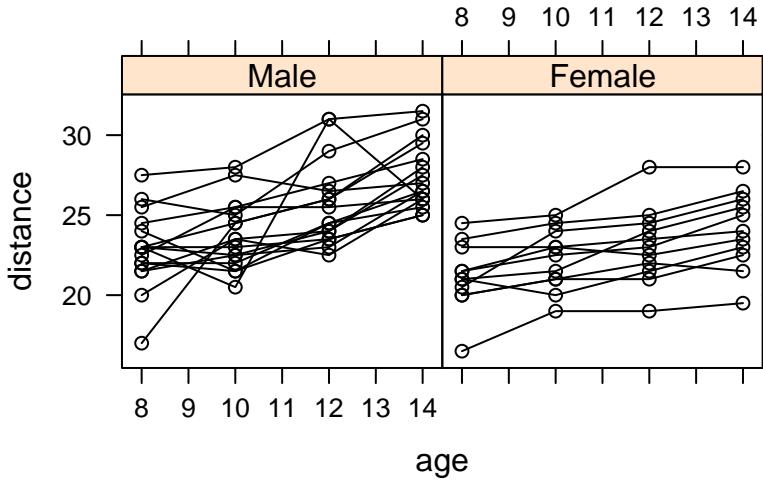


Figure 7.1: Plots of the Orthodont data: distance as a function of age, grouped by Subject, separated by Sex.

Here is the key feature: we want  $\beta_0$  to represent the average intercept of all males and  $\beta_0 + \beta_1$  to represent the average intercept of all females, not just those in this study; and we want to think of the  $\delta_i$ 's in this study as random draws from a population of  $\delta$ 's representing all children. The terminology is that  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  are *fixed* effects because they describe the general population, not the particular individuals in the sample, while  $\delta_1, \dots, \delta_n$  are *random* effects because they are random draws from the general population. It is customary to add a term to the model to describe the distribution of random effects. In our case we shall model the  $\delta_i$ 's as  $N(0, \sigma_{\text{ran. eff.}})$ . Thus Model 7.1 becomes

$$\begin{aligned}
 Y_{i,t} &= \text{distance for subject } i \text{ at time } t \\
 Y_{i,t} &\sim N(\mu_{i,t}, \sigma) \\
 \mu_{i,t} &= \beta_0 + \beta_1 \text{Sex}_i + \delta_i + \beta_2 t \\
 \delta_1, \dots, \delta_n &\sim \text{i.i.d. } N(0, \sigma_{\text{ran. eff.}})
 \end{aligned} \tag{7.2}$$

Modelling the random effects as Normal is an arbitrary choice. We could have used a different distribution, but the Normal is convenient and not obviously contraindicated by Figure 7.1. Choosing the mean of the Normal distribution to be 0 is not arbitrary. It's the result of thinking of the  $\delta_i$ 's as draws from a larger population. In any population, the average departure from the mean must be 0.

Model 7.2 is called a *mixed* effects model because it contains both fixed and random effects. Often, in mixed effects models, we are interested in  $\sigma_{\text{ran. eff.}}$ . Whether we are interested in the individual  $\delta_i$ 's depends on the purpose of the investigation.

We can fit Model 7.2 by the following R code.

```
ortho.fit1 <- lme (distance ~ age + Sex, random = ~ 1,
 data = Orthodont)
```

- `lme` stands for “linear mixed effects model.” It is in the `nlme` package.
- The formula `distance ~ age + Sex` is just like formulas for linear models.
- `random = ~ 1` specifies the random effects. In this case, the random effects are intercepts, or coefficients of 1. If we thought that each child had his or her own slope, then we would have said `random = ~ age - 1` to say that the random effects are coefficients of age but not of 1.

Printing `ortho.fit1` yields

```
Linear mixed-effects model fit by REML
 Data: Orthodont
 Log-restricted-likelihood: -218.7563
 Fixed: distance ~ age + Sex
 (Intercept) age SexFemale
 17.7067130 0.6601852 -2.3210227
```

Random effects:

```
 Formula: ~1 | Subject
 (Intercept) Residual
 StdDev: 1.807425 1.431592
```

Number of Observations: 108

Number of Groups: 27

The fixed effects part of the output is just like the fixed effects part of linear models. The estimates of  $(\beta_0, \beta_1, \beta_2)$  are around  $(17.7, -2.32, 0.66)$ . The random effects part of the output shows that the estimate of  $\sigma$  is about 1.43 while the estimate of  $\sigma_{\text{ran. eff.}}$  is about 1.81. In other words, most of the intercepts are within about 3.6 or so of their mean and the differences from child to child are about the same size as any unexplained variation with each child's data. We can see whether that's sensible by inspecting Figure 7.1. It also

appears, from Figure 7.1, that one of the males doesn't fit the general pattern. Because that subject's data fluctuates wildly from the pattern, it may have inflated the estimate of  $\sigma$ . We might want to remove that subject's data and refit the model, just to see how influential it really is. That's a topic we don't pursue here.

We can see estimates of  $\delta_1, \dots, \delta_{27}$  by typing `random.effects(ortho.fit1)`, which yields

```
(Intercept)
M16 -1.70183357
M05 -1.70183357
M02 -1.37767479
M11 -1.16156894
M07 -1.05351602
M08 -0.94546309
M03 -0.62130432
M12 -0.62130432
M13 -0.62130432
M14 -0.08103969
M09 0.13506616
M15 0.78338371
M06 1.21559540
M04 1.43170125
M01 2.40417758
M10 3.91691853
F10 -3.58539251
F09 -1.31628108
F06 -1.31628108
F01 -1.10017523
F05 -0.01964599
F07 0.30451279
F02 0.30451279
F08 0.62867156
F03 0.95283034
F04 1.92530666
F11 3.22194176
```

The reasonableness of these estimates can be judged by comparison to Figure 7.2, where we can also see that the strange male is M09.

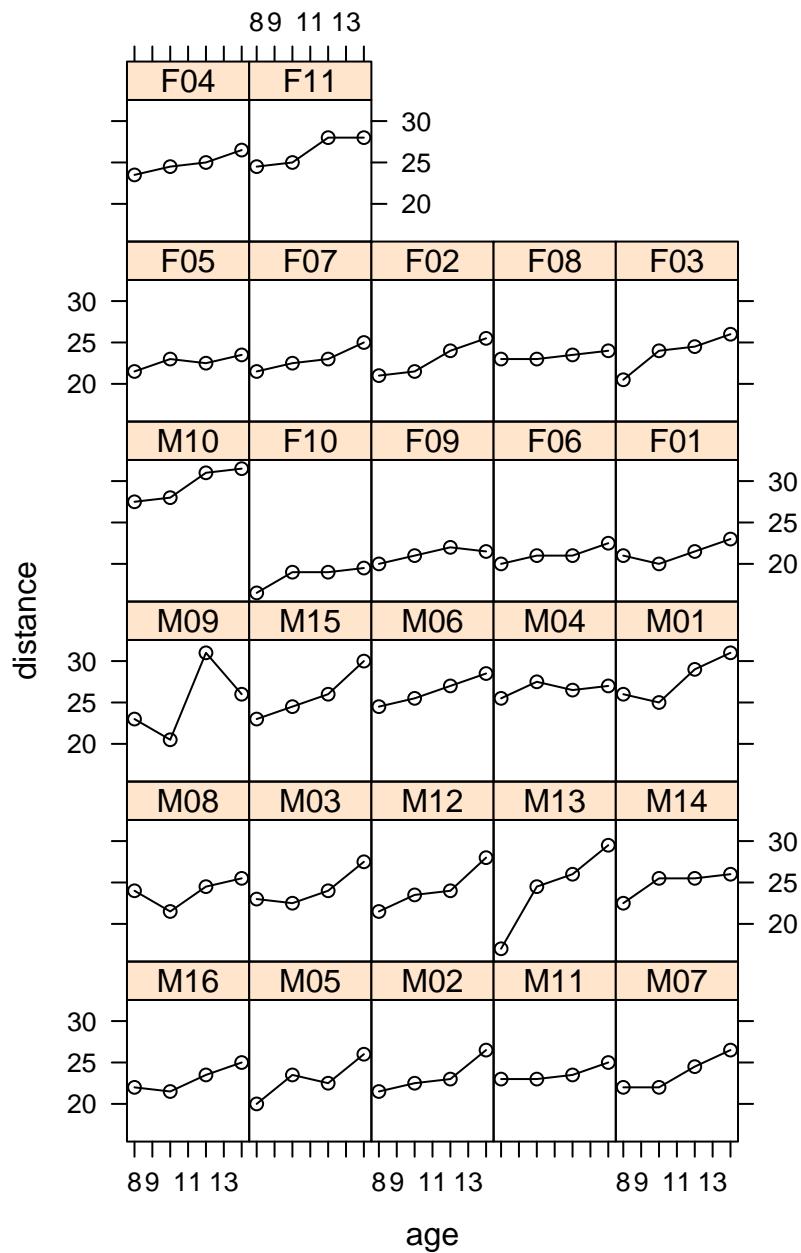


Figure 7.2: Plots of the Orthodont data: distance as a function of age, separated by Subject.

To see what is gained by fitting the mixed effects model, we compare it to two other models that have fixed effects only.

```
ortho.fit2 <- lm (distance ~ age + Sex, data = Orthodont)
ortho.fit3 <- lm (distance ~ age+Sex+Subject, data=Orthodont)
```

The summary of `ortho.fit2` is

```
...
Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.70671 1.11221 15.920 < 2e-16 ***
age 0.66019 0.09776 6.753 8.25e-10 ***
SexFemale -2.32102 0.44489 -5.217 9.20e-07 ***
...
Residual standard error: 2.272 on 105 degrees of freedom
...
```

The estimates of the coefficients are exactly the same as in `ortho.fit1`. But, since `ortho.fit2` ignores differences between subjects, it attributes those differences to the observation SD  $\sigma$ . That's why the estimate of  $\sigma$  from `ortho.fit1`, 1.43, is smaller than the estimate from `ortho.fit2`, 2.27.

The summary of `ortho.fit3` is

```
Coefficients: (1 not defined because of singularities)
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.19192 1.85453 12.506 < 2e-16 ***
age 0.66019 0.06161 10.716 < 2e-16 ***
SexFemale -15.78472 4.22373 -3.737 0.000348 ***
Subject.L 33.91970 9.21165 3.682 0.000418 ***
Subject.Q 6.64753 2.31508 2.871 0.005228 **
Subject.C -5.80525 2.99918 -1.936 0.056448 .
Subject^4 -3.18380 2.17427 -1.464 0.147028
Subject^5 3.26565 1.39346 2.344 0.021584 *
...
Subject^25 -0.14291 0.89189 -0.160 0.873100
Subject^26 NA NA NA NA
```

We see that the estimates of  $\beta_0$  and  $\beta_1$  have changed. That's because, in the fixed effects model, those estimates are too dependent on the particular subjects in the study. The

estimates from `ortho.fit1` are preferred, as they come from a model that describes the data more correctly.

Finally, in Figure 7.1 there is a slight suggestion that the slope for males is slightly larger than the slope for females. Model 7.3 allows for that possibility:  $\beta_3$  is the difference in slope between males and females.

$$\begin{aligned} Y_{i,t} &= \text{distance for subject } i \text{ at time } t \\ Y_{i,t} &\sim N(\mu_{i,t}, \sigma) \\ \mu_{i,t} &= \beta_0 + \beta_1 \text{Sex}_i + \delta_i + \beta_2 t + \beta_3 t \times \text{Sex}_i \\ \delta_1, \dots, \delta_n &\sim \text{i.i.d.} N(0, \sigma_{\text{ran. eff.}}) \end{aligned} \tag{7.3}$$

The model can be fit by `ortho.fit4 <- lme ( distance ~ age * Sex, random = ~ 1, data = Orthodont )` and yields

```
...
Random effects:
Formula: ~1 | Subject
 (Intercept) Residual
StdDev: 1.816214 1.386382

Fixed effects: distance ~ age * Sex
 Value Std.Error DF t-value p-value
(Intercept) 16.340625 0.9813122 79 16.651810 0.0000
age 0.784375 0.0775011 79 10.120823 0.0000
SexFemale 1.032102 1.5374208 25 0.671321 0.5082
age:SexFemale -0.304830 0.1214209 79 -2.510520 0.0141
...
```

This model suggests that the slope for males is about 0.78 while the slope for females is about 0.3 less, plus or minus about 0.25 or so. The evidence is not overwhelming in either direction whether males and females have the same slope. But if they differ, it could be by an amount (roughly,  $.30 \pm 2 \times .12$ ) that is large compared to the slope for males. For its extra complexity, this model has reduced the residual SD, or our accuracy of prediction, from about 1.43 to about 1.39. At this point, we don't have a strong preference for either model and, if we were investigating further, would keep both of them in mind.

The next example comes from a study of how ants store nutrients over the winter.

### **Example 7.1** (Ant Fat)

The data in this example were collected to examine the strategy that ants of the

species *Pheidole morrisi* employ to store nutrients, specifically fat, over the winter; the study was reported by YANG [2006]. Many animals need to store nutrients over the winter when food is scarce. For most species, individual animals store their own nutrients. But for some species, nutrients can be stored collectively. As YANG explains, "Among ants, a common mechanism of colony fat storage is for workers of both castes (majors and minors) to uniformly increase the amount of fat they hold ... The goal of this study is to better understand the specific mechanisms by which ants use division of labor to store colony fat ...." The need to store fat varies with the severity of winter, which in turn varies with latitude. So Yang studied ants at three sites, one in Florida, one in North Carolina, and one in New York. At each site he dug up several ant colonies in the Spring and another several colonies in the Fall. From each colony, the ants were separated into their two castes, majors and minors, and the fat content of each caste was measured. The first several lines of data look like this.

|   | colony | season | site     | caste  | fat       |
|---|--------|--------|----------|--------|-----------|
| 1 | 1      | Spring | New York | minors | 15.819209 |
| 2 | 2      | Spring | New York | minors | 10.526316 |
| 3 | 3      | Spring | New York | minors | 18.534483 |
| 4 | 4      | Spring | New York | minors | 21.467098 |
| 5 | 5      | Spring | New York | minors | 9.784946  |
| 6 | 6      | Spring | New York | minors | 20.138289 |

...

There are 108 lines in all. Figure 7.3 shows the data along with kernel density estimates. It was produced by `densityplot (~ fat | site+season, groups=caste, data=ants, adjust=1.5)`. Purple lines are for minors; blue lines for majors. In New York and North Carolina, it appears that minors have, on average, less fat than majors and less variability. There is also some suggestion that fat content increases with increasing latitude. The main question is how the average percent fat differs by site, season, and caste. Because the predictors are categorical, we need to define indicator variables:  $NY_i = 1$  if the  $i$ 'th observation was from NY;  $NC_i = 1$  if the  $i$ 'th observation was from NC;  $Spring_i = 1$  if the  $i$ 'th observation was from Spring; and  $minors_i = 1$  if the  $i$ 'th observation was on the minor caste. With those conventions, we begin with the following linear model.

$$\begin{aligned}
 y_i = & \beta_0 + \beta_1 NY_i + \beta_2 NC_i + \beta_3 Spring_i + \beta_4 minors_i + \beta_5 NY_i Spring_i \\
 & + \beta_6 NC_i Spring_i + \beta_7 NY_i minors_i + \beta_8 NC_i minors_i + \beta_9 Spring_i minors_i \\
 & + \beta_{10} NY_i Spring_i minors_i + \beta_{11} NC_i Spring_i minors_i + \epsilon_i. \quad (7.4)
 \end{aligned}$$

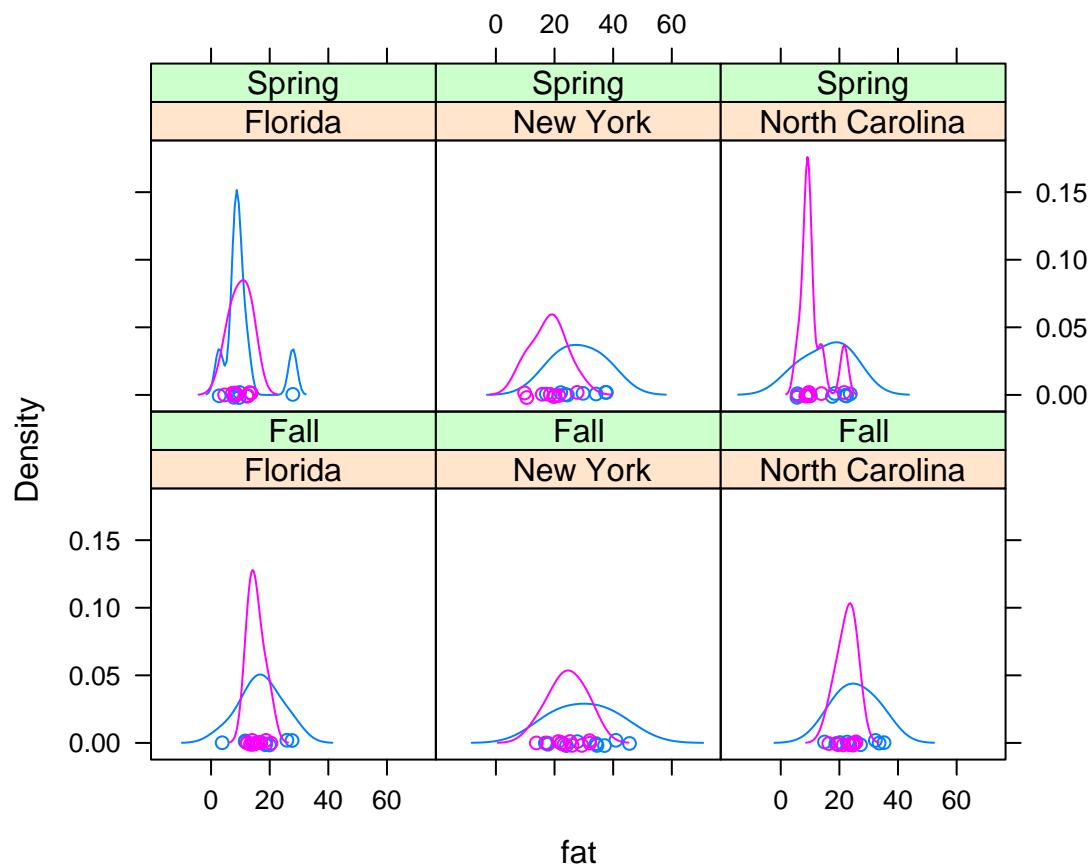


Figure 7.3: Percent body fat of major (blue) and minor (purple) *Pheidole morrisi* ants at three sites in two seasons.

In Model 7.4,  $\beta_0$  is the average fat content of major, Florida ants in the Fall;  $\beta_1$  is the difference between New York majors and Florida majors in the Fall;  $\beta_2$  is the difference between North Carolina majors and Florida majors in the Fall;  $\dots$ ;  $\beta_9$  is the difference between Florida majors in the Fall and Florida minors in the Spring;  $\dots$ ; etc. Model 7.4 can be fit by

```
fit1 <- lm (fat ~ site * season * caste, data=ants)
```

The parameter estimates and their SD's, given by `summary ( fit1 )`, are

|                                             |         |        |
|---------------------------------------------|---------|--------|
| (Intercept)                                 | 16.8857 | 2.0016 |
| siteNew York                                | 13.2084 | 2.8307 |
| siteNorth Carolina                          | 8.6182  | 2.8307 |
| seasonSpring                                | -6.0555 | 3.0024 |
| casteminors                                 | -1.5489 | 2.8307 |
| siteNew York:seasonSpring                   | 4.2587  | 4.2460 |
| siteNorth Carolina:seasonSpring             | -3.7789 | 4.2460 |
| siteNew York:casteminors                    | -4.2622 | 4.0032 |
| siteNorth Carolina:casteminors              | -1.6258 | 4.0032 |
| seasonSpring:casteminors                    | 0.6997  | 4.2460 |
| siteNew York:seasonSpring:casteminors       | -5.2305 | 6.0048 |
| siteNorth Carolina:seasonSpring:casteminors | -2.2812 | 6.0048 |

We see that New York majors are much fatter, by about 13.2 percentage points on average, than Florida majors in the Fall, while North Carolina majors are in between. That seems consistent with the theory that New York ants need to store more fat in the Fall because they are going to face a longer, harder winter than Florida ants. In the Spring, Florida majors have about  $16.9\% - 6.1\% \approx 13.8\%$  body fat while New York majors have about  $16.9\% + 13.2\% - 6.1\% + 4.3\% \approx 28.2\%$  body fat. It appears that New York majors did not lose as much fat over the winter as Florida majors. Perhaps they didn't need to store all that fat after all.

Before addressing that question more thoroughly, we want to examine one possible inadequacy of the model. The data sets contains two data points from each colony — one for majors, one for minors — and it's possible that those two data points are not independent. In particular, there might be colony effects; one colony might be fatter, on average, than another from the same site and season, and that extra fatness might apply to both castes. To see whether that's true, we'll examine residuals from `fit1`. Specifically, we'll plot residuals for minors on the abscissa and residuals for majors

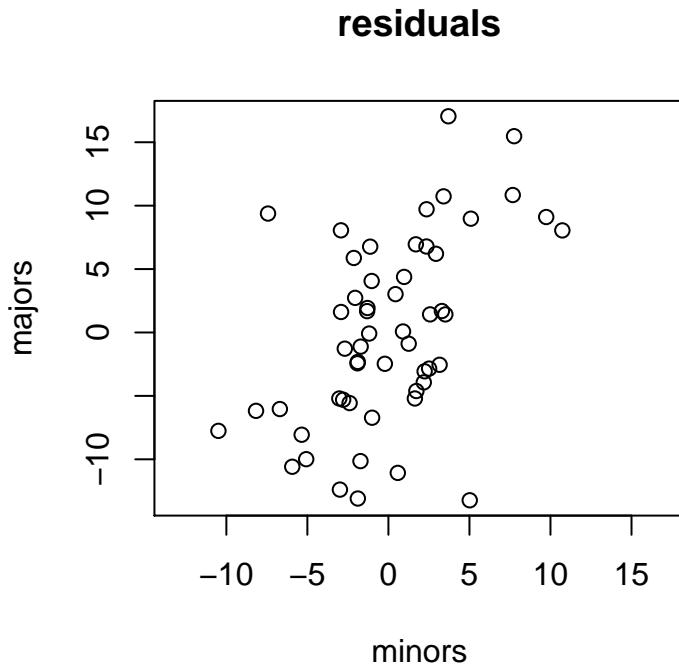


Figure 7.4: Residuals from Model 7.4. Each point represents one colony. There is an upward trend, indicating the possible presence of colony effects.

on the ordinate. Figure 7.4 is the plot. There is a clear, upward, approximately linear trend, indicating that majors and minors from one colony tend to be thin or fat together. We can capture that tendency by including random colony effects in the model. That leads to Model 7.5:

$$\begin{aligned}
 y_i = & \beta_0 + \beta_1 NY_i + \beta_2 NC_i + \beta_3 Spring_i + \beta_4 minors_i + \beta_5 NY_i Spring_i \\
 & + \beta_6 NC_i Spring_i + \beta_7 NY_i minors_i + \beta_8 NC_i minors_i + \beta_9 Spring_i minors_i \\
 & + \beta_{10} NY_i Spring_i minors_i + \beta_{11} NC_i Spring_i minors_i + \delta_{col(i)} + \epsilon_i
 \end{aligned} \quad (7.5)$$

which differs from Model 7.4 by the presence of the  $\delta$ 's, which are assumed to be distributed i.i.d. $N(0, \sigma_{\text{ran. eff.}})$ . Model 7.5 can be fit by

```
fit2 <- lme (fat~site*season*caste, data=ants, random = ~1|colony)
```

The R code (...random = ~1|colony) says to give each colony its own, random, intercept. The summary of fit2 is, in part,

```
Linear mixed-effects model fit by REML
Data: ants
 AIC BIC logLik
 671.0577 706.9585 -321.5288

Random effects:
Formula: ~1 | colony
 (Intercept) Residual
StdDev: 4.164122 4.76696

Fixed effects: fat ~ site * season * caste
 Value Std.Error
(Intercept) 16.885710 2.001595
siteNew York 13.208390 2.830683
siteNorth Carolina 8.618221 2.830683
seasonSpring -6.055542 3.002392
casteminors -1.548911 2.131849
siteNew York:seasonSpring 4.258693 4.246024
siteNorth Carolina:seasonSpring -3.778862 4.246024
siteNew York:casteminors -4.262151 3.014890
siteNorth Carolina:casteminors -1.625755 3.014890
seasonSpring:casteminors 0.699721 3.197774
siteNew York:seasonSpring:casteminors -5.230472 4.522335
siteNorth Carolina:seasonSpring:casteminors -2.281162 4.522335
...
Number of Observations: 108
Number of Groups: 54
```

There are several things to note about fit2. First,  $\hat{\sigma} \approx 4.77$  is somewhat smaller than the estimate from fit1. That's because some of the variability that fit1 attributes to the  $\epsilon$ 's, fit2 attributes to the  $\delta$ 's. Second,  $\hat{\sigma}_{\text{ran. eff.}} \approx 4.16$  is about the same size as  $\hat{\sigma}$ , indicating that colony effects explain a sizable portion of the variation in the data. Third, estimates of the  $\beta$ 's are unchanged. And fourth, some of the estimates of SD's of the  $\beta$ 's are changed while others are not. Specifically, all the SD's involving caste effects have decreased. That's because some of the variability that fit1 attributes to

| Site           | Season | Majors |        | Minors |        |
|----------------|--------|--------|--------|--------|--------|
|                |        | Fall   | Spring | Fall   | Spring |
| Florida        |        | 16.9   | 10.8   | 15.3   | 10.1   |
| North Carolina |        | 25.0   | 15.7   | 24.0   | 11.0   |
| New York       |        | 30.1   | 28.3   | 24.3   | 18.0   |

Table 7.1: Fat as a percentage of body weight in ant colonies. Three sites, two seasons, two castes.

the  $\epsilon$ 's, `fit2` attributes to the  $\delta$ 's.

With `fit2` in hand, we can turn our attention to the question of interest: what are the fat storage strategies and how do they differ from New York to Florida? Table 7.1 summarizes the analysis; the numbers come from adding the appropriate  $\hat{\beta}$ 's. The main points are

- Both castes in all locations stored more fat in the Fall than in the Spring.
- The amount of fat stored increases with increasing latitude.
- Majors store more fat than minors, in some cases by a lot, in other cases by a little.
- Moving from Florida to North Carolina to New York, majors increase their fat content in both Fall and Spring. But moving from Florida to North Carolina, minors increase their fat content only in the Fall while from North Carolina to New York, they increase only in the Spring.

The summary in Table 7.1 and the preceding bullets could have been carried out with no formal statistical analysis. That's the way it goes, sometimes. In this example, we could use the statistical analysis to be more precise about how accurately we can estimate each of the effects noted in the bullets. Here, there seems little need to do that.

YANG [2006] carries the analysis further. For one thing, he is more formal about statistics. In addition, he notes that majors can be further divided into repletes and non-repletes. According to YANG, “[R]epletes carry and store a disproportionate amount of nutrients relative to other individuals in a colony and provide it to other colony members through trophallaxis in times of food scarcity.” And according to Wikipedia, “Trophallaxis is the transfer of food or other fluids among members of a community through mouth-to-mouth (stomodeal) or anus-to-mouth (proctodeal) feeding. It is most highly

developed in social insects such as ants, termites, wasps and bees." YANG finds interesting differences in fat storage among repletes and other majors, differences that vary according to season and location.

## 7.2 Time Series and Markov Chains

Figure 7.5 shows some data sets that come with R. The following descriptions are taken from the R help pages.

**Beaver** The data are a small part of a study of the long-term temperature dynamics of beaver *Castor canadensis* in north-central Wisconsin. Body temperature was measured by telemetry every 10 minutes for four females, but data from one period of less than a day is shown here.

**Mauna Loa** Monthly atmospheric concentrations of CO<sub>2</sub> are expressed in parts per million (ppm) and reported in the preliminary 1997 SIO manometric mole fraction scale.

**DAX** The data are the daily closing prices of Germany's DAX stock index. The data are sampled in business time; i.e., weekends and holidays are omitted.

**UK Lung Disease** The data are monthly deaths from bronchitis, emphysema and asthma in the UK, 1974 – 1979.

**Canadian Lynx** The data are annual numbers of lynx trappings for 1821 – 1934 in Canada.

**Presidents** The data are (approximately) quarterly approval rating for the President of the United states from the first quarter of 1945 to the last quarter of 1974.

**UK drivers** The data are monthly totals of car drivers in Great Britain killed or seriously injured Jan 1969 to Dec 1984. Compulsory wearing of seat belts was introduced on 31 Jan 1983.

**Sun Spots** The data are monthly numbers of sunspots. They come from the World Data Center-C1 For Sunspot Index Royal Observatory of Belgium, Av. Circulaire, 3, B-1180 BRUSSELS [HTTP://WWW.OMA.BE/KSB-ORB/SIDC/SIDC\\_TXT.HTML](http://www.oma.be/KSB-ORB/SIDC/SIDC_TXT.HTML).

What these data sets have in common is that they were all collected sequentially in time. Such data are known as *time series data*. Because each data point is related to the ones before and the ones after, they usually cannot be treated as independent random variables. Methods for analyzing data of this type are called *time series methods*. More formally,

a time series is a sequence  $Y_1, \dots, Y_T$  of random variables indexed by time. The generic element of the series is usually denoted  $Y_t$ .

Figure 7.5 was produced by the following snippet.

```
par (mfrow=c(4,2), mar=c(3,4,1,1)+.1)
plot.ts (beaver1$temp, main="Beaver", xlab="Time",
 ylab="Temperature")
plot.ts (co2, main="Mauna Loa", ylab="CO2 (ppm)")
plot.ts (EuStockMarkets[,1], main="DAX",
 ylab="Closing Price")
plot.ts (ldeaths, main="UK Lung Disease",
 ylab="monthly deaths")
plot.ts (lynx, main="Canadian Lynx", ylab="trappings")
plot.ts (presidents, main="Presidents", ylab="approval")
plot.ts (Seatbelts[, "DriversKilled"], main="UK drivers",
 ylab="deaths")
plot.ts (sunspot.month, main="Sun Spots",
 ylab="number of sunspots")
```

- `plot.ts` is the command for plotting time series.
- The `mar` argument in the `par` command says how many lines are in the margins of each plot. Those lines are used for titles and axis labels. The command is used here to decrease the default so there is less white space between the plots, hence more room for each plot.

The data sets in Figure 7.5 exhibit a feature common to many time series: if one data point is large, the next tends to be large, and if one data point is small, the next tends to be small; i.e.,  $Y_t$  and  $Y_{t+1}$  are dependent. The dependence can be seen in Figure 7.6 which plots  $Y_{t+1}$  vs.  $Y_t$ , for the Beaver and President datasets. The upward trend in each panel shows the dependence. Time series analysts typically use the term *autocorrelation* — the prefix *auto* refers to the fact that the time series is correlated with itself — even though they mean dependence. R has the built-in function `acf` for computing autocorrelations. The following snippet shows how it works.

```
> acf (beaver1$temp, plot=F, lag.max=5)
```

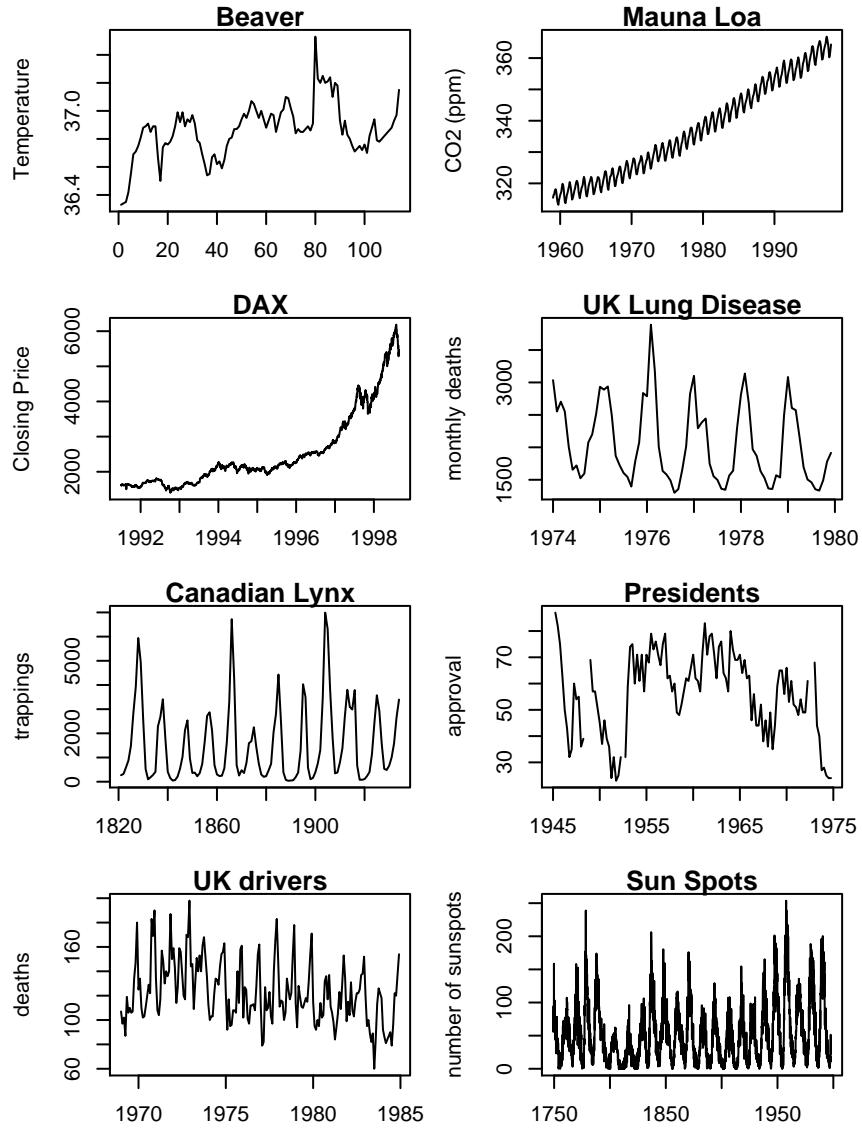


Figure 7.5: Time series. **Beaver:** Body temperature of a beaver, recorded every 10 minutes; **Mauna Loa:** Atmospheric concentration of CO<sub>2</sub>; **DAX:** Daily closing prices of the DAX stock exchange in Germany; **UK Lung Disease:** monthly deaths from bronchitis, emphysema and asthma; **Canadian Lynx:** annual number of trappings; **Presidents:** quarterly approval ratings; **UK drivers:** deaths of car drivers; **Sun Spots:** monthly sunspot numbers. In all cases the abscissa is time.

Autocorrelations of series 'beaver1\$temp', by lag

|       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|
| 0     | 1     | 2     | 3     | 4     | 5     |
| 1.000 | 0.826 | 0.686 | 0.580 | 0.458 | 0.342 |

The six numbers in the bottom line are  $\text{Cor}(Y_t, Y_t)$ ,  $\text{Cor}(Y_t, Y_{t+1})$ , ...,  $\text{Cor}(Y_t, Y_{t+5})$  and are referred to as autocorrelations of lag 0, lag 1, ..., lag 5. Those autocorrelations can, as usual, be visualized with plots as in Figure 7.7.

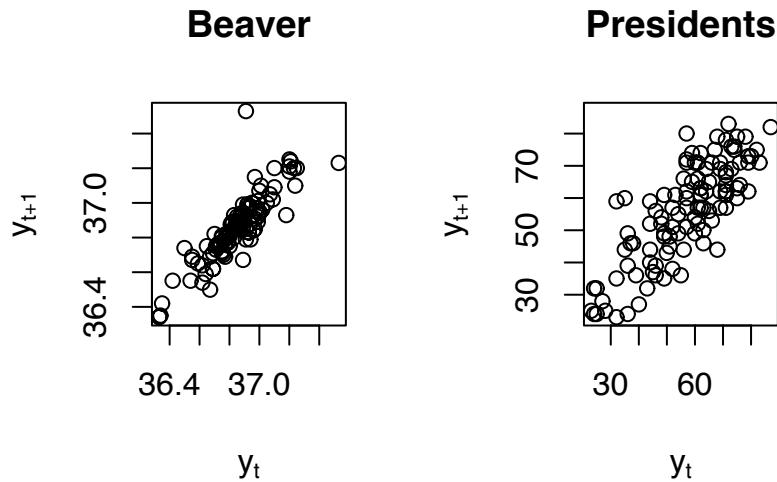


Figure 7.6:  $Y_{t+1}$  plotted against  $Y_t$  for the Beaver and Presidents data sets

Figure 7.6 was produced by the following snippet.

```
dim (beaver1)
plot (beaver1$temp[-114], beaver1$temp[-1], main="Beaver",
 xlab=expression(y[t]), ylab=expression(y[t+1]))
length (presidents)
plot (presidents[-120], presidents[-1], main="Presidents",
 xlab=expression(y[t]), ylab=expression(y[t+1]))
```

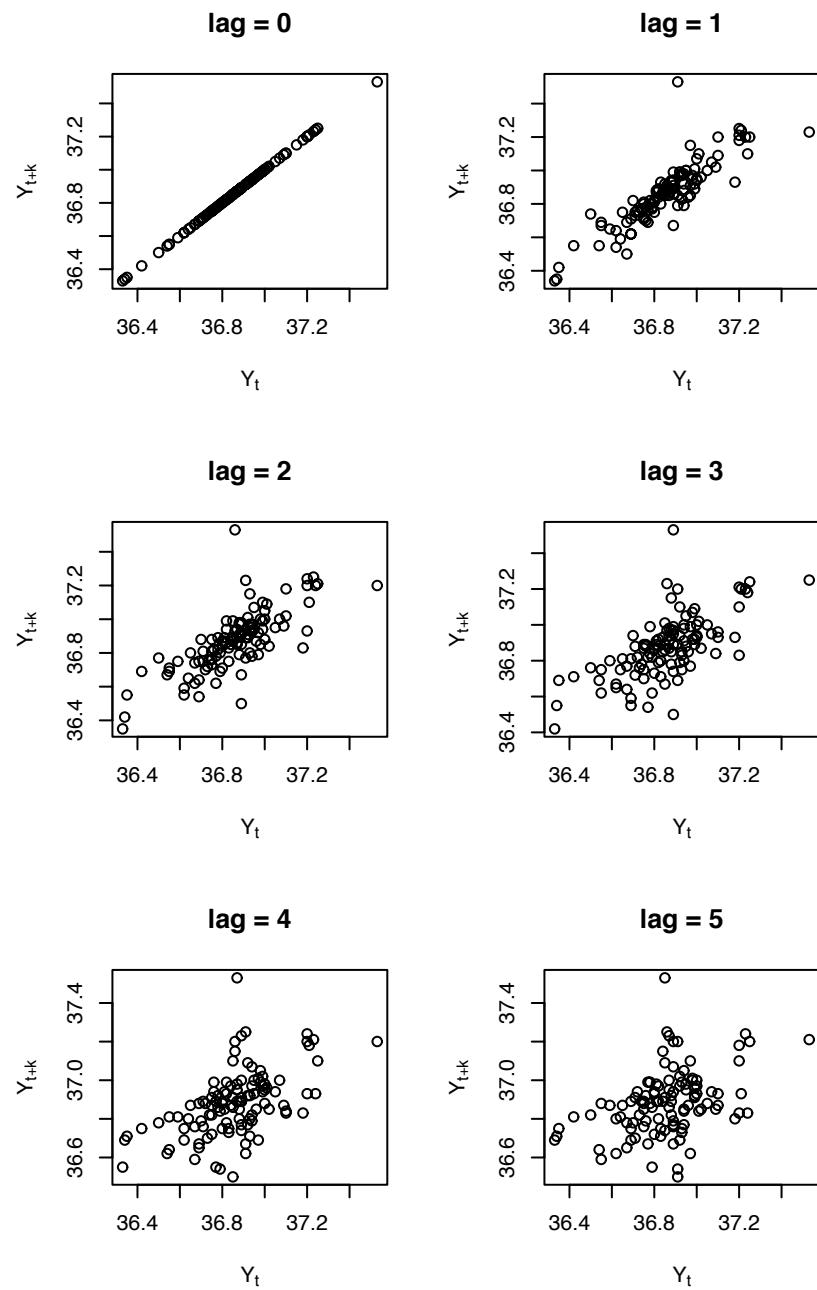


Figure 7.7:  $Y_{t+k}$  plotted against  $Y_t$  for the Beaver data set and lags  $k = 0, \dots, 5$

Figure 7.7 was produced by the following snippet.

```
par (mfrow=c(3,2))
temp <- beaver1$temp
n <- length(temp)
for (k in 0:5) {
 x <- temp[1:(n-k)]
 y <- temp[(1+k):n]
 plot (x, y, xlab=expression(Y[t]),
 ylab=expression(Y[t+k]), main=paste("lag =", k))
}
```

Because time series data cannot usually be treated as independent, we need special methods to deal with them. It is beyond the scope of this book to present the major theoretical developments of time series methods. As Figure 7.5 shows, there can be a wide variety of structure in time series data. In particular, the Beaver, and Presidents data sets have no structure readily apparent to the eye; DAX has seemingly minor fluctuations imposed on a general increasing trend; UK Lung Disease and UK drivers have an annual cycle; Mauna Loa has an annual cycle imposed on a general increasing trend; and Canadian Lynx and Sun Spots are cyclic, but for no obvious reason and with no obvious length of the cycle. In the remainder of this section we will show, by analyzing some of the data sets in Figure 7.5, some of the possibilities.

**Beaver** Our goal is to develop a more complete picture of the probabilistic structure of the  $\{Y_t\}$ 's. To that end, consider the following question. If we're trying to predict  $Y_{t+1}$ , and if we already know  $Y_t$ , does it help us also to know  $Y_{t-1}$ ? I.e., are  $Y_{t-1}$  and  $Y_{t+1}$  conditionally independent given  $Y_t$ ? That question can be answered visually with a coplot (Figures 2.15 and 2.16). Figure 7.8 shows the coplot for the Beaver data.

Figure 7.8 was produced by the following snippet.

```
temp <- beaver1$temp
n <- length (temp)
coplot (temp[3:n] ~ temp[1:(n-2)] | temp[2:(n-1)],
 xlab=c (expression(Y[t-1]), expression(Y[t])),
 ylab=expression(Y[t+1]))
```

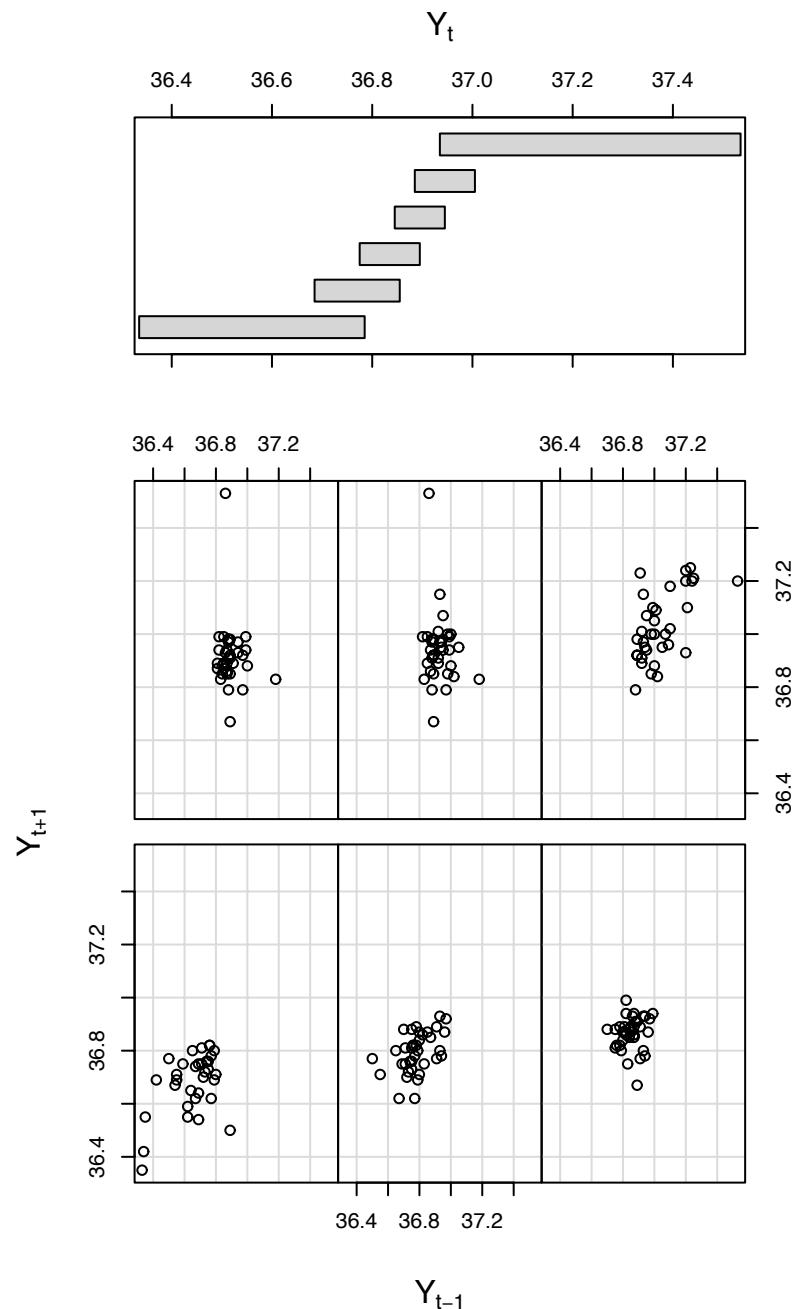


Figure 7.8: coplot of  $Y_{t+1}$  as a function of  $Y_{t-1}$  given  $Y_t$  for the Beaver data set

The figure is ambiguous. In the first, second, and sixth panels,  $Y_{t+1}$  and  $Y_{t-1}$  seem to be linearly related given  $Y_t$ , while in the third, fourth, and fifth panels,  $Y_{t+1}$  and  $Y_{t-1}$  seem to be independent given  $Y_t$ . We can examine the question numerically with the *partial autocorrelation*, the conditional correlation of  $Y_{t+1}$  and  $Y_{t-1}$  given  $Y_t$ . The following snippet shows how to compute partial autocorrelations in R using the function `pacf`.

```
> pacf (temp, lag.max=5, plot=F)
Partial autocorrelations of series 'temp', by lag

 1 2 3 4 5
0.826 0.014 0.031 -0.101 -0.063
```

The numbers in the bottom row are  $\text{Cor}(Y_t, Y_{t+k} | Y_{t+1}, \dots, Y_{t+k-1})$ . Except for the first, they're small. Figure 7.8 and the partial autocorrelations suggest that a model in which  $Y_{t+1} \perp Y_{t-1} | Y_t$  would fit the data well. And the second panel in Figure 7.7 suggests that a model of the form  $Y_{t+1} = \beta_0 + \beta_1 Y_t + \epsilon_{t+1}$  might fit well. Such a model is called an *autoregression*. R has a function `ar` for fitting them. Here's how it works with the Beaver data.

```
> fit <- ar (beaver1$temp, order.max=1)
> fit # see what we've got

Call:
ar(x = beaver1$temp, order.max = 1)

Coefficients:
1
0.8258

Order selected 1 sigma^2 estimated as 0.01201
```

The 0.8258 means that the fitted model is  $Y_{t+1} = \beta_0 + 0.8258Y_t + \epsilon_{t+1}$ . You can see whether the 0.8258 makes sense by examining the second panel of Figure 7.7. The  $\epsilon_t$ 's have an estimated variance of 0.012. `fit$x.mean` shows that  $\hat{\beta}_0 = 36.86$ . Finally, `qqnorm(fit$resid)` (Try it.) shows a nearly linear plot, except for one point, indicating that  $Y_{t+1} \sim N(36.86 + .8258Y_t, \sqrt{0.012})$  is a reasonably good model, except for one outlier.

**Mauna Loa** The Mauna Loa data look like an annual cycle superimposed on a steadily increasing long term trend. Our goal is to estimate both components and decompose the data as

$$Y_t = \text{long term trend} + \text{annual cycle} + \text{unexplained variation}.$$

Our strategy, because it seems easiest, is to estimate the long term trend first, then use deviations from the long term trend to estimate the annual cycle. A sensible estimate of the long term trend at time  $t$  is the average of a year's CO<sub>2</sub> readings, for a year centered at  $t$ . Thus, let

$$\hat{g}(t) = \frac{.5y_{t-6} + y_{t-5} + \dots + y_{t+5} + .5y_{t+6}}{12} \quad (7.6)$$

where  $g(t)$  represents the long term trend at time  $t$ . R has the built-in command `filter` to compute  $\hat{g}$ . The result is shown in Figure 7.9 (a) which also shows how to use `filter`. Deviations from  $\hat{g}$  are `co2 - g.hat`. See Figure 7.9 (b). The deviations can be grouped by month, then averaged. The average of the January deviations, for example, is a good estimate of how much the January CO<sub>2</sub> deviates from the long term trend, and likewise for other months. See Figure 7.9 (c). Finally, Figure 7.9 (d) shows the data,  $\hat{g}$ , and the fitted values  $\hat{g} + \text{monthly effects}$ . The fit is good: the fitted values differ very little from the data.

Figure 7.9 was produced by the following snippet.

```
filt <- c (.5, rep(1,11), .5) / 12
g.hat <- filter (co2, filt)
par (mfrow=c(2,2))
plot.ts (co2, main="(a)")
lines (g.hat)
resids <- co2 - g.hat
plot.ts (resids, main="(b)")
resids <- matrix (resids, nrow=12)
cycle <- apply (resids, 1, mean, na.rm=T)
plot (cycle, type="b", main="(c)")
plot.ts (co2, type="p", pch=". ", main="(d)")
lines (g.hat)
lines (g.hat + cycle)
```

**DAX**  $Y_t$  is the closing price of the German stock exchange DAX on day  $t$ . Investors often care about the rate of return  $Y_t^* = Y_{t+1}/Y_t$ , so we'll have to consider whether to

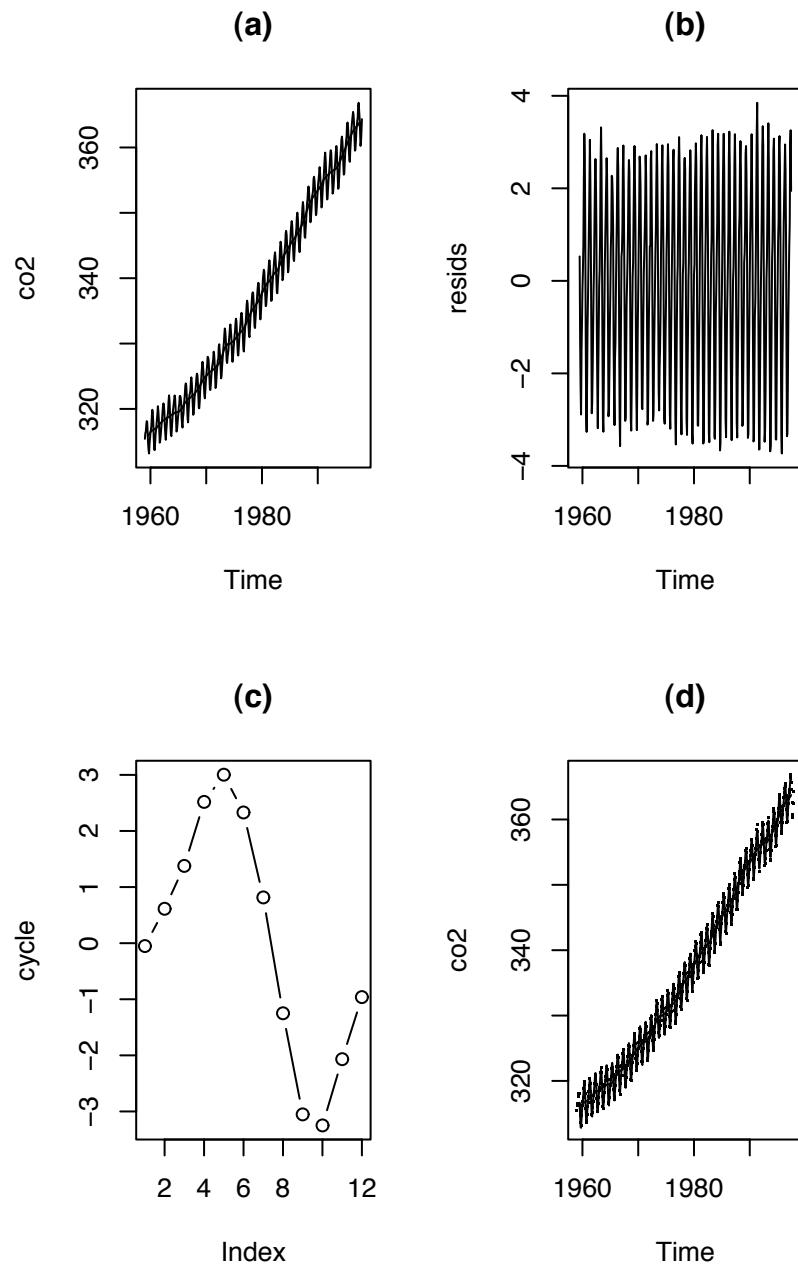


Figure 7.9: (a): CO<sub>2</sub> and  $\hat{g}$ ; (b): residuals; (c): residuals averaged by month; (d): data,  $\hat{g}$ , and fitted values

analyze the  $Y_t$ 's directly or convert them to  $Y_t^*$ 's first. Figure 7.10 is for the DAX prices directly. Panel **(a)** shows the  $Y_t$ 's. It seems to show minor fluctuations around a steadily increasing trend. Panel **(b)** shows the time series of  $Y_t - Y_{t-1}$ . It seems to show a series of fluctuations approximately centered around 0, with no apparent pattern, and with larger fluctuations occurring later in the series. Panel **(c)** shows  $Y_t$  versus  $Y_{t-1}$ . It shows a strong linear relationship between  $Y_t$  and  $Y_{t-1}$ . Two lines are drawn on the plot: the lines  $Y_t = \beta_0 + \beta_1 Y_{t-1}$  for  $(\beta_0, \beta_1) = (0, 1)$  and for  $(\beta_0, \beta_1)$  set equal to the ordinary regression coefficients found by `lm`. The two lines are indistinguishable, suggesting that  $Y_t \approx Y_{t-1}$  is a good model for the data. Panel **(d)** is a Q-Q plot of  $Y_t - Y_{t-1}$ . It is not approximately linear, suggesting that  $Y_t \sim N(Y_{t-1}, \sigma)$  is not a good model for the data.

Figure 7.10 was produced by the following snippet.

```
par (mfrow=c(2,2))
plot.ts (DAX, main="(a)")
plot.ts (diff(DAX), ylab = expression (DAX[t] - DAX[t-1]),
 main="(b)")
plot (DAX[-n], DAX[-1], xlab = expression (DAX[t-1]),
 ylab = expression (DAX[t]), main="(c)", pch=".")
abline (0, 1)
abline (lm (DAX[-1] ~ DAX[-n])$coef, lty=2)
qqnorm (diff(DAX), main="(d)", pch=".")
```

- The R command `diff` is for taking differences, typically of time series. `diff(y)` yields  $y[2]-y[1]$ ,  $y[3]-y[2]$ , ... which could also be accomplished easily enough without using `diff`:  $y[-1] - y[-n]$ . But additional arguments, as in `diff ( y, lag, differences )` make it much more useful. For example, `diff(y,lag=2)` yields  $y[3]-y[1]$ ,  $y[4]-y[2]$ , ... while `diff ( y, differences=2 )` is the same as `diff ( diff(y) )`. The latter is a construct very useful in time series analysis.

Figure 7.11 is for the  $Y_t^*$ 's. Panel **(a)** shows the time series. It shows a seemingly patternless set of data centered around 1. Panel **(b)** shows the time series of  $Y_t^* - Y_{t-1}^*$ , a seemingly patternless set of data centered at 0. Panel **(c)** shows  $Y_t^*$  versus  $Y_{t-1}^*$ . It shows no apparent relationship between  $Y_t^*$  and  $Y_{t-1}^*$ , suggesting that  $Y_t^* \perp Y_{t-1}^*$  is a good model for the data. Panel **(d)** is a Q-Q plot of  $Y_t^*$ . It is approximately linear, suggesting that  $Y_t^* \sim N(\mu, \sigma)$  is a good model for the data, with a few outliers on both the high and low ends. The mean and SD of the  $Y_t^*$ 's are about 1.000705 and 0.01028; so  $Y_t^* \sim N(1.0007, 0.01)$

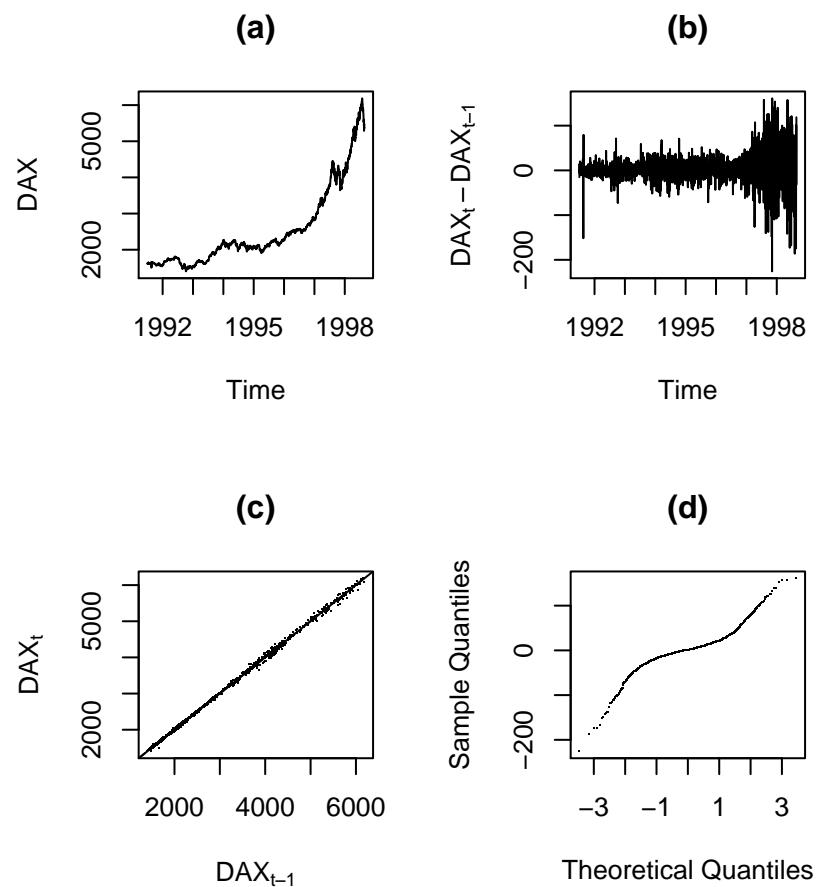


Figure 7.10: DAX closing prices. (a): the time series of  $Y_t$ 's; (b):  $Y_t - Y_{t-1}$ ; (c):  $Y_t$  versus  $Y_{t-1}$ ; (d): QQ plot of  $Y_t - Y_{t-1}$ .

might be a good model.

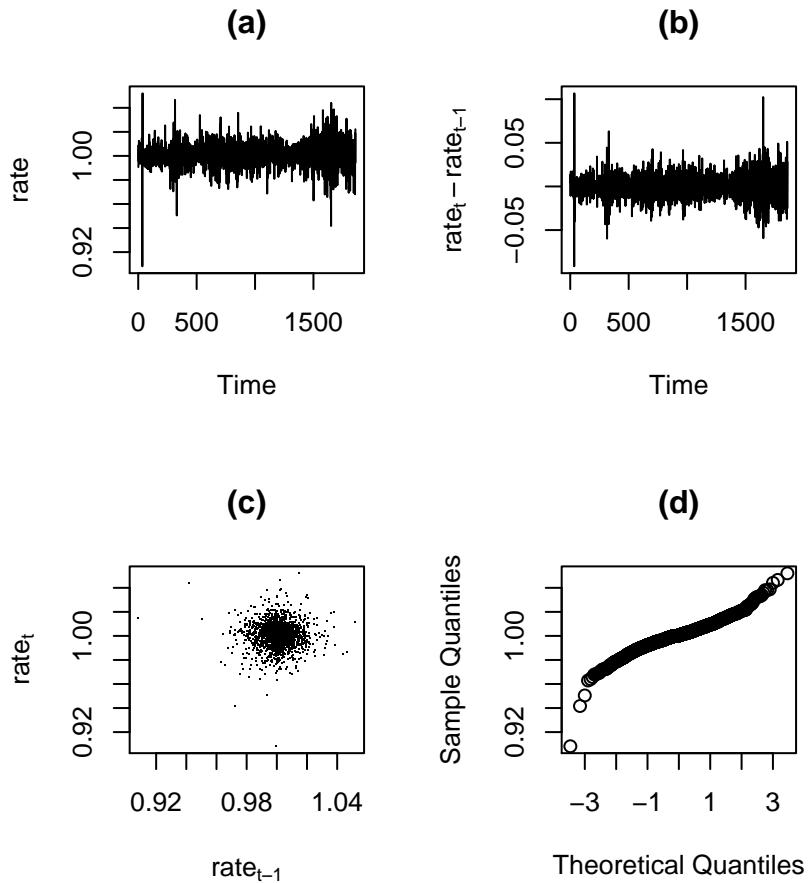


Figure 7.11: DAX returns. (a): the time series of  $Y_t^*$ 's; (b):  $Y_t^* - Y_{t-1}^*$ ; (c):  $Y_t^*$  versus  $Y_{t-1}^*$ ; (d): QQ plot of  $Y_t^*$ .

Figure 7.11 was produced by the following snippet.

```
par (mfrow=c(2,2))
plot.ts (rate, main="(a)")
plot.ts (diff(rate), ylab = expression (rate[t] - rate[t-1]),
 main="(b)")
```

```
plot (rate[-n2], rate[-1], xlab = expression (rate[t-1]),
 ylab = expression (rate[t]), main="(c)", pch="."
 qqnorm (rate, main="(d)")
```

We now have two possible models for the DAX data:  $Y_t \approx Y_{t-1}$  with a still to be determined distribution and  $Y^* \sim N(1.0007, 0.01)$  with the  $Y_t^*$ 's mutually independent. Both seem plausible on statistical grounds. (But see Exercise 8 for further development.) It is not necessary to choose one or the other. Having several ways of describing a data set is useful. Each model gives us another way to view the data. Economists and investors might prefer one or the other at different times or for different purposes. There might even be other useful models that we haven't yet considered. Those would be beyond the scope of this book, but could be covered in texts on time series, financial mathematics, econometrics, or similar topics.

## 7.3 Survival analysis

In many studies, the random variable is the time at which an event occurs. For example,

**medicine** The time until a patient dies.

**neurobiology** The time until a neuron fires.

**quality control** The time until a computer crashes.

**higher education** The time until an Associated Professor is promoted to Full Professor.

Such data are called *survival data*. For the  $i$ 'th person, neuron, computer, etc., there is a random variable

$$y_i = \text{time of event on } i\text{'th unit.}$$

We usually call  $y_i$  the *lifetime*, even though the event is not necessarily death. It is often the case with survival data that some measurements are *censored*. For example, if we study a university's records to see how long it takes to get promoted from Associate to Full Professor, we will find some Associate Professors leave the university — either through retirement or by taking another job — before they get promoted while others are still Associate Professors at the time of our study. For these people we don't know their time of promotion. If either (a) person  $i$  left the university after five years, or (b) person  $i$  became Associate Professor five years prior to our study, then we don't know  $y_i$  exactly. All we know is  $y_i > 5$ . This form of censoring is called *right censoring*. In some data sets

there may also be *left censoring* or *interval censoring*. Survival analysis typically requires specialized statistical techniques. R has a package of functions for this purpose; the name of the package is **survival**. The **survival** package is automatically distributed with R. To load it into your R session, type `library(survival)`. The package comes with functions for survival analysis and also with some example data sets. Our next example uses one of those data sets.

### **Example 7.2** (Bladder Tumors)

This example comes from a study of bladder tumors, originally published in BYAR [1980] and later reanalyzed in WEI ET AL. [1989]. Patients had bladder tumors. The tumors were removed and the patients were randomly assigned to one of three treatment groups (placebo, thiotepa, pyridoxine). Then the patients were followed through time to see whether and when bladder tumors would recur. R's **survival** package has the data for the first two treatment groups, placebo and thiotepa. Type `bladder` to see it. (Remember to load the **survival** package first.) The last several lines look like this.

|     | <code>id</code> | <code>rx</code> | <code>number</code> | <code>size</code> | <code>stop</code> | <code>event</code> | <code>enum</code> |
|-----|-----------------|-----------------|---------------------|-------------------|-------------------|--------------------|-------------------|
| 341 | 83              | 2               | 3                   | 4                 | 54                | 0                  | 1                 |
| 342 | 83              | 2               | 3                   | 4                 | 54                | 0                  | 2                 |
| 343 | 83              | 2               | 3                   | 4                 | 54                | 0                  | 3                 |
| 344 | 83              | 2               | 3                   | 4                 | 54                | 0                  | 4                 |
| 345 | 84              | 2               | 2                   | 1                 | 38                | 1                  | 1                 |
| 346 | 84              | 2               | 2                   | 1                 | 54                | 0                  | 2                 |
| 347 | 84              | 2               | 2                   | 1                 | 54                | 0                  | 3                 |
| 348 | 84              | 2               | 2                   | 1                 | 54                | 0                  | 4                 |
| 349 | 85              | 2               | 1                   | 3                 | 59                | 0                  | 1                 |
| 350 | 85              | 2               | 1                   | 3                 | 59                | 0                  | 2                 |
| 351 | 85              | 2               | 1                   | 3                 | 59                | 0                  | 3                 |
| 352 | 85              | 2               | 1                   | 3                 | 59                | 0                  | 4                 |

- `id` is the patient's id number. Note that each patient has four lines of data. That's to record up to four recurrences of tumor.
- `rx` is the treatment: 1 for placebo; 2 for thiotepa.
- `number` is the number of tumors the patient had at the initial exam when the patient joined the study.
- `size` is the size (cm) of the largest initial tumor.

- `stop` is the time (months) of the observation.
- `event` is 1 if there's a tumor; 0 if not.
- `enum` line 1, 2, 3, or 4 for each patient

For example, patient 83 was followed for 54 months and had no tumor recurrences; patient 85 was followed for 59 months and also had no recurrences. But patient 84, who was also followed for 54 months, had a tumor recurrence at month 38 and no further recurrences after that. Our analysis will look at the time until the first recurrence, so we want `bladder[bladder$enum==1, ]`, the last several lines of which are

|     | <code>id</code> | <code>rx</code> | <code>number</code> | <code>size</code> | <code>stop</code> | <code>event</code> | <code>enum</code> |
|-----|-----------------|-----------------|---------------------|-------------------|-------------------|--------------------|-------------------|
| 329 | 80              | 2               | 3                   | 3                 | 49                | 0                  | 1                 |
| 333 | 81              | 2               | 1                   | 1                 | 50                | 0                  | 1                 |
| 337 | 82              | 2               | 4                   | 1                 | 4                 | 1                  | 1                 |
| 341 | 83              | 2               | 3                   | 4                 | 54                | 0                  | 1                 |
| 345 | 84              | 2               | 2                   | 1                 | 38                | 1                  | 1                 |
| 349 | 85              | 2               | 1                   | 3                 | 59                | 0                  | 1                 |

Patients 80, 81, 83, and 85 had no tumors for as long as they were followed; their data is right-censored. The data for patients 82 and 84 is not censored; it is observed exactly.

Figure 7.12 is a plot of the data. The solid line is for placebo; the dashed line for thiotepa. The abscissa is in months. The ordinate shows the fraction of patients who have survived without a recurrence of bladder tumors. The plot shows, for example, that at 30 months, the survival rate without recurrence is about 50% for thiotepa patients compared to a little under 40% for placebo patients. The circles on the plot show censoring. I.e., the four circles on the solid curve between 30 and 40 months represent four placebo patients whose data was right-censored. There is a circle at every censoring time that is not also the time of a recurrence (for a different patient).

Figure 7.12 was produced with the snippet

```
event.first <- bladder[, "enum"] == 1
blad.surv <- Surv (bladder[event.first, "stop",],
 bladder[event.first, "event"])
blad.fit <- survfit (blad.surv ~ bladder[event.first, "rx"])
plot (blad.fit, conf.int=FALSE, mark=1, xlab="months",
```

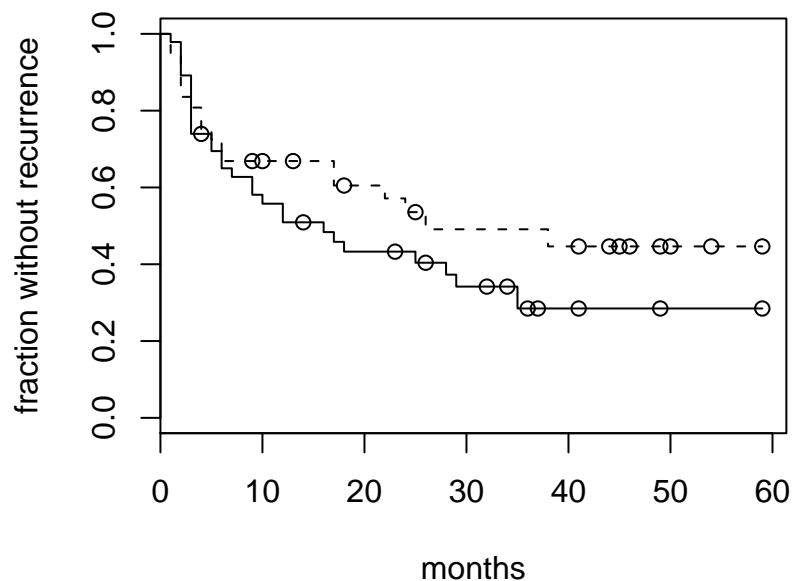


Figure 7.12: Survival curve for bladder cancer. Solid line for placebo; dashed line for thiotepa.

```
ylab="fraction without recurrence", lty=1:2)
```

- `Surv` is R's function for creating a survival object. You can type `print(blad.surv)` and `summary(blad.surv)` to learn more about survival objects.
- `survfit` computes an estimate of a survival curve.

In survival analysis we think of  $y_1, \dots, y_n$  as a sample from a distribution  $F$ , with density  $f$ , of lifetimes. In survival analysis, statisticians often work with the *survivor function*  $S(t) = 1 - F(t) = P[y_i > t]$ , the probability that a unit survives beyond time  $t$ . The lines in Figure 7.12 are the so-called *Kaplan-Meier* estimates of  $S$  for patients in the thiotepa and placebo groups, which arise from the following argument. Partition  $\mathbb{R}^+$  into intervals  $(0, t_1], (t_1, t_2], \dots$ , and let  $p_i = P[y > t_i | y > t_{i-1}]$ . Then for each  $i$ ,  $S(t_i) = \prod_{j=1}^i p_i$ . The  $p_i$ 's can be estimated from data as  $\hat{p}_i = (r(t_i) - d_i)/r(t_i)$  where  $r(t_i)$  is the number of people at risk (in the study but not yet dead) at time  $t_{i-1}$  and  $d_i$  is the number of deaths in the interval  $(t_{i-1}, t_i]$ . Thus,  $\hat{S}(t_i) = \prod_{j=1}^i (r(t_i) - d_i)/r(t_i)$ . As the partition becomes finer, most terms in the product are equal to one; only those intervals with a death contribute a term that is not one. The limit yields the Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{i:y_i < t} \frac{r(y_i) - d_i}{r(y_i)}.$$

This estimate is reasonably accurate so long as  $r(t)$  is reasonably large;  $\hat{S}(t)$  is more accurate for small values of  $t$  than for large values of  $t$ ; and there is no information at all for estimating  $S(t)$  for  $t > \max\{y_i\}$ .

Survival data is often modelled in terms of the hazard function

$$h(t) = \lim_{h \rightarrow 0} \frac{P[y \in [t, t+h] | y \geq t]}{h} = \lim_{h \rightarrow 0} \frac{P[y \in [t, t+h]]}{h P[y \geq t]} = \frac{f(t)}{S(t)}. \quad (7.7)$$

The interpretation of  $h(t)$  is the fraction, among people who have survived to time  $t$ , of those who will die soon thereafter. There are several parametric families of distributions for lifetimes in use for survival analysis. The most basic is the exponential —  $f(y) = (1/\lambda) \exp^{-y/\lambda}$  — which has hazard function  $h(y) = 1/\lambda$ , a constant. A constant hazard function says, for example, that young people are just as likely to die as old people, or that new air conditioners are just as likely to fail as old air conditioners. For many applications that assumption is unreasonable, so statisticians may work with other parametric families

for lifetimes, especially the Weibull, which has  $h(y) = \lambda\alpha(\lambda y)^{\alpha-1}$ , an increasing function of  $y$  if  $\alpha > 1$ . We will not dwell further on parametric models; the interested reader should refer to a more specialized source.

However, the goal of survival analysis is not usually to estimate  $S$  and  $h$ , but to compare the survivor and hazard functions for two groups such as treatment and placebo or to see how the survivor and hazard functions vary as functions of some covariates. Therefore it is not usually necessary to estimate  $S$  and  $h$  well, as long as we can estimate how  $S$  and  $h$  differ between groups, or as a function of the covariates. For this purpose it has become common to adopt a *proportional hazards* model:

$$h(y) = h_0(y) \exp(\beta' x) \quad (7.8)$$

where  $h_0$  is the baseline hazard function that is adjusted according to  $x$ , a vector of covariates and  $\beta$ , a vector of coefficients. Equation 7.8 is known as the Cox proportional hazards model. The goal is usually to estimate  $\beta$ . R's *survival* package has a function for fitting Equation 7.8 to data.

### Example 7.3 (Bladder Tumors, cont.)

This continues Example 7.2. Here we adopt the Cox proportional hazards model and see how well we can estimate the effect of the treatment thiotepa compared to placebo in preventing recurrence of bladder tumors. We will also examine the effects of other potential covariates.

We'd like to fit the Cox proportional hazards model  $h(y) = h_0(y) \exp(\beta_{\text{trt}} \cdot \text{trt})$  where  $\text{trt}$  is an indicator variable that is 1 for patients on thiotepa and 0 for patients on placebo; but first we check whether such a model looks plausible; i.e. whether the hazards look proportional. Starting from Equation 7.7 we can integrate both sides to get  $H(y) \equiv \int_0^y h(z) dz = -\log S(y)$ .  $H(y)$  is called the *cumulative hazard* function. Thus, if the two groups have proportional hazards, they also have proportional cumulative hazard functions and log survivor functions. Figure 7.13 plots the estimated cumulative hazard and log (cumulative hazard) functions for the bladder tumor data. The log (cumulative hazard) functions look parallel, so the proportional hazards assumption looks reasonable.

Figure 7.13 was produced with the snippet

```
plot (blad.fit, conf.int=FALSE, mark=1, xlab="months",
 ylab="cumulative hazard", lty=1:2, fun="cumhaz")
plot (blad.fit, conf.int=FALSE, mark=1, xlab="months",
 ylab="log(cumulative hazard)", lty=1:2, fun="cloglog")
```

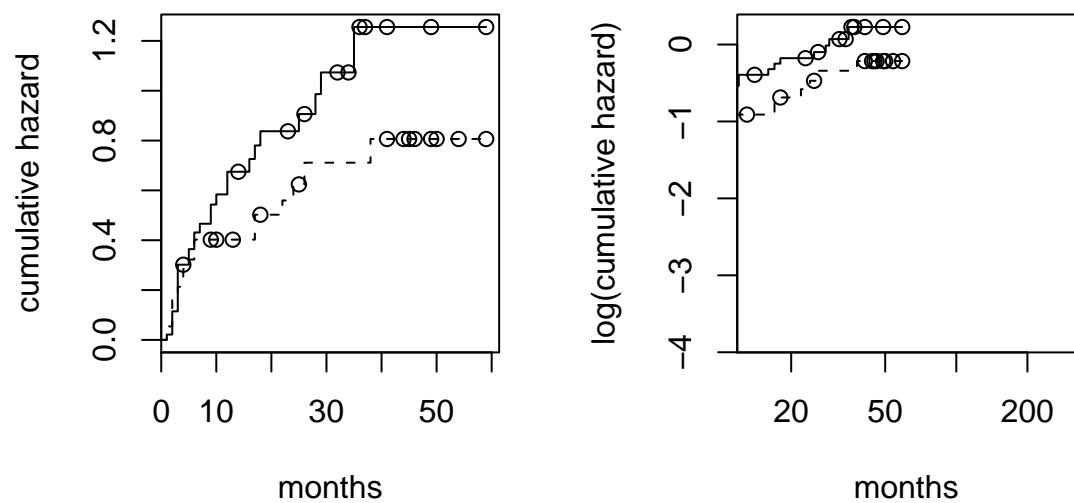


Figure 7.13: Cumulative hazard and  $\log(\text{hazard})$  curves for bladder cancer. Solid line for thiotepa; dashed line for placebo.

- The `fun` argument allows transformations of the survival curve. `fun="cumhaz"` plots the cumulative hazard function and `fun="cloglog"` plots the log (cumulative hazard) function.

Since the proportional hazards model looks reasonable, we fit it:

```
blad.cox <- coxph (blad.surv ~ bladder[event.first,"rx"]).
Printing blad.cox yields
```

Call:

```
coxph(formula = blad.surv ~ bladder[event.first, "rx"])
```

|                            | coef   | exp(coef) | se(coef) | z     | p    |
|----------------------------|--------|-----------|----------|-------|------|
| bladder[event.first, "rx"] | -0.371 | 0.69      | 0.303    | -1.22 | 0.22 |

```
Likelihood ratio test=1.54 on 1 df, p=0.215 n= 85
```

The estimated coefficient is  $\hat{\beta}_{\text{trt}} = -0.371$ . Thus the hazard function for thiotepa patients is estimated to be  $\exp(-0.371) = 0.69$  times that for placebo patients. The standard error of  $\hat{\beta}_{\text{trt}} = -0.371$  is about 0.3; so  $\hat{\beta}_{\text{trt}}$  is accurate to about  $\pm 0.6$  or so.

## 7.4 Exercises

- In the `orthodont` data, M09 doesn't follow the general pattern. Maybe there's an error in the data or maybe that person really did grow in an unusual way. Fit a model similar to `ortho.fit1` but excluding the data from M09. Does anything change in an important way?
- In the `orthodont` data, it seems clear that different individuals have different intercepts. It's not as clear whether they have different slopes. Conduct an analysis to investigate that possibility. Your analysis must address at least two questions. First, do different individuals seem to have different slopes? Second, is the apparent difference in Male and Female slopes found by `ortho.fit4` really a difference between males and females on average, or is it due to the particular males and females who happened to be chosen for this study?
- Carry out a Bayesian analysis of the model in Equation 7.2.

4. (a) Make plots analogous to Figures 7.6 and 7.7, compute autocorrelations, and interpret for the other datasets in Figure 7.5.
- (b) Make plots analogous to Figure 7.8, compute partial autocorrelations, and interpret for the other data sets in Figure 7.5.
5. Create and fit a good model for body temperatures of the second beaver. Use the dataset `beaver2`.
6. (a) Why does Equation 7.6 average over a year? Why isn't it, for example,

$$\hat{g}(t) = \frac{.5y_{t-k} + y_{t-k+1} + \cdots + y_{t+k-1} + .5y_{t+k}}{2k}$$

for some  $k \neq 6$ ?

- (b) Examine  $\hat{g}$  in Equation 7.6. Use R if necessary. Why are some of the entries `NA`?
7. The R code for Figure 7.9 contains the lines

```
resids <- matrix (resids, nrow=12)
cycle <- apply (resids, 1, mean, na.rm=T)
```

Would the following lines work instead?

```
resids <- matrix (resids, ncol=12)
cycle <- apply (resids, 2, mean, na.rm=T)
```

Why or why not?

8. Figure 7.10 and the accompanying text suggest that  $Y_t \approx Y_{t-1}$  is a good model for the DAX data. But that doesn't square with the observation that the  $Y_t$ 's have a generally increasing trend.
  - (a) Find a quantitative way to show the trend.
  - (b) Say why the DAX analysis missed the trend.
  - (c) Improve the analysis so it's consistent with the trend.

9. Figures 7.10 and 7.11 and the accompanying text analyze the DAX time series as though it has the same structure throughout the entire time. Does that make sense? Think of and implement some way of investigating whether the structure of the series changes from early to late.
10. Choose one or more of the other EU Stock Markets that come with the DAX data. Investigate whether it has the same structure as the DAX.
11.
  - (a) Make a plausible analysis of the UK Lung Disease data.
  - (b) R has the three data sets `ldeaths`, `fdeaths`, and `mdeaths` which are the total deaths, the deaths of females, and the deaths of males. Do the deaths of females and males follow similar distributional patterns? Justify your answer.
12. Make a plausible analysis of the Presidents approval ratings.
13.
  - (a) Make a plausible analysis of the UK drivers deaths.
  - (b) According to R, “Compulsory wearing of seat belts was introduced on 31 Jan 1983.” Did that effect the number of deaths? Justify your answer.
  - (c) Is the number of deaths related to the number of kilometers driven? (Use the variable `kms` in the `Seatbelts` data set.) Justify your answer.
14. This question follows up Example 7.3. In the example we analyzed the data to learn the effect of thioteipa on the recurrence of bladder tumors. But the data set has two other variables that might be important covariates: the number of initial tumors and the size of the largest initial tumor.
  - (a) Find the distribution of the numbers of initial tumors. How many patients had 1 initial tumor, how many had 2, etc?
  - (b) Divide patients, in a sensible way, into groups according to the number of initial tumors. You must decide how many groups there should be and what the group boundaries should be.
  - (c) Make plots similar to Figures 7.12 and 7.13 to see whether a proportional hazard model looks sensible for number of initial tumors.
  - (d) Fit a proportional hazard model and report the results.
  - (e) Repeat the previous analysis, but for size of largest initial tumor.
  - (f) Fit a proportional hazard model with three covariates: treatment, number of initial tumors, size of largest initial tumor. Report the results.

## CHAPTER 8

# MATHEMATICAL STATISTICS

## 8.1 Properties of Statistics

### 8.1.1 Sufficiency

Consider the following two facts.

1. Let  $Y_1, \dots, Y_n \sim \text{i.i.d. Poi}(\lambda)$ . Chapter 2, Exercise 7 showed that  $\ell(\lambda)$  depends only on  $\sum Y_i$  and not on the specific values of the individual  $Y_i$ 's.
2. Let  $Y_1, \dots, Y_n \sim \text{i.i.d. Exp}(\lambda)$ . Chapter 2, Exercise 20 showed that  $\ell(\lambda)$  depends only on  $\sum Y_i$  and not on the specific values of the individual  $Y_i$ 's.

Further, since  $\ell(\lambda)$  quantifies how strongly the data support each value of  $\lambda$ , other aspects of  $\mathbf{y}$  are irrelevant. Thus, for inference about  $\lambda$  it suffices to know  $\ell(\lambda)$ , and therefore, for Poisson and Exponential data, it suffices to know  $\sum Y_i$ . We don't need to know the individual  $Y_i$ 's. We say that  $\sum Y_i$  is a sufficient statistic for  $\lambda$ .

Section 8.1.1 examines the general concept of sufficiency. We work in the context of a parametric family. The idea of sufficiency is formalized in Definition 8.1.

**Definition 8.1.** Let  $\{p(\cdot | \theta)\}$  be a family of probability densities indexed by a parameter  $\theta$ . Let  $\mathbf{y} = (y_1, \dots, y_n)$  be a sample from  $p(\cdot | \theta)$  for some unknown  $\theta$ . Let  $T(\mathbf{y})$  be a statistic such that the joint distribution factors as

$$\prod p(y_i | \theta) = g(T(\mathbf{y}), \theta)h(\mathbf{y}).$$

for some functions  $g$  and  $h$ . Then  $T$  is called a *sufficient statistic for  $\theta$* .

The idea is that once the data have been observed,  $h(\mathbf{y})$  is a constant that does not depend of  $\theta$ , so  $\ell(\theta) \propto \prod p(y_i | \theta) = g(T, \theta)h(\mathbf{y}) \propto g(T, \theta)$ . Therefore, in order to know the likelihood function and make inference about  $\theta$ , we need only know  $T(\mathbf{y})$ , not anything else about  $\mathbf{y}$ . For our Poisson and Exponential examples we can take  $T(\mathbf{y}) = \sum y_i$ .

For a more detailed look at sufficiency, think of generating three  $\text{Bern}(\theta)$  trials  $\mathbf{y} \equiv (y_1, y_2, y_3)$ .  $\mathbf{y}$  can be generated, obviously, by generating  $y_1, y_2, y_3$  sequentially. The possible outcomes and their probabilities are

$$\begin{aligned} (0, 0, 0) & \quad (1 - \theta)^3 \\ (1, 0, 0) \\ (0, 1, 0) & \quad \theta(1 - \theta)^2 \\ (0, 0, 1) \\ (1, 1, 0) \\ (1, 0, 1) & \quad \theta^2(1 - \theta) \\ (0, 1, 1) \\ (1, 1, 1) & \quad \theta^3 \end{aligned}$$

But  $\mathbf{y}$  can also be generated by a two-step procedure:

1. Generate  $\sum y_i = 0, 1, 2, 3$  with probabilities  $(1 - \theta)^3, 3\theta(1 - \theta)^2, 3\theta^2(1 - \theta), \theta^3$ , respectively.
2. (a) If  $\sum y_i = 0$ , generate  $(0, 0, 0)$   
(b) If  $\sum y_i = 1$ , generate  $(1, 0, 0), (0, 1, 0)$ , or  $(0, 0, 1)$  each with probability  $1/3$ .  
(c) If  $\sum y_i = 2$ , generate  $(1, 1, 0), (1, 0, 1)$ , or  $(0, 1, 1)$  each with probability  $1/3$ .  
(d) If  $\sum y_i = 3$ , generate  $(1, 1, 1)$

It is easy to check that the two-step procedure generates each of the 8 possible outcomes with the same probabilities as the obvious sequential procedure. For generating  $\mathbf{y}$  the two procedures are equivalent. But in the two-step procedure, only the first step depends on  $\theta$ . So if we want to use the data to learn about  $\theta$ , we need only know the outcome of the first step. The second step is irrelevant. I.e., we need only know  $\sum y_i$ . In other words,  $\sum y_i$  is sufficient.

For an example of another type, let  $y_1, \dots, y_n \sim \text{i.i.d. } U(0, \theta)$ . What is a sufficient statistic for  $\theta$ ?

$$\begin{aligned} p(\mathbf{y} | \theta) &= \begin{cases} \frac{1}{\theta^n} & \text{if } y_i < \theta \text{ for } i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{\theta^n} \mathbf{1}_{(0,\theta)}(y_{(n)}) \end{aligned}$$

shows that  $y_{(n)}$ , the maximum of the  $y_i$ 's, is a one dimensional sufficient statistic for  $\theta$ .

### Example 8.1

In World War II, when German tanks came from the factory they had serial numbers labelled consecutively from 1. I.e., the numbers were 1, 2, .... The Allies wanted to estimate  $T$ , the total number of German tanks and had, as data, the serial numbers of captured tanks. See Exercise 23 in Chapter 5. Assume that tanks were captured independently of each other and that all tanks were equally likely to be captured. Let  $x_1, \dots, x_n$  be the serial numbers of the captured tanks. Then  $x_{(n)}$  is a sufficient statistic. Inference about the total number of German tanks should be based on  $x_{(n)}$  and not on any other aspect of the data.

If  $y$  is a random variable whose values are in a space  $\mathcal{Y}$ , then  $\mathbf{y}$  is a random variable whose values are in  $\mathcal{Y}^n$ . For any statistic  $T$  we can divide  $\mathcal{Y}^n$  into subsets indexed by  $T$ . I.e., for each value  $t$ , we define the subset

$$\mathcal{Y}_t^n = \{\mathbf{y} \in \mathcal{Y}^n : T(\mathbf{y}) = t\}$$

Then  $T$  is a sufficient statistic if and only if

$$p(\mathbf{y} | \mathbf{y} \in \mathcal{Y}_t^n)$$

does not depend on  $\theta$ .

Sometimes sufficient statistics are higher dimensional. For example, let  $y_1, \dots, y_n \sim \text{i.i.d. } \text{Gam}(\alpha, \beta)$ . Then

$$\prod p(y_i | \alpha, \beta) = \prod \frac{1}{\Gamma(\alpha)\beta^\alpha} y_i^{\alpha-1} e^{-y_i/\beta} = \left( \frac{1}{\Gamma(\alpha)\beta^\alpha} \right)^n \left( \prod y_i \right)^{\alpha-1} e^{-\sum y_i/\beta}$$

so  $T(\mathbf{y}) = (\prod y_i, \sum y_i)$  is a two dimensional sufficient statistic.

Sufficient statistics are not unique. If  $T = T(\mathbf{y})$  is a sufficient statistic, and if  $f$  is a 1-1 function, then  $f(T)$  is also sufficient. So in the Poisson, Exponential, and Bernoulli examples where  $\sum y_i$  was sufficient,  $\bar{y} = \sum y_i/n$  is also sufficient. But the lack of uniqueness is

even more severe. The whole data set  $T(\mathbf{y}) = (\mathbf{y})$  is an  $n$ -dimensional sufficient statistic because

$$\prod p(y_i | \theta) = g(T(\mathbf{y}), \theta)h(\mathbf{y})$$

where  $g(T(\mathbf{y}), \theta) = p(\mathbf{y} | \theta)$  and  $h(\mathbf{y}) = 1$ . The *order statistic*  $T(\mathbf{y}) = (y_{(1)}, \dots, y_{(n)})$  is another  $n$ -dimensional sufficient statistic. Also, if  $T$  is any sufficient one dimensional statistic then  $T_2 = (y_1, T)$  is a two dimensional sufficient statistic. But it is intuitively clear that these sufficient statistics are higher-dimensional than necessary. They can be reduced to lower dimensional statistics while retaining sufficiency, that is, without losing information.

The key idea in the preceding paragraph is that the high dimensional sufficient statistics can be transformed into the low dimensional ones, but not *vice versa*. E.g.,  $\bar{y}$  is a function of  $(y_{(1)}, \dots, y_{(n)})$  but  $(y_{(1)}, \dots, y_{(n)})$  is not a function of  $\bar{y}$ . Definition 8.2 is for statistics that have been reduced as much as possible without losing sufficiency.

**Definition 8.2.** A sufficient statistic  $T(\mathbf{y})$  is called *minimal sufficient* if, for any other sufficient statistic  $T_2$ ,  $T(\mathbf{y})$  is a function of  $T_2(\mathbf{y})$ .

This book does not delve into methods for finding minimal sufficient statistics. In most cases the user can recognize whether a statistic is minimal sufficient.

Does the theory of sufficiency imply that statisticians need look only at sufficient statistics and not at other aspects of the data? Not quite. Let  $y_1, \dots, y_n$  be binary random variables and suppose we adopt the model  $y_1, \dots, y_n \sim \text{i.i.d. Bern}(\theta)$ . Then for estimating  $\theta$  we need look only at  $\sum y_i$ . But suppose  $(y_1, \dots, y_n)$  turn out to be

$$\underbrace{0 0 \cdots 0}_{\text{many 0's}} \underbrace{1 1 \cdots 1}_{\text{many 1's}},$$

i.e., many 0's followed by many 1's. Such a dataset would cast doubt on the assumption that the  $y_i$ 's are independent. Judging from this dataset, it looks much more likely that the  $y_i$ 's come in streaks or that  $\theta$  is increasing over time. So statisticians should look at all the data, not just sufficient statistics, because looking at all the data can help us create and critique models. But once a model has been adopted, then inference should be based on sufficient statistics.

### 8.1.2 Consistency, Bias, and Mean-squared Error

**Consistency** Heuristically speaking, as we collect ever more data we should be able to learn the truth ever more accurately. This heuristic is captured formally, at least for parameter estimation, by the notion of *consistency*. To say whether an estimator is consistent

we have to define it for every sample size. To that end, let  $Y_1, Y_2, \dots \sim \text{i.i.d. } f$  for some unknown density  $f$  having finite mean  $\mu$  and SD  $\sigma$ . For each  $n \in \mathbb{N}$  let  $T_n : \mathbb{R}^n \rightarrow \mathbb{R}$ . I.e.  $T_n$  is a real-valued function of  $(y_1, \dots, y_n)$ . For example, if we're trying to estimate  $\mu$  we might take  $T_n = n^{-1} \sum_1^n y_i$ .

**Definition 8.3.** The sequence of estimators  $T_1, T_2, \dots$  is said to be *consistent for the parameter  $\theta$*  if for every  $\theta$  and for every  $\epsilon > 0$ .

$$\lim_{n \rightarrow \infty} P[|T_n - \theta| < \epsilon] = 1.$$

For example, the Law of Large Numbers, Theorem 1.12, says the sequence of sample means  $\{T_n = n^{-1} \sum_1^n y_i\}$  is consistent for  $\mu$ . Similarly, let  $S_n = n^{-1} \sum_i (y_i - T_n)^2$  be the sample variance. Then  $\{S_n\}$  is consistent for  $\sigma^2$ . More generally, m.l.e.'s are consistent.

**Theorem 8.1.** Let  $Y_1, Y_2, \dots \sim \text{i.i.d. } p_Y(y|\theta)$  and let  $\hat{\theta}_n$  be the m.l.e. from the sample  $(y_1, \dots, y_n)$ . Further, let  $g$  be a continuous function of  $\theta$ . Then, subject to regularity conditions,  $\{g(\hat{\theta}_n)\}$  is a consistent sequence of estimators for  $g(\theta)$ .

*Proof.* The proof requires regularity conditions relating to differentiability and the interchange of integral and derivative. It is beyond the scope of this book.  $\square$

Consistency is a good property; one should be wary of an inconsistent estimator. On the other hand, consistency alone does not guarantee that a sequence of estimators is optimal, or even sensible. For example, let  $R_n(y_1, \dots, y_n) = ([n/2])^{-1}(y_1 + \dots + y_{[n/2]})$ , the mean of the first half of the observations. ( $[w]$  is the *floor* of  $w$ , the largest integer not greater than  $w$ .) The sequence  $\{R_n\}$  is consistent for  $\mu$  but is not as good as the sequence of sample means.

**Bias** It seems natural to want the sampling distribution of an estimator to be centered around the parameter being estimated. This desideratum is captured formally, at least for centering in the sense of expectation, by the notion of *bias*.

**Definition 8.4.** Let  $\hat{\theta} = \hat{\theta}(y_1, \dots, y_n)$  be an estimator of a parameter  $\theta$ . The quantity  $\mathbb{E}[\hat{\theta}] - \theta$  is called the *bias* of  $\hat{\theta}$ . An estimator whose bias is 0 is called *unbiased*.

Here are some examples.

**An unbiased estimator** Let  $y_1, \dots, y_n \sim \text{i.i.d. } N(\mu, \sigma^2)$  and consider  $\hat{\mu} = \bar{y}$  as an estimate of  $\mu$ . Because  $\mathbb{E}[\bar{y}] = \mu$ ,  $\bar{y}$  is an unbiased estimate of  $\mu$ .

**A biased estimator** Let  $y_1, \dots, y_n \sim \text{i.i.d. } N(\mu, \sigma^2)$  and consider

$$\hat{\sigma}^2 = n^{-1} \sum (y_i - \bar{y})^2$$

as an estimate of  $\sigma^2$ . Theorem 5.31 says that  $\sum (y_i - \bar{y})^2 / \sigma^2 \sim \chi_{n-1}^2$  and therefore  $\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2$ . I.e.,  $\hat{\sigma}^2$  is a biased estimator of  $\sigma^2$ . Its bias is  $-\sigma^2/n$ . Some statisticians prefer to use the unbiased estimator  $\tilde{\sigma}^2 = (n-1)^{-1} \sum (y_i - \bar{y})^2$ .

**A biased estimator** Let  $x_1, \dots, x_n \sim \text{i.i.d. } U(0, \theta)$  and consider  $\hat{\theta} = x_{(n)}$  as an estimate of  $\theta$ . ( $\hat{\theta}$  is the m.l.e.; see Section 5.4.) But  $x_{(n)} < \theta$ ; therefore  $\mathbb{E}[x_{(n)}] < \theta$ ; therefore  $x_{(n)}$  is a biased estimator of  $\theta$ .

**Mean Squared Error** If  $\hat{\theta}$  is an estimate of  $\theta$ , then the *mean squared error* (MSE) of  $\hat{\theta}$  is  $\mathbb{E}[(\hat{\theta} - \theta)^2]$ . MSE is a combination of variance and bias:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + 2(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta) + (\mathbb{E}\hat{\theta} - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})^2] + (\mathbb{E}\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2. \end{aligned}$$

Whenever possible, one would like to have an estimator with minimum bias and minimum variance; those two properties would also imply minimum MSE. But it is not always possible to achieve both desiderata simultaneously. For example, let  $Y_1, \dots, Y_n \sim \text{i.i.d. } N(\mu, \sigma^2)$  and suppose we want an estimate of  $\sigma^2$ . We have looked at two estimators so far: the mle  $\hat{\sigma}^2 = n^{-1} \sum (y_i - \bar{y})^2$  and the unbiased estimator  $\tilde{\sigma}^2 = (n/(n-1))\hat{\sigma}^2$ . We already know the bias of  $\tilde{\sigma}^2$  is 0 and the bias of  $\hat{\sigma}^2$  is  $-\sigma^2/n$ . We also know that  $\sum (y_i - \bar{y})^2 / \sigma^2 \sim \chi_{n-1}^2$ . Therefore  $\text{Var}(\sum (y_i - \bar{y})^2 / \sigma^2) = 2(n-1)$ ;  $\text{Var}(\hat{\sigma}^2) = 2(n-1)\sigma^4/n^2$  and  $\text{Var}(\tilde{\sigma}^2) = 2\sigma^4/(n-1)$ . Thus,

$$\begin{aligned} \text{MSE}(\hat{\sigma}^2) &= \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2} = \frac{(2n^2 - 3n + 1)\sigma^4}{n^2(n-1)} \\ &< \frac{2n^2\sigma^4}{n^2(n-1)} = \frac{2\sigma^4}{n-1} = \text{MSE}(\tilde{\sigma}^2), \end{aligned}$$

showing that the unbiased estimator does not minimize MSE. In choosing an estimator for  $\sigma^2$  one must trade bias for variance. That is a common problem in statistics to which there is no universal solution.

We now investigate the variance-bias tradeoff for estimating  $\theta$  in a Binomial distribution. Let  $X_1, \dots, X_n \sim \text{i.i.d. } \text{Bin}(n, \theta)$ . A sensible estimator of  $\theta$  is  $\bar{X} \equiv \sum X_i/n$ . We know that  $\bar{X}$  is unbiased, and its variance is  $\theta(1-\theta)/n$ , so its MSE is also  $\theta(1-\theta)/n$ .

But we could take a Bayesian approach to the problem instead and use the posterior mean,  $\mathbb{E}[\theta|X_1, \dots, X_n]$ , as an estimator of  $\theta$ . For convenience, adopt the prior distribution  $\theta \sim \text{Be}(\alpha, \beta)$ . Then the posterior is given by  $[\theta|X_1, \dots, X_n] \sim \text{Be}(\alpha + X, \beta + n - X)$  and the posterior mean is  $\tilde{\theta} \equiv \mathbb{E}[\theta|X_1, \dots, X_n] = (\alpha + X)/(\alpha + \beta + n)$ . We want the MSE of  $\tilde{\theta}$ , which we shall calculate from its bias and variance.

$$\mathbb{E}[\tilde{\theta}] = \mathbb{E}\left[\frac{X + \alpha}{\alpha + \beta + n}\right] = \frac{n\theta + \alpha}{\alpha + \beta + n}$$

$$\text{Var}[\tilde{\theta}] = \text{Var}\left[\frac{X + \alpha}{\alpha + \beta + n}\right] = \frac{n\theta(1 - \theta)}{(\alpha + \beta + n)^2}.$$

Therefore

$$\text{MSE}(\tilde{\theta}) = \frac{n\theta(1 - \theta)}{(\alpha + \beta + n)^2} + \left(\frac{n\theta + \alpha}{\alpha + \beta + n} - \theta\right)^2.$$

We want to compare  $\text{MSE}(\bar{X})$  to  $\text{MSE}(\tilde{\theta})$ , but the comparison might depend on  $n$ ,  $\alpha$ , and  $\beta$ . Figure 8.1 shows the comparison for four values of  $n$  and four values of  $(\alpha, \beta)$ , always keeping  $\alpha = \beta$ . Note: the Bayes estimator when  $\alpha = \beta = 1$  is the same as  $\bar{X}$ . Using large values of  $(\alpha, \beta)$  looks advantageous when  $\theta \approx .5$ ; small values are advantageous when  $\theta$  is near 0 or 1. There is no general rule for choosing which estimator to use or for choosing  $(\alpha, \beta)$  if we decide to use the Bayes estimator. We might make the choice according to a prior guess of whether  $\theta$  is likely to be near 0, .5, or 1.

Figure 8.1 was produced with the following snippet.

```
theta <- seq(0, 1, length=60)

mse <- function(a, b, n, theta) {
 (n*theta*(1-theta)) / (a+b+n)^2
 + ((n*theta + a)/(a+b+n) - theta)^2
}

mse.out <- data.frame(ab=NULL, n=NULL, mse=NULL)
for(ab in c(0, .5, 1, 4))
 for(n in c(5, 20, 100, 1000)) {
 mse.out <- rbind(mse.out, cbind(a=ab, b=ab, n=n,
 theta=theta, mse=mse(ab, ab, n, theta)))
 }
```

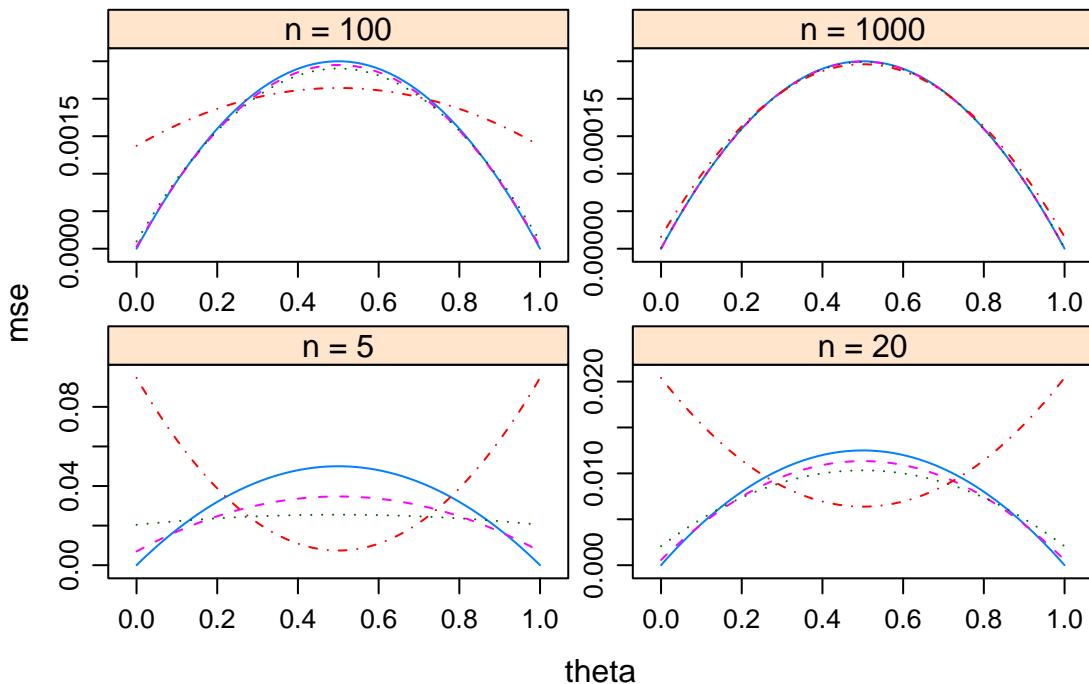


Figure 8.1: Mean Squared Error for estimating Binomial  $\theta$ . Sample size = 5, 20, 100, 1000.  $\alpha = \beta = 0$ : solid line.  $\alpha = \beta = 0.5$ : dashed line.  $\alpha = \beta = 1$ : dotted line.  $\alpha = \beta = 4$ : dash-dotted line.

```

mse.out$n <- factor (paste("n =", mse.out$n),
 levels = paste("n =", c(5,20,100,1000)))

xyplot (mse ~ theta | n, groups=a, data=mse.out, type="l",
 scales = list (relation="free"), lty=1:4)

```

## 8.2 Information

We have talked loosely about the amount of information in a sample and we now want to make the notion more precise. This is an important idea in statistics that can be found in many texts. One devoted wholly to information is KULLBACK [1968].

Suppose that  $Y_1, \dots, Y_n$  is a random sample from an unknown density. And suppose there are only two densities under consideration,  $p_1$  and  $p_2$ . We know that the  $Y_i$ 's are a sample from one of them, but we don't know which. Now consider the first observation,  $Y_1$ . How can we be precise about the amount of information that  $Y_1$  gives us for distinguishing between  $p_1$  and  $p_2$ ? We have already seen that the strength of evidence provided by  $Y_1 = y_1$  is the likelihood ratio  $p_1(y_1)/p_2(y_1)$ . The strength of evidence provided by the whole sample is the product  $\prod_i [p_1(y_i)/p_2(y_i)]$  of the evidence provided by the individual observations. If we transform to a log scale, then the evidence is additive:

$$\log \frac{p_1(y_1, \dots, y_n)}{p_2(y_1, \dots, y_n)} = \sum \log \frac{p_1(y_i)}{p_2(y_i)}.$$

This seems a suitable quantity to call *information*. Therefore we make the following definition.

**Definition 8.5.** The *information* in a datum  $y$  for distinguishing between densities  $p_1$  and  $p_2$  is  $\log \frac{p_1(y)}{p_2(y)}$ .

The information in a random sample is the sum of the information in the individual elements. The  $i$ 'th term in the sum is  $\log \frac{p_1(y_i)}{p_2(y_i)}$ . That term is a function of  $y_i$  and is therefore a random variable whose distribution is determined by either  $p_1$  or  $p_2$ , whichever is the true density generating the sample. If  $p_1$  is the true distribution, then the expected value of that random variable is  $\int \log \frac{p_1(y)}{p_2(y)} p_1(y) dy$  and is the expected amount of information from a single observation. The Law of Large Numbers says that a large sample will contain approximately  $n$  times that amount of information. This view of information was originally discussed thoroughly in KULLBACK AND LEIBLER [1951]. Accordingly, we make the following definition.

**Definition 8.6.**  $I(p_1, p_2) \equiv \int \log \frac{p_1(y)}{p_2(y)} p_1(y) dy$  is called the *Kullback-Leibler divergence* from  $p_1$  to  $p_2$ .

For example, suppose that we're tossing a coin, and we know the probability of Heads is either 0.4 or 0.8. Over a sequence of many tosses, how will information accumulate to distinguish the two possibilities? First, if we're tossing the 0.4 coin, then the likelihood ratio on a single toss is either 0.4/0.8 (if the coin lands Heads) or 0.6/0.2 (if the coin lands Tails). The two possibilities have probabilities 0.4 and 0.6, respectively. So  $I(\text{Bern}(.4), \text{Bern}(.8)) = 0.4 \log(.5) + 0.6 \log(3) \approx .382$ . But for tossing the 0.8 coin,  $I(\text{Bern}(.8), \text{Bern}(.4)) = 0.8 \log(2) + 0.2 \log(1/3) \approx .335$ . Notice that the divergence is not symmetric. The calculation shows that information accumulates slightly faster if we're tossing the .4 coin than if we're tossing the .8 coin. To carry the example to an extreme, suppose one coin is fair but the other is two-headed. For tossing the fair coin,  $I(\text{Bern}(.5), \text{Bern}(1)) = .5 \log(.5) + .5 \log(\infty) = \infty$ ; but  $I(\text{Bern}(1), \text{Bern}(.5)) = \log(2) \approx .693$ . The asymmetry and the infinity make sense: if we're tossing the fair coin then eventually we will toss a Tail and we'll know for certain which coin it is. But if we're tossing the two-headed coin, we'll always toss Heads and never know for certain which coin it is.

**Theorem 8.2.** *The Kullback-Leibler divergence between two distributions is non-negative.*

*Proof.* We prove the theorem in the continuous case; the discrete case is similar. Our proof follows Theorem 3.1 in KULLBACK [1968]. Let  $f_1$  and  $f_2$  be the two densities. We want

$$I(f_1, f_2) = \int \log \left( \frac{f_1(x)}{f_2(x)} \right) f_1(x) dx = \int g(x) \log g(x) f_2(x) dx$$

where  $g(x) = f_1(x)/f_2(x)$ . Let  $\phi(t) = t \log(t)$  and use Taylor's Theorem to get

$$\phi(g(x)) = \phi(1) + [g(x) - 1]\phi'(1) + \frac{1}{2}[g(x) - 1]^2\phi''(h(x))$$

where for every  $x$ ,  $h(x) \in [1, g(x)]$ . Integrate to get

$$\begin{aligned} I(f_1, f_2) &= \int \phi(g(x)) f_2(x) dx = \\ &\int \phi(1) f_2(x) dx + \int [g(x) - 1]\phi'(1) f_2(x) dx + \int \frac{1}{2}[g(x) - 1]^2\phi''(h(x)) f_2(x) dx \\ &= 0 + 0 + \int \frac{1}{2}[g(x) - 1]^2\phi''(h(x)) f_2(x) dx \end{aligned}$$

But  $g(x) \geq 0$  and  $h(x) \in [1, g(x)]$  so  $h(x) \geq 0$  and therefore  $I(f_1, f_2) \geq 0$ .  $\square$

Theorem 8.2 shows that Kullback-Leibler divergences are non-negative. See Exercise 5 for conditions under which they are zero.

While the Kullback-Leibler divergence measures the ability to discriminate between two arbitrary densities, the usual situation in statistics is that we are working with a parametric family of distributions, so we want to look at the ability to discriminate just among distributions in the family. And when the sample size is large, we can usually focus attention on a small neighborhood of  $\theta$ 's. To that end, note first that

$$\begin{aligned}\frac{d^2}{d\theta^2} \log f(x|\theta) &= \frac{d}{d\theta} \left[ \frac{1}{f(x|\theta)} \frac{d}{d\theta} f(x|\theta) \right] \\ &= -\frac{1}{f(x|\theta)^2} \left( \frac{d}{d\theta} f(x|\theta) \right)^2 + \frac{1}{f(x|\theta)} \frac{d^2}{d\theta^2} f(x|\theta).\end{aligned}$$

Now we want to find  $I(f(x|\theta), f(x|\theta + \delta))$ . Expand in a Taylor's series around  $f(x|\theta)$  to get

$$\begin{aligned}I(f(x|\theta), f(x|\theta + \delta)) &= \int f(x|\theta) \{\log f(x|\theta) - \log f(x|\theta + \delta)\} dx \\ &\approx \int f(x|\theta) \left\{ \log f(x|\theta) - \left[ \log f(x|\theta) + \frac{\delta}{f(x|\theta)} \frac{d}{d\theta} f(x|\theta) + \frac{\delta^2}{2} \frac{d^2}{d\theta^2} \log f(x|\theta) \right] \right\} dx \\ &= \delta \int \frac{d}{d\theta} f(x|\theta) + \int f(x|\theta) \frac{\delta^2}{2} \frac{d^2}{d\theta^2} \log f(x|\theta) dx.\end{aligned}$$

Assume we can differentiate under the integral so that

$$\int \frac{d}{d\theta} f(x|\theta) dx = \frac{d}{d\theta} \int f(x|\theta) dx = \frac{d}{d\theta} 1 = 0$$

and the first term in the previous expression vanishes to yield

$$\begin{aligned}I(f(x|\theta), f(x|\theta + \delta)) &\approx \frac{1}{2} \int \delta^2 f(x|\theta) \frac{d^2}{d\theta^2} \log f(x|\theta) dx \\ &= \frac{1}{2} \int \delta^2 f(x|\theta) \left\{ -\frac{1}{f(x|\theta)^2} \left( \frac{d}{d\theta} f(x|\theta) \right)^2 + \frac{1}{f(x|\theta)} \frac{d^2}{d\theta^2} f(x|\theta) \right\} dx \\ &= -\frac{\delta^2}{2} \int f(x|\theta) \left[ \frac{1}{f(x|\theta)} \frac{d}{d\theta} f(x|\theta) \right]^2 dx = -\frac{\delta^2}{2} \mathbb{E} \left[ \frac{d}{d\theta} \log f(x|\theta) \right]^2 dx.\end{aligned}$$

**Definition 8.7.** The expectation on the right in the previous expression,

$$\mathbb{E} \left[ \frac{d}{d\theta} \log p(x|\theta) \right]^2$$

is called the *Fisher Information* for sampling from the family  $f(x|\theta)$  and is denoted  $I(\theta)$ .

Fisher Information is the most studied form of information in statistics and is relevant for inference in parametric families when the sample size is large. The penultimate integral in the previous derivation shows that  $I(\theta)$  is also equal to  $-\mathbb{E}[\frac{d^2}{d\theta^2} \log f(x|\theta)]$ . The word *information* is justified because the Fisher Information also tells us the maximum precision (minimum variance) with which we can estimate parameters, in a limiting, asymptotic sense to be made precise in Section 8.4.3. Next we calculate the information in some common parametric families. Others are in the exercises.

Let  $X \sim N(\mu, \sigma)$  where  $\sigma$  is fixed and  $\mu$  is the unknown parameter.

$$I(\mu) = \mathbb{E} \left[ \frac{d}{d\mu} \log f(x|\mu) \right]^2 = \mathbb{E} \left[ \frac{1}{f(x|\mu)} \frac{d}{d\mu} f(x|\mu) \right]^2 = \mathbb{E} \left[ \frac{1}{f(x|\mu)} \frac{(x-\mu)}{\sigma^2} f(x|\mu) \right]^2 = \frac{1}{\sigma^2}.$$

Let  $X \sim \text{Poi}(\lambda)$ .

$$\begin{aligned} I(\lambda) &= \mathbb{E} \left[ \frac{d}{d\lambda} \log f(x|\lambda) \right]^2 = \mathbb{E} \left[ \frac{1}{f(x|\lambda)} \frac{d}{d\lambda} \frac{e^{-\lambda} \lambda^x}{x!} \right]^2 \\ &= \mathbb{E} \left[ \frac{1}{f(x|\lambda)} \left( -f(x|\lambda) + \frac{x e^{-\lambda} \lambda^{x-1}}{x!} \right) \right]^2 = \mathbb{E} \left[ -1 + \frac{x}{\lambda} \right]^2 = \mathbb{E} \left[ \frac{x-\lambda}{\lambda} \right]^2 = \frac{\text{Var}(x|\lambda)}{\lambda^2} = \frac{1}{\lambda}. \end{aligned}$$

## 8.3 Exponential families

Section 8.3 examines a structure that is shared by many common parametric families of distributions, including the Normal, Gamma, Beta, Binomial, and Poisson. We illustrate with the Normal and Poisson. If  $X \sim \text{Poi}(\lambda)$  then

$$p(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{e^{-\lambda} e^{x \log \lambda}}{x!},$$

which we can write as  $h(x)c(\lambda) \exp(w(\lambda)x)$  for some functions  $h$ ,  $c$ , and  $w$ . If  $X \sim N(\mu, \sigma)$  then

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2)} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2\sigma^2}} e^{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}}$$

which we can write as  $h(x)c(\mu, \sigma) \exp(\sum w_i(\mu, \sigma)t_i(x))$  for some functions  $h$ ,  $c$ ,  $w_i$ , and  $t_i$ . These two examples reveal the structure we're looking for.

A parametric family of distributions  $p(x|\theta)$  is called an *exponential family* if its pdf's can be written as

$$p(x|\theta) = h(x)c(\theta)e^{\sum_{i=1}^d w_i(\theta)t_i(x)}.$$

We study exponential families because they have many features in common. A good source for an advanced treatment of exponential families is BROWN [1986].

The normalizing constant  $c(\theta)$  can be absorbed into the exponent, so it is common to see exponential families written as  $p(x|\theta) = h(x)e^{\sum w_i(\theta)t_i(x)-c^*(\theta)}$ . We have just seen that the family of Poisson distributions is an exponential family with  $d = 1$ . The family of Normal distributions is an exponential family with  $d = 2$ . See Exercise 9 for other examples of exponential families. We may rewrite the parameters as  $\eta_i = w_i(\theta)$ . The  $\eta_i$ 's are called the *natural parameters* of the family. For Poisson distributions, the natural parameter is  $\eta = \log \lambda$ . For Normal distributions the natural parameters are  $(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$ .

A function of the form  $p(x|\eta) \propto h(x)e^{\sum \eta_i t_i(x)}$  can be made a probability density by dividing by its integral, provided that its integral is finite. The set of  $\eta$ 's for which  $\int h(x)e^{\sum \eta_i t_i(x)} < \infty$  is called the *natural parameter space* and denoted  $H$ . For every  $\eta \in H$ , there is a number  $c(\eta)$  such that  $p(x|\eta) = h(x)c(\eta)e^{\sum \eta_i t_i(x)}$  is a probability density.

Suppose  $X_1, \dots, X_n$  is a sample from an exponential family. Then

$$p(x_1, \dots, x_n|\theta) = \prod_{j=1}^n p(x_j|\theta) = \left( \prod_{j=1}^n h(x_j) \right) (c(\theta))^n e^{\sum_{i=1}^d w_i(\theta) \sum_{j=1}^n t_i(x_j)}$$

which shows that the collection  $T_1(x) = \sum_j t_1(x_j), \dots, T_d(x) = \sum_j t_d(x_j)$  is a  $d$ -dimensional sufficient statistic.

In derivations — as in Section 8.4 — we sometimes need to study forms like  $\frac{d}{dy} \left[ \int g(x, y) dx \right]$ . For arbitrary functions  $g$ ,  $\frac{d}{dy} \left[ \int g(x, y) dx \right] \neq \int \frac{d}{dy} [g(x, y)] dx$ . But exponential families obey the following theorem, which we state without proof. The statement comes from LEHMANN [1983].

**Theorem 8.3.** *For any integrable function  $g$  and any  $\eta$  in the interior of  $H$ , the integral*

$$\int g(x)h(x)c(\eta)e^{\sum \eta_i t_i(x)} dx$$

*is continuous and has derivatives of all orders with respect to the  $\eta$ 's, and these can be obtained by differentiating under the integral sign.*

For example, take  $g(x) = 1$ . Write a one-parameter exponential family in the form  $p(x|\eta) = h(x)e^{\eta t(x)-c^*(\eta)}$ . The integral of the density is, as always, 1, so taking derivatives yields

$$0 = \int \frac{d}{d\eta} h(x)e^{\eta t(x)-c^*(\eta)} dx = \int t(x)h(x)e^{\eta t(x)-c^*(\eta)} dx - \int c^{*\prime} h(x)e^{\eta t(x)-c^*(\eta)} dx, \quad (8.1)$$

or  $\mathbb{E}[t(x)] = c^{*\prime}$ .

It is sometimes useful and natural to consider the random variable  $T = t(X)$ . We have just derived its expectation.

## 8.4 Asymptotics

In real life, data sets are finite:  $(y_1, \dots, y_n)$ . Yet we often appeal to the Law of Large Numbers or the Central Limit Theorem, Theorems 1.12, 1.13, and 1.14, which concern the limit of a sequence of random variables as  $n \rightarrow \infty$ . The hope is that when  $n$  is large those theorems will tell us something, at least approximately, about the distribution of the sample mean. But we're faced with the questions "How large is large?" and "How close is the approximation?"

To take an example, we might want to apply the Law of Large Numbers or the Central Limit Theorem to a sequence  $Y_1, Y_2, \dots$  of random variables from a distribution with mean  $\mu$  and SD  $\sigma$ . Here are a few instances of the first several elements of such a sequence.

|       |       |       |       |       |       |       |       |      |     |
|-------|-------|-------|-------|-------|-------|-------|-------|------|-----|
| 0.70  | 0.29  | 0.09  | -0.23 | -0.30 | -0.79 | -0.72 | -0.35 | 1.79 | ... |
| -0.23 | -0.24 | 0.29  | -0.16 | 0.37  | -0.01 | -0.48 | -0.59 | 0.39 | ... |
| -1.10 | -0.91 | -0.34 | 0.22  | 1.07  | -1.51 | -0.41 | -0.65 | 0.07 | ... |
| :     | :     | :     | :     | :     | :     | :     | :     | :    | ..  |

Each sequence occupies one row of the array. The "..." indicates that the sequence continues infinitely. The ":" indicates that there are infinitely many such sequences. The numbers were generated by

```
y <- matrix(NA, 3, 9)
for (i in 1:3) {
 y[i,] <- rnorm(9)
 print(round(y[i,], 2))
}
```

- I chose to generate  $Y_i$ 's from the  $N(0, 1)$  distribution so I used `rnorm`, and so, for this example,  $\mu = 0$  and  $\sigma = 1$ . Those are arbitrary choices. I could have used any values of  $\mu$  and  $\sigma$  and any distribution for which I know how to generate random variables on the computer.
  - `round` does rounding. In this case we're printing each number to two decimal places.

Because there are multiple sequences, each with multiple elements, we need two subscripts to keep track of things properly. Let  $Y_{ij}$  be the  $j$ 'th element of the  $i$ 'th sequence. A real data set  $X_1, \dots, X_n$  is analogous to the first  $n$  observations,  $\{Y_{i,j}\}_{j=1}^n$ , along one row of the array. For each sequence of random variables (each row of the array), we're interested in things like the behavior as  $n \rightarrow \infty$  of the sequence of sample means  $\bar{Y}_{i1}, \bar{Y}_{i2}, \dots$  where  $\bar{Y}_{in} = (Y_{i1} + \dots + Y_{in})/n$ . And for the Central Limit Theorem, we're also interested in the sequence  $Z_{i1}, Z_{i2}, \dots$  where  $Z_{in} = \sqrt{n}(\bar{Y}_{in} - \mu)$ . For the three instances above, the  $\bar{Y}_{in}$ 's and  $Z_{in}$ 's can be printed with

```

for (i in 1:3) {
 print (round (cumsum(y[i,]) / 1:9, 2))
 print (round (cumsum(y[i,]) / (sqrt(1:9)), 2))
}

```

- `cumsum` computes a cumulative sum; so `cumsum(y[1,])` yields the vector  $y[1,1]$ ,  $y[1,1]+y[1,2]$ , ...,  $y[1,1]+\dots+y[1,9]$ . (Print out `cumsum(y[1,])` if you're not sure what it is.) Therefore, `cumsum(y[i,])/1:9` is the sequence of  $\bar{Y}_{in}$ 's.
  - `sqrt` computes the square root. So the second `print` statement prints the sequence of  $Z_{in}$ 's.

The results for the  $\bar{Y}_{in}$ 's are

and for the  $Z_{in}$ 's are

We're interested in the following questions.

1. Will every sequence of  $\bar{Y}_i$ 's or  $Z_i$ 's converge? This is a question about the limit along each row of the array.
2. If they converge, do they all have the same limit?
3. If not every sequence converges, what fraction of them converge; or what is the probability that a randomly chosen sequence of  $\bar{Y}_i$ 's or  $Z_i$ 's converges?
4. For a fixed  $n$ , the collection of random variables  $\{\bar{Y}_{i,n}\}_{i=1}^{\infty}$  are a set of i.i.d. draws from some distribution. What is it? And what is the analogous distribution for  $\{Z_{i,n}\}_{i=1}^{\infty}$ ? This is a question about the distribution along columns of the array.
5. Does the distribution of  $\bar{Y}_n$  or  $Z_n$  depend on  $n$ ?
6. Is there a limiting distribution as  $n \rightarrow \infty$ ?

Some simple examples and the Strong Law of Large Numbers, Theorem 1.13, answer questions 1, 2, and 3 for the sequences of  $\bar{Y}_i$ 's. The Central Limit Theorem, Theorem 1.14, answers question 6 for the sequences of  $Z_i$ 's.

1. Will every sequence of  $\bar{Y}_i$ 's converge? No. Suppose the sequence of  $Y_i$ 's is 1, 2, 3, .... Then  $\{\bar{Y}_i\}$  increases without limit and does not converge.
2. If they converge, do they have the same limit? No. Here are two sequences of  $Y_i$ 's.

$$\begin{array}{cccc} 1 & 1 & 1 & \dots \\ -1 & -1 & -1 & \dots \end{array}$$

The corresponding sequences  $\{\bar{Y}_i\}$  converge to different limits.

3. What is the probability of convergence? The probability of convergence is 1. That's the Strong Law of Large Numbers. In particular, the probability of randomly getting a sequence like 1, 2, 3, ... that doesn't converge is 0. But the Strong Law of Large Numbers says even more. It says

$$P[\lim_{n \rightarrow \infty} \bar{Y}_n = \mu] = 1.$$

So the probability of getting sequences like 1, 1, 1, ... or -1, -1, -1, ... that converges to something other than  $\mu$  is 0.

4. What is the distribution of  $Z_n$ ? We cannot say in general. It depends on the distribution of the individual  $Y_i$ 's.
5. Does the distribution of  $Z_n$  depend on  $n$ ? Yes, except in the special case where  $Y_i \sim N(0, 1)$  for all  $i$ .
6. Is there a limiting distribution? Yes. That's the Central Limit Theorem. Regardless of the distribution of the  $Y_{ij}$ 's, as long as  $\text{Var}(Y_{ij}) < \infty$ , the limit, as  $n \rightarrow \infty$ , of the distribution of  $Z_n$  is  $N(0, 1)$ .

The Law of Large Numbers and the Central Limit Theorem are theorems about the limit as  $n \rightarrow \infty$ . When we use those theorems in practice we hope that our sample size  $n$  is large enough that  $\bar{Y}_{in} \approx \mu$  and  $Z_{in} \sim N(0, 1)$ , approximately. But how large should  $n$  be before relying on these theorems, and how good is the approximation? The answer is, “*It depends on the distribution of the  $Y_{ij}$ 's*”. That's what we look at next.

To illustrate, we generate sequences of  $Y_{ij}$ 's from two distributions, compute  $\bar{Y}_{in}$ 's and  $Z_{in}$ 's for several values of  $n$ , and compare. One distribution is  $U(0, 1)$ ; the other is a recentered and rescaled version of  $Be(.39, .01)$ .

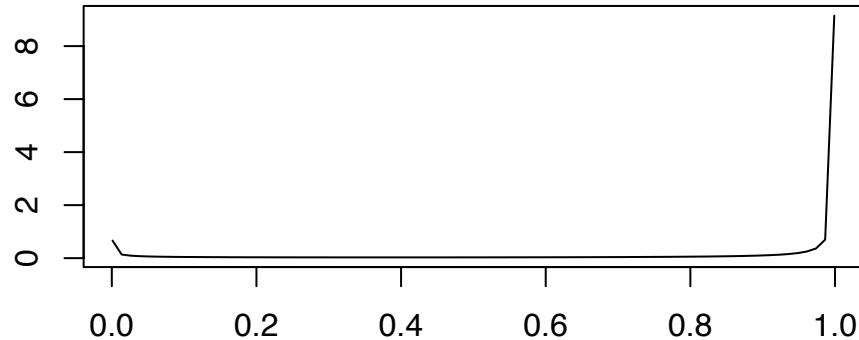
The  $Be(.39, .01)$  density, shown in Figure 8.2, was chosen for its asymmetry. It has a mean of  $.39/.40 = .975$  and a variance of  $(.39)(.01)/((.40)^2(1.40)) \approx .017$ . It was recentered and rescaled to have a mean of  $.5$  and variance of  $1/12$ , the same as the  $U(0, 1)$  distribution.

Densities of the  $\bar{Y}_{in}$ 's are in Figure 8.3. As the sample size increases from  $n = 10$  to  $n = 270$ , the  $\bar{Y}_{in}$ 's from both distributions get closer to their expected value of  $0.5$ . That's the Law of Large Numbers at work. The amount by which they're off their mean goes from about  $\pm .2$  to about  $\pm .04$ . That's Corollary 1.10 at work. And finally, as  $n \rightarrow \infty$ , the densities get more Normal. That's the Central Limit Theorem at work.

Note that the density of the  $\bar{Y}_{in}$ 's derived from the  $U(0, 1)$  distribution is close to Normal even for the smallest sample size, while the density of the  $\bar{Y}_{in}$ 's derived from the  $Be(.39, .01)$  distribution is way off. That's because  $U(0, 1)$  is symmetric and unimodal, and therefore close to Normal to begin with, while  $Be(.39, .01)$  is far from symmetric and unimodal, and therefore far from Normal, to begin with. So  $Be(.39, .01)$  needs a larger  $n$  to make the Central Limit Theorem work; i.e., to be a good approximation.

Figure 8.4 is for the  $Z_{in}$ 's. It's the same as Figure 8.3 except that each density has been recentered and rescaled to have mean  $0$  and variance  $1$ . When put on the same scale we can see that all densities are converging to  $N(0, 1)$ .

Figure 8.2 was produced by

Figure 8.2: The  $\text{Be}(.39, .01)$  density

```
x <- seq (.01, .99, length=80)
plot (x, dbeta(x,.39,.01), type="l", ylab="", xlab="")
```

Figure 8.3 was generated by the following R code.

```
samp.size <- c (10, 30, 90, 270)
n.reps <- 500
Y.1 <- matrix (NA, n.reps, max(samp.size))
Y.2 <- matrix (NA, n.reps, max(samp.size))
for (i in 1:n.reps) {
 Y.1[i,] <- runif (max(samp.size), 0, 1)
 Y.2[i,] <- (rbeta (max(samp.size), 0.39, .01) - .975) *
 sqrt(.4^2*1.4 / (.39*.01*12)) + .5
}
par (mfrow=c(2,2))
for (n in 1:length(samp.size)) {
 Ybar.1 <- apply (Y.1[,1:samp.size[n]], 1, mean)
```

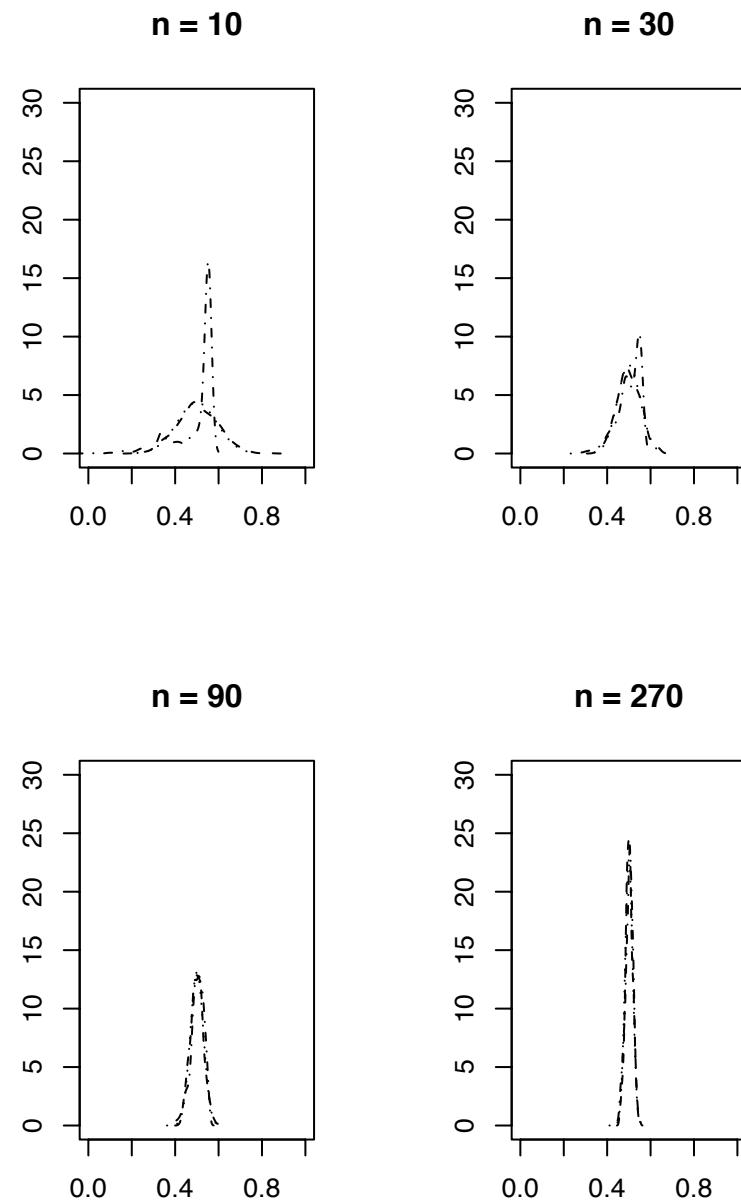


Figure 8.3: Densities of  $\bar{Y}_{in}$  for the  $U(0, 1)$  (dashed), modified  $Be(.39, .01)$  (dash and dot), and Normal (dotted) distributions.

```

Ybar.2 <- apply (Y.2[,1:samp.size[n]], 1, mean)
sd <- sqrt (1 / (12 * samp.size[n]))
x <- seq (.5-3*sd, .5+3*sd, length=60)
y <- dnorm (x, .5, sd)
den1 <- density (Ybar.1)
den2 <- density (Ybar.2)
ymax <- max (y, den1$y, den2$y)
plot (x, y, ylim=c(0,ymax), type="l", lty=3, ylab="",
 xlab="", main=paste("n =", samp.size[n]))
lines (den1, lty=2)
lines (den2, lty=4)
}

```

- The manipulations in the line  $Y.2[i,] <- \dots$  are so  $Y.2$  will have mean 1/2 and variance 1/12.

### 8.4.1 Modes of Convergence

There are at least three different meanings for convergence of a sequence of random variables. Section 8.4.1 describes *convergence in distribution*, *convergence in probability*, and *convergence almost surely*.

**Definition 8.8** (Convergence in Distribution). Let  $Y$  be a random variable and  $Y_1, Y_2, \dots$  be a sequence of random variables. The sequence  $Y_1, Y_2, \dots$  is said to *converge in distribution* to  $Y$  if, at every point  $y$  where  $F_Y$  is continuous,

$$\lim_{n \rightarrow \infty} F_{Y_n}(y) = F_Y(y).$$

**Definition 8.9** (Convergence in Probability). The sequence  $Y_1, Y_2, \dots$  is said to *converge in probability* to  $Y$  if, for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P[|Y_n - Y| < \epsilon] = 1.$$

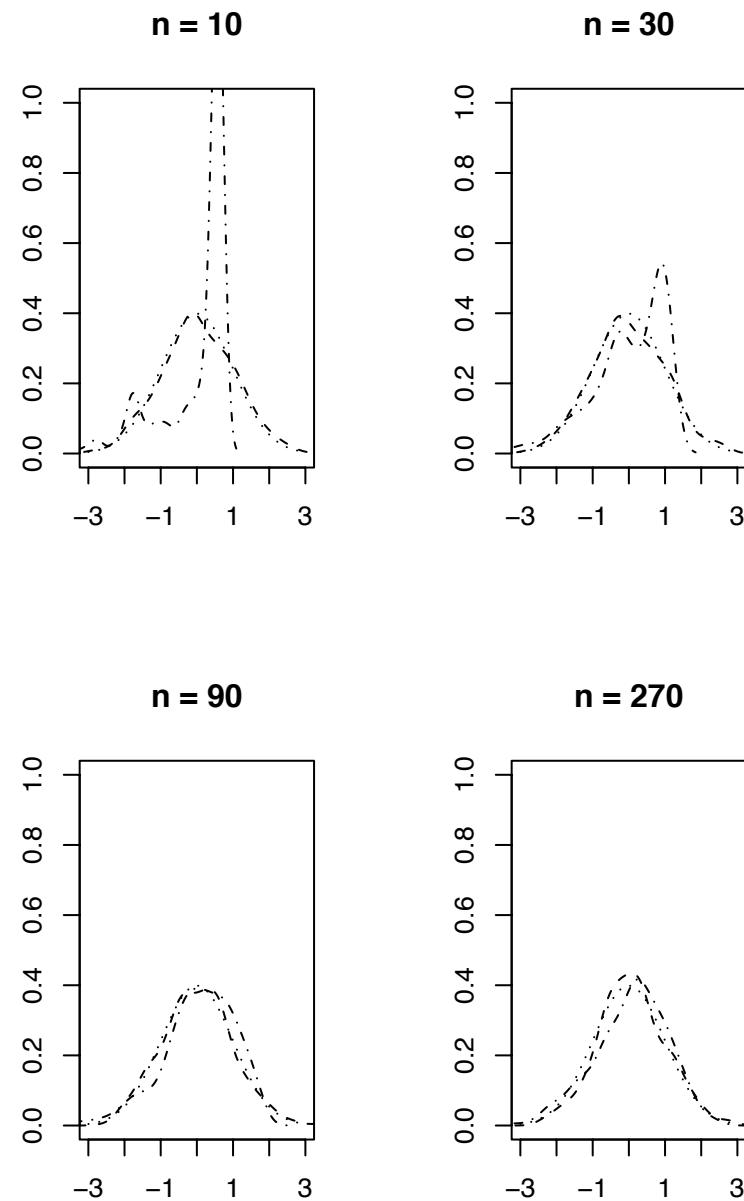


Figure 8.4: Densities of  $Z_{in}$  for the  $U(0, 1)$  (dashed), modified  $Be(.39, .01)$  (dash and dot), and Normal (dotted) distributions.

**Definition 8.10** (Convergence almost surely). The sequence  $Y_1, Y_2, \dots$  is said to *converge almost surely* to  $Y$  if, for every  $\epsilon > 0$ ,

$$P[\lim_{n \rightarrow \infty} |Y_n - Y| < \epsilon] = 1.$$

Almost sure (a.s.) convergence is also called convergence almost everywhere (a.e.) and convergence with probability 1 (w.p.1).

The three modes of convergence are related by the following Theorems, stated here without proof.

**Theorem 8.4.** *If  $Y_n \rightarrow Y$  almost surely, then  $Y_n \rightarrow Y$  in probability.*

**Theorem 8.5.** *If  $Y_n \rightarrow Y$  in probability, then  $Y_n \rightarrow Y$  in distribution.*

One result we will need later is that if  $g$  is a function, continuous at  $c$ , and if  $Y_1, Y_2, \dots$  converges to  $c$  in probability, then  $g(Y_1), g(Y_2), \dots$  converges to  $g(c)$  in probability. The proof is left as an exercise.

The converses to Theorems 8.4 and 8.5 are not true, as the following examples demonstrate.

### Convergence in Distribution

1. Toss a fair penny. For all  $i = 0, 1, \dots$ , let  $X_i = 1$  if the penny lands heads and  $X_i = 0$  if the penny lands tails. (All  $X_i$ 's are equal to each other.) Toss a fair dime. Let  $Y = 1$  if the dime lands heads and  $Y = 0$  if the dime lands tails. Then the sequence  $X_1, X_2, \dots$  converges to  $X_0$  in distribution, in probability and almost surely. The cdf's of  $X_1, X_2, \dots$  and  $Y$  are all the same, so the  $X_i$ 's converge to  $Y$  in distribution. But the  $X_i$ 's do not converge to  $Y$  in probability or almost surely.
2. See the Exercises.

**Convergence in Probability** See Exercise 20 and others.

**Convergence Almost Surely** See Exercise 19 and others.

One reason that convergence in distribution is sometimes useful is the following theorem.

**Theorem 8.6.** *The sequence  $X_n \rightarrow X$  in distribution if and only if  $\mathbb{E}[g(X_n)] \rightarrow \mathbb{E}[g(X)]$  for every bounded, continuous, real-valued function  $g$ .*

The proof is beyond the scope of this book, but may be found in advanced texts on probability.

Sometimes we know that a sequence  $X_n \rightarrow X$  in distribution. But we may be interested in  $Y = g(X)$  for some known function  $g$ . Does  $g(X_n) \rightarrow g(X)$ ? Theorems 8.7 and 8.8 provide answers in some important special cases.

**Theorem 8.7** (Slutsky). *Let  $X_n \rightarrow X$  in distribution and  $Y_n \rightarrow c$  in probability, where  $|c| < \infty$ . Then,*

1.  $X_n + Y_n \rightarrow X + c$  in distribution;
2.  $X_n Y_n \rightarrow cX$  in distribution; and
3.  $X_n / Y_n \rightarrow X/c$  in distribution, if  $c \neq 0$ .

*Proof.* This proof follows SERFLING [1980]. Choose  $t$  such that  $F_X$  is continuous at  $t - c$  and let  $\epsilon > 0$  be such that  $F_X$  is continuous at  $t - c - \epsilon$  and  $t - c + \epsilon$ . (Such a  $t$  and  $\epsilon$  can always be found if  $F_X$  is continuous except at finitely many points.) Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= \Pr[X_n + Y_n \leq t] \\ &= \Pr[X_n + Y_n \leq t; |Y_n - c| \leq \epsilon] + \Pr[X_n + Y_n \leq t; |Y_n - c| > \epsilon] \\ &\leq \Pr[X_n + Y_n \leq t; |Y_n - c| \leq \epsilon] + \Pr[|Y_n - c| > \epsilon] \\ &\leq \Pr[X_n \leq t - c + \epsilon] + \Pr[|Y_n - c| > \epsilon]. \end{aligned}$$

So

$$\begin{aligned} \limsup F_{X_n+Y_n}(t) &\leq \limsup \Pr[X_n \leq t - c + \epsilon] + \Pr[|Y_n - c| > \epsilon] \\ &\leq \limsup \Pr[X_n \leq t - c + \epsilon] = F_X(t - c + \epsilon). \end{aligned}$$

Letting  $\epsilon \downarrow 0$  gives

$$\limsup F_{X_n+Y_n}(t) \leq F_X(t - c) = F_{X+c}(t). \quad (8.2)$$

Similarly,

$$\begin{aligned} \Pr[X_n \leq t - c - \epsilon] &= \Pr[X_n \leq t - c - \epsilon; |Y_n - c| < \epsilon] + \Pr[X_n \leq t - c - \epsilon; |Y_n - c| \geq \epsilon] \\ &\leq \Pr[X_n \leq t - c - \epsilon; |Y_n - c| < \epsilon] + \Pr[|Y_n - c| \geq \epsilon] \\ &= \Pr[X_n + Y_n \leq t - c - \epsilon + Y_n; |Y_n - c| < \epsilon] + \Pr[|Y_n - c| \geq \epsilon] \\ &\leq \Pr[X_n + Y_n \leq t - c - \epsilon + (c + \epsilon); |Y_n - c| < \epsilon] + \Pr[|Y_n - c| \geq \epsilon] \\ &= \Pr[X_n + Y_n \leq t; |Y_n - c| < \epsilon] + \Pr[|Y_n - c| \geq \epsilon] \\ &\leq \Pr[X_n + Y_n \leq t] + \Pr[|Y_n - c| \geq \epsilon]. \end{aligned}$$

So,

$$\liminf \Pr[X_n + Y_n \leq t] + \Pr[|Y_n - c| \geq \epsilon] \geq \liminf \Pr[X_n \leq t - c - \epsilon] = F_X(t - c - \epsilon).$$

Letting  $\epsilon \downarrow 0$  gives

$$\liminf F_{X_n+Y_n}(t) \geq F_X(t - c) = F_{X+c}(t). \quad (8.3)$$

Combining 8.2 and 8.3 yields

$$\lim F_{X_n+Y_n}(t) = F_{X+c}(t),$$

showing that  $X_n + Y_n \rightarrow X + c$  in distribution.

We leave the proofs of parts 2 and 3 as exercises.  $\square$

One application of Slutsky's Theorem is the following, which comes from CASELLA AND BERGER [2002]. Let  $X_1, X_2, \dots$  be a sample from a distribution whose mean  $\mu$  and SD  $\sigma$  are finite. The Central Limit Theorem says

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow N(0, 1)$$

in distribution. But in almost all practical problems we don't know  $\sigma$ . However, we can estimate  $\sigma$  by either the m.l.e.  $\hat{\sigma}$  or the unbiased estimator  $\tilde{\sigma}$ . Both of them are consistent, meaning they converge to  $\sigma$  in probability. Thus,

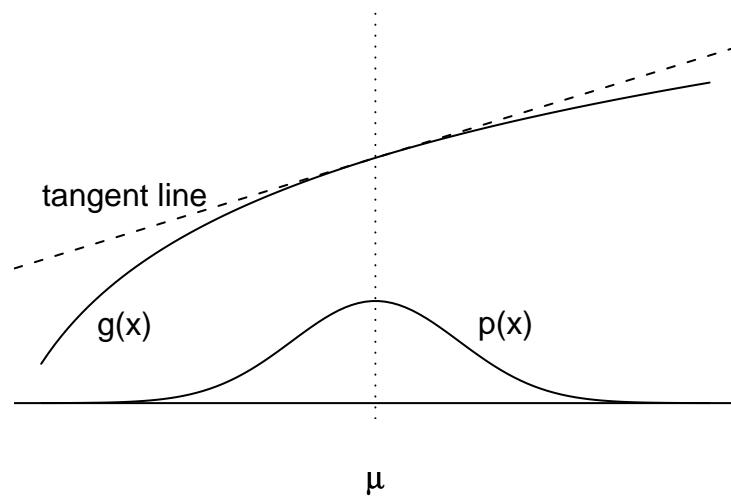
$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} = \frac{\sigma}{\hat{\sigma}} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow N(0, 1) \quad \text{in distribution}$$

by Slutsky's Theorem.

### 8.4.2 The $\delta$ -method

For inferences about  $\mu$ , the Central Limit Theorem tells us that  $\sqrt{n}(\bar{X}_n - \mu)/\sigma \rightarrow N(0, 1)$  in distribution. But sometimes we're interested in a function of  $\mu$ , say  $g(\mu)$ . That's where the  $\delta$ -method comes in. It translates the Central Limit Theorem into a result about  $g(X_n)$  and  $g(\mu)$ . Before stating and proving the  $\delta$ -method, let's try to get some intuition.

Suppose  $X \sim N(\mu, \sigma)$ . Figure 8.5 shows  $p(x)$ , a smooth function  $g(x)$ , and the tangent line at  $x = \mu$ . Now consider a sequence of random variables  $X_n \sim N(\mu, \sigma/n)$ . They would have pictures similar to 8.5, but the densities would become more concentrated around  $\mu$ . As the densities become more concentrated, the relevant portion of  $g$  would become more concentrated around the point of tangency and  $g(X)$  would be more like the transformation represented by the tangent line  $\ell(X) = g(\mu) + g'(\mu)(X - \mu)$ . In the limit,  $g(X) \sim N(g(\mu), \sigma g'(\mu))$ . The  $\delta$ -method formalizes this intuition.

Figure 8.5: The  $\delta$ -method

**Theorem 8.8** (The  $\delta$ -method). *Let  $\sqrt{n}(X_n - \mu)/\sigma \rightarrow N(0, 1)$  in distribution. Let  $g$  be a function with a continuous derivative such that  $g'(\mu)$  is finite and not zero. Then*

$$\sqrt{n} \frac{g(X_n) - g(\mu)}{\sigma g'(\mu)} \rightarrow N(0, 1).$$

*Proof.* Expand  $g$  in a Taylor's series about  $\mu$ :

$$g(X_n) = g(\mu) + g'(\tilde{\mu})(X_n - \mu)$$

for some  $\tilde{\mu} \in [\mu, X_n]$  and therefore

$$\sqrt{n} \frac{g(X_n) - g(\mu)}{\sigma} = \sqrt{n} \frac{g'(\tilde{\mu})(X_n - \mu)}{\sigma} \rightarrow N(0, g'(\mu))$$

by Slutsky's theorem and because  $g'(\tilde{\mu}) \rightarrow g'(\mu)$ . Now use Slutsky's theorem again to get

$$\sqrt{n} \frac{g(X_n) - g(\mu)}{\sigma g'(\mu)} \rightarrow N(0, 1).$$

□

To illustrate the  $\delta$ -method, suppose we're testing a new medical treatment to learn its probability of success,  $\theta$ . Subjects,  $n$  of them, are recruited into a trial and we observe  $X_n$ , the number of successes. We could use the m.l.e.  $\hat{\theta}_n = X_n/n$  as an estimator of  $\theta$ . And we already know that  $\sqrt{n}(\hat{\theta}_n - \theta)/\sqrt{\theta(1-\theta)} \rightarrow N(0, 1)$ . But researchers sometimes want to phrase their results in terms of the odds of success,  $\theta/(1-\theta)$ , or the log odds,  $\log(\theta/(1-\theta))$ . We could use  $\hat{\theta}_n/(1-\hat{\theta}_n)$  to estimate the odds of success, but how would we know its accuracy; what do we know about its distribution? The  $\delta$ -method gives the answer, at least approximately for large  $n$ .

Let  $g(x) = x/(1-x)$ . Then  $g'(\theta) = (1-\theta)^{-2}$ . Therefore,

$$\sqrt{n} \left[ \frac{g(\hat{\theta}_n) - g(\theta)}{\sqrt{\theta(1-\theta)} g'(\theta)} \right] = \sqrt{n} \left[ \frac{\frac{\hat{\theta}_n}{1-\hat{\theta}_n} - \frac{\theta}{1-\theta}}{\sqrt{\theta(1-\theta)}(1-\theta)^{-2}} \right] \rightarrow N(0, 1),$$

or equivalently,

$$\sqrt{n} \left[ \frac{\hat{\theta}_n}{1-\hat{\theta}_n} - \frac{\theta}{1-\theta} \right] \rightarrow N\left(0, \frac{\theta^{1/2}}{(1-\theta)^{3/2}}\right).$$

### 8.4.3 The Asymptotic Behavior of Estimators

Suppose that  $X_1, X_2, \dots$  is an i.i.d. sequence of draws from a distribution  $p(x|\theta)$  and we are trying to estimate  $\theta$ . From  $X_1, \dots, X_n$  we construct an estimate  $\delta_n$ . (We use the notation  $\delta_n$  rather than  $\hat{\theta}_n$  so as not to imply that  $\delta_n$  is the m.l.e.) In most cases, we would choose a consistent estimator:  $\delta_n \rightarrow \theta$  in probability. Typically, the Central Limit Theorem would also apply:  $\sqrt{n}(\delta_n - \theta_n) \rightarrow N(0, \sigma)$  in distribution.

Since the sequence  $\delta_n$  is consistent, then for any  $d \geq 0$ ,  $\Pr[|\delta_n - \theta| > d] \rightarrow 0$ . In the Central Limit Theorem, we multiply the sequences of differences by the sequence  $\sqrt{n}$ , which magnifies it just enough so  $\sqrt{n}(\delta_n - \theta_n)$  doesn't go to 0, doesn't go to infinity, but goes to a distribution, in this case  $N(0, \sigma)$ . But more general behavior is possible. Suppose there is a sequence of constants,  $c_1, \dots$  and a cdf  $G$  such that for all  $x$ ,

$$\Pr[c_n(\delta_n - \theta) \leq x] \rightarrow G(x).$$

The sequence of constants tells us the rate of convergence. We say the errors  $(\delta_n - \theta)$  converge to zero at the rate  $1/c_n$ . For example, in cases where the Central Limit Theorem applies,  $(\delta_n - \theta) \rightarrow 0$  in distribution at the rate of  $1/\sqrt{n}$ . The  $\delta$ -method is another example. When it applies,  $\sqrt{n}(g(\delta_n) - g(\theta)) \rightarrow N(0, \sigma g'(\theta))$ , so the errors  $(g(\delta_n) - g(\theta)) \rightarrow 0$  at the rate of  $1/\sqrt{n}$ .

When there are competing estimators, say  $\delta_n$  and  $\delta'_n$ , we might want to choose the one whose errors converge to zero at a faster rate. But the situation can be more subtle than that. It might be that  $\sqrt{n}(\delta_n - \theta) \rightarrow N(0, \sigma)$  while  $\sqrt{n}(\delta'_n - \theta) \rightarrow N(0, \sigma')$ . In that case, the two estimators converge at the same rate and we might want to know which has the smaller asymptotic SD.

This section of the book is about rates of convergence and comparison of asymptotic SD's. It will guide us in choosing estimators. Of course, with any finite data set, rates do not tell the whole story and can be only a guide. A more advanced treatment of these topics can be found in many texts on mathematical statistics. LEHMANN [1983] is a good example.

First, we have to see whether the concept of rate is well-defined. The following theorem gives that assurance under some mild conditions.

**Theorem 8.9.** *Suppose that for two sequences of constants,  $c_n$  and  $c'_n$ ,  $c_n(\delta_n - \theta) \rightarrow G$  and  $c'_n(\delta_n - \theta) \rightarrow G'$  in distribution. Then, under conditions given in LEHMANN [1983] (pg. 347), there exists a constant  $k$  such that (a)  $c'_n/c_n \rightarrow k$  and (b)  $G'(x) = G(x/k)$  for all  $x$ .*

The proof is beyond our scope but can be found in LEHMANN. Theorem 8.9 says that if two sequences of constants lead to convergence then (a) the sequences themselves differ

asymptotically by only a scale factor and (b), the corresponding limits differ by only that same scale factor.

Now let's look at an illustration of two estimators  $\delta_n$  and  $\delta'_n$  that both converge at the rate of  $1/\sqrt{n}$ . The illustration comes from Example 2.1 in Chapter 5 of LEHMANN. Suppose  $X_1, X_2, \dots$  are a sample from an unknown distribution  $F$ . We want to estimate  $\theta = F(a) = \Pr[X_i \leq a]$ . If we think that  $F$  is approximately  $N(\mu, \sigma)$ , and if we know or can estimate  $\sigma$ , then we might take  $\bar{X}_n$  as an estimator of  $\mu$  and  $\delta_n = \Phi((a - \bar{X}_n)/\sigma)$ , where  $\Phi(x)$  is the  $N(0, 1)$  cdf:  $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$ . But if we don't know that  $F$  is approximately Normal, then we might instead use the estimator  $\delta'_n = n^{-1}(\# \text{ of } X_i \leq a)$ . How can we compare  $\delta_n$  to  $\delta'_n$ ?

First, the Central Limit Theorem says  $\sqrt{n} \left( \frac{\bar{X}_n - \mu}{\sigma} \right) \rightarrow N(0, 1)$ . Then, since  $\delta_n$  is a known transformation of  $\bar{X}_n$ , we can get its asymptotic distribution by Theorem 8.7:

$$\sqrt{n} \frac{\Phi\left(\frac{a-\bar{X}_n}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{\sigma \phi\left(\frac{a-\mu}{\sigma}\right) \frac{1}{\sigma}} \rightarrow N(0, 1)$$

where  $\phi = \Phi'$  is the standard Normal pdf. Equivalently,

$$\sqrt{n} (\delta_n - \theta) \rightarrow N\left(0, \phi\left(\frac{a-\mu}{\sigma}\right)\right). \quad (8.4)$$

On the other hand,  $\delta'_n$  is a transformed version of a Binomial random variable. Let  $Y_n = (\# \text{ of } X_i \leq a)$ . Then  $Y_n \sim \text{Bin}(n, \theta)$ . The Central Limit Theorem says

$$\sqrt{n} \frac{\frac{Y_n}{n} - \theta}{\sqrt{\theta(1-\theta)}} \rightarrow N(0, 1)$$

or, equivalently,

$$\sqrt{n} (\delta'_n - \theta) \rightarrow N\left(0, \sqrt{\theta(1-\theta)}\right). \quad (8.5)$$

Equations 8.4 and 8.5 hold regardless of whether  $F$  really is close to Normal. We can compare their asymptotic variances. Since there is a 1-to-1 correspondence between values of  $\theta$  and values of  $\frac{a-\mu}{\sigma}$ , we can plot  $\phi\left(\frac{a-\mu}{\sigma}\right)$  against  $\sqrt{\theta(1-\theta)}$ . Figure 8.6 shows the plot. The top panel, a plot of the asymptotic SD of  $\delta_n$  as a function of the asymptotic SD of  $\delta'_n$ , shows that  $\delta_n$  has the smaller SD. The bottom panel, a plot of the ratio of the SD's, shows that the advantage of  $\delta_n$  over  $\delta'_n$  grows as  $\theta$  goes to either 0 or 1. Therefore, we would normally prefer  $\delta_n$ , especially for extreme values of  $a$ , at least when the sample size is large enough that the asymptotics are reliable.

When there are competing estimators, there may be one or more that have faster rate, or smaller asymptotic SD, than the others. If so, those estimators would be the best. Our

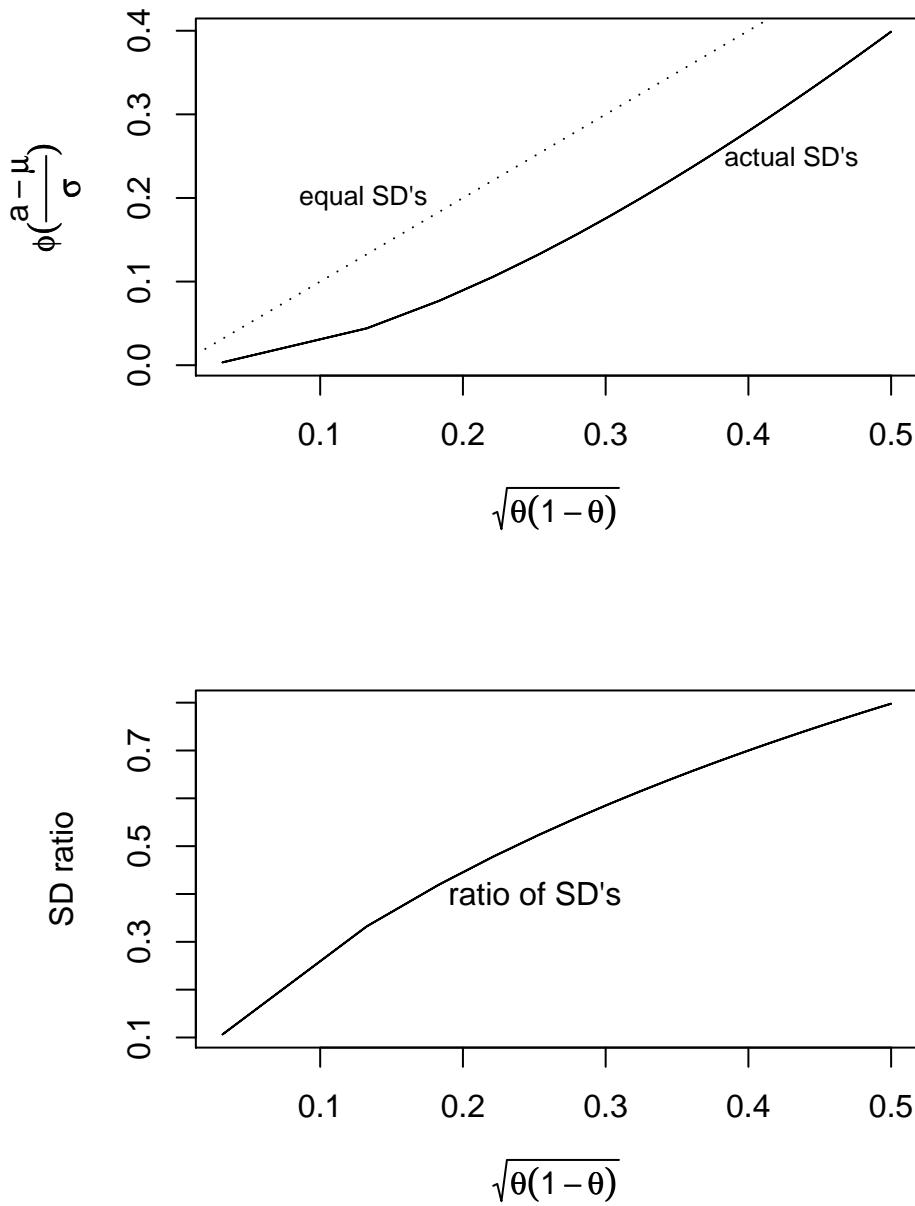


Figure 8.6: Top panel: asymptotic standard deviations of  $\delta_n$  and  $\delta'_n$  for  $\Pr[X \leq a]$ . The solid line shows the actual relationship. The dotted line is the line of equality. Bottom panel: the ratio of asymptotic standard deviations.

next task is to show that under some mild regularity conditions there is, in fact, a lower bound on the asymptotic SD, find what that bound is, and show that the sequence  $\hat{\theta}_n$  of m.l.e.'s achieves that lower bound. Thus the sequence of m.l.e.'s is, in this sense, a good sequence of estimators. There may be other estimators that achieve the same bound, but no sequence can do better. The main theorem is the following, which we state without proof.

**Theorem 8.10.** *Under suitable regularity conditions concerning continuity, differentiability, support of the parametric family, and interchanging the order of differentiation and integration (See LEHMANN, pg. 406.), if  $\delta_n$  is a sequence of estimators satisfying  $\sqrt{n}(\delta_n - \theta) \rightarrow N(0, \text{SD}(\theta))$  in distribution, then  $\text{SD}(\theta) \geq I(\theta)^{-1/2}$ , except possibly for a set of  $\theta$ 's of Lebesgue measure 0.*

Theorem 8.10 shows the sense in which Fisher Information quantifies the ability to estimate a parameter, at least in large samples. Theorem 8.10 provides only a bound. It is quite typical that there are sequences of estimators that achieve the bound. Such estimators are called *asymptotically efficient*. Our next task is to show that under suitable conditions, the sequence  $\hat{\theta}_n$  of m.l.e.'s is asymptotically efficient. The material here follows the development in Chapter 6, Section 2 of LEHMANN.

Suppose that  $X_1, \dots$  is an i.i.d. sequence from  $f(x|\theta_0)$  and that  $f(x|\theta_0)$  is a member of a parametric family  $f(x|\theta)$  that satisfies suitable regularity conditions. (We will not mention the conditions further. A precise statement can be found in LEHMANN and elsewhere.)

**Theorem 8.11.** *For any  $\theta \neq \theta_0$ ,*

$$\Pr\left[\prod_{i=1}^n f(x_i|\theta_0) > \prod_{i=1}^n f(x_i|\theta)\right] \rightarrow 1 \text{ as } n \rightarrow \infty.$$

*Proof.*

$$\prod_{i=1}^n f(x_i|\theta_0) > \prod_{i=1}^n f(x_i|\theta) \iff n^{-1} \sum_{i=1}^n \log \frac{f(x_i|\theta_0)}{f(x_i|\theta)} > 0.$$

The Law of Large Numbers says that the right-hand side goes to  $\mathbb{E}\left[\log \frac{f(x|\theta_0)}{f(x|\theta)}\right]$ , which Theorem 8.2 and the exercises say is positive.  $\square$

Theorem 8.11 says that for any fixed  $\theta$ ,  $\ell(\theta_0) > \ell(\theta)$  with high probability. It does not say that  $\operatorname{argmax} \ell(\theta) = \theta_0$  with high probability. In fact,  $\operatorname{argmax} \ell(\theta) \equiv \hat{\theta}$  is a random variable with a continuous distribution, so  $\Pr[\operatorname{argmax} \ell(\theta) = \theta_0] = 0$ .

**Theorem 8.12.** *The sequence of maximum likelihood estimators,  $\hat{\theta}_n$ , is consistent; i.e.  $\hat{\theta}_n \rightarrow \theta_0$  in probability.*

*Proof.* Choose  $\epsilon > 0$ , let  $\vec{x}_n = x_1, \dots, x_n$ , and let

$$S_n = \{\vec{x}_n : p(\vec{x}_n | \theta_0) \geq p(\vec{x}_n | \theta_0 - \epsilon) \text{ and } p(\vec{x}_n | \theta_0) \geq p(\vec{x}_n | \theta_0 + \epsilon)\}.$$

For any  $\vec{x} \in S_n$ , the likelihood function  $p(\vec{x}_n | \theta)$  has a local maximum in  $[\theta_0 - \epsilon, \theta_0 + \epsilon]$ . And by Theorem 8.11,  $\Pr[S_n] \rightarrow 1$ . Therefore  $\Pr[|\hat{\theta}_n - \theta_0| < \epsilon] \rightarrow 1$ .  $\square$

**Theorem 8.13.** *If there exists a number  $c$  and a function  $M(x)$  such that*

$$\left| \frac{d^3}{d\theta^3} \log p(x | \theta) \right| \leq M(x)$$

*for all  $x$  and for all  $\theta \in (\theta_0 - c, \theta_0 + c)$ , then the sequence of maximum likelihood estimators,  $\hat{\theta}_n$ , is asymptotically efficient; i.e.*

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow N\left(0, \frac{1}{\sqrt{I(\theta_0)}}\right) \text{ in distribution.}$$

*Proof.* Let  $g(\theta)$  be the derivative of the log-likelihood function:

$$g(\theta) = \frac{d}{d\theta} \log p(\vec{x}_n | \theta) = \sum \frac{p'(x_i | \theta)}{p(x_i | \theta)}.$$

We want to examine  $g(\hat{\theta}_n)$ ; we know  $g(\hat{\theta}_n)$  is close to  $g(\hat{\theta}_0)$ , so expand in a Taylor's series.

$$g(\hat{\theta}_n) = g(\theta_0) + (\hat{\theta}_n - \theta_0)g'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 g^{(2)}(\theta^*)$$

for some  $\theta^* \in [\hat{\theta}_n, \theta_0]$ . By assumption, the left-hand side is zero. Rearrange terms to get

$$(\hat{\theta}_n - \theta_0) = -\frac{g(\theta_0)}{g'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)g^{(2)}(\theta^*)}$$

or

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}}g(\theta_0)}{-\frac{1}{n}g'(\theta_0) - \frac{1}{2n}(\hat{\theta}_n - \theta_0)g^{(2)}(\theta^*)}$$

Now we're going to analyze the numerator and the two terms in the denominator separately, then use Slutsky's Theorem (Thm 8.7).

1.  $g(\theta_0)$  is the sum of terms like  $p'(x_i | \theta)/p(x_i | \theta)$ . Because  $x_i$  is a random variable, each term is a random variable. Because the  $x_i$ 's are i.i.d., the terms are i.i.d. The expected value of each term is

$$\mathbb{E}\left[\frac{p'(x_i | \theta)}{p(x_i | \theta)}\right] = \int \frac{d}{d\theta} p(x_i | \theta) d\theta = \frac{d}{d\theta} \int p(x_i | \theta) d\theta = 0.$$

The Central Limit Theorem applies so, by the definition of  $I(\theta)$ ,  $\frac{1}{\sqrt{n}}g(\theta_0) \rightarrow N(0, \sqrt{I(\theta)})$  in probability.

2.  $g'(\theta_0)$  is also the sum of i.i.d.summands. Specifically,

$$\frac{1}{n}g'(\theta) = \frac{1}{n} \sum \frac{p^{(2)}(x_i | \theta)}{p(x_i | \theta)^2} - \frac{[p'(x_i | \theta)]^2}{p(x_i | \theta)^2}.$$

The expectation of the first term is 0; the expectation of the second is  $I(\theta)$ . Therefore  $\frac{1}{n}g'(\theta) \rightarrow -I(\theta)$  by the Law of Large Numbers.

3. By assumption of the theorem,  $g^{(2)}(\theta^*) \leq M(x)$  for all  $\theta$ 's in an interval around  $\theta_0$ . And  $\theta^* \rightarrow \theta_0$  in probability, so  $\frac{1}{2n}(\hat{\theta}_n - \theta_0)g^{(2)}(\theta^*) \rightarrow 0$  in probability.

Now Slutsky's Theorem gives the desired result.  $\square$

The condition of Theorem 8.13 may seem awkward, but it is often easy to check. The following corollary is an example.

**Corollary 8.14.** *If  $\{p(x | \theta)\}$  is a one-dimensional exponential family, then  $\hat{\theta}_n$  is asymptotically efficient.*

*Proof.* If  $\{p(x | \theta)\}$  is a one-dimensional exponential family, then

$$\left| \frac{d^3}{d\theta^3} \log p(x | \theta) \right| = \left| \frac{d^3}{d\theta^3} \log h(x) + \log c(\theta) + w(\theta)t(x) \right|$$

which satisfies the conditions of the theorem, at least when  $c$ ,  $w$ , and  $t$  are continuous functions.  $\square$

## 8.5 Exercises

1. Let  $Y_1, \dots, Y_n$  be a sample from  $N(\mu, \sigma^2)$ .

- (a) Suppose  $\mu$  is unknown but  $\sigma$  is known. Find a one dimensional sufficient statistic for  $\mu$ .
- (b) Suppose  $\mu$  is known but  $\sigma$  is unknown. Find a one dimensional sufficient statistic for  $\sigma$ .
- (c) Suppose  $\mu$  and  $\sigma$  are both unknown. Find a two dimensional sufficient statistic for  $(\mu, \sigma)$ .
2. Let  $Y_1, \dots, Y_n$  be a sample from  $\text{Be}(\alpha, \beta)$ . Find a two dimensional sufficient statistic for  $(\alpha, \beta)$ .
3. Let  $Y_1, \dots, Y_n \sim \text{i.i.d. } U(-\theta, \theta)$ . Find a low dimensional sufficient statistic for  $\theta$ .
4. Let  $Y_1, \dots, Y_n \sim \text{i.i.d. } U(0, \theta)$ .
- Find the m.l.e.  $\hat{\theta}$ .
  - Find the distribution of  $\hat{\theta}$ .
  - Find  $\mathbb{E}[\hat{\theta}]$ .
  - Find  $\text{Var}[\hat{\theta}]$ .
  - Find  $\text{MSE}(\hat{\theta})$ .
  - Because  $\mathbb{E}[\bar{Y}] = \theta/2$ , we might consider whether  $\tilde{\theta} \equiv 2\bar{Y}$  as an estimator of  $\theta$ .
    - Is  $\tilde{\theta}$  a function of a sufficient statistic?
    - Is  $\tilde{\theta}$  unbiased?
    - Find  $\text{MSE}(\tilde{\theta})$ . Compare to  $\text{MSE}(\hat{\theta})$ .
5. Theorem 8.2 shows that Kullback-Leibler divergences are non-negative. This exercise investigates conditions under which they are zero.
- Let  $F$  be a distribution. Find the Kullback-Leibler divergence from  $F$  to itself.
  - Look at the proof of Theorem 8.2 to see what conditions are necessary for  $I(f_1, f_2)$  to be zero. Prove that if  $f_1$  and  $f_2$  differ on any interval, then  $I(f_1, f_2) > 0$ .
6. (a) Find the Kullback-Leibler divergence from  $\text{Bern}(p_1)$  to  $\text{Bern}(p_2)$  and from  $\text{Bern}(p_2)$  to  $\text{Bern}(p_1)$ .
- (b) Find the Kullback-Leibler divergence from  $\text{Bin}(n, p_1)$  to  $\text{Bin}(n, p_2)$  and from  $\text{Bin}(n, p_2)$  to  $\text{Bin}(n, p_1)$ .

7. (a) Let  $X \sim N(\mu, \sigma)$  where  $\mu$  is fixed. Find  $I(\sigma)$ .  
 (b) Let  $X \sim \text{Bin}(n, \theta)$ . Find  $I(\theta)$ .
8. (a) Let  $X \sim \text{Poi}(\lambda)$ . We know  $I(\lambda) = \lambda^{-1}$ . But we may be interested in  $\lambda^* \equiv \log \lambda$ . Find  $I(\lambda^*)$ .  
 (b) Let  $X \sim f(x|\theta)$ . Let  $\phi = h(\theta)$ . Show  $I(\phi) = (\frac{d\theta}{d\phi})^2 I(\theta)$ .
9. Show that the following are exponential families of distributions. In each case, identify the functions  $h$ ,  $c$ ,  $w_i$ , and  $t_i$  and find the natural parameters.
  - (a)  $\text{Bin}(n, \theta)$  where  $n$  is known and  $\theta$  is the parameter.
  - (b)  $\text{Gam}(\alpha, \beta)$ .
  - (c)  $\text{Be}(\alpha, \beta)$ .
10. Verify that Equation 8.1 gives the correct value for the means of the following distributions.
  - (a)  $\text{Poi}(\lambda)$ .
  - (b)  $\text{Exp}(\theta)$ .
  - (c)  $\text{Bin}(n, \theta)$ .
11. Differentiate Equation 8.1 to show  $\text{Var}(t(x)) = c^{*(2)}$ .
12. Derive the two-parameter version of Equation 8.1.
13. In a one-parameter exponential family, it is sometimes natural and useful to consider the random variable  $T = t(x)$ . Equation 8.1 gives  $\mathbb{E}[T]$ .
  - (a) Use the method of transformations to find  $p(t|\eta)$ . Show that it is an exponential family.
  - (b) Find the moment generating function  $M_T(s)$  of  $T$ .
14. Prove that if  $g$  is a function continuous at a number  $c$ , and if  $\{Y_n\} \rightarrow c$  in probability, then  $\{g(Y_n)\} \rightarrow c$  in probability.
15. Prove the claims in item 1 on page 434 that  $X_n \rightarrow X_0$  in distribution, in probability, and almost surely, but  $X_n \rightarrow Y$  in distribution only.

16. Let  $X_n \sim N(0, 1/\sqrt{n})$ . Does the sequence  $\{X_n\}_{n=1}^{\infty}$  converge? Explain why or why not. If yes, also explain in what sense it converges — distribution, probability or almost sure — and find its limit.
17. Let  $X_1, X_2, \dots \sim$  i.i.d.  $N(\mu, \sigma)$  and let  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ . Does the sequence  $\{\bar{X}_n\}_{n=1}^{\infty}$  converge? In what sense? To what limit? Justify your answer.
18. Let  $X_1, X_2, \dots$  be an i.i.d. random sample from a distribution  $F$  with mean  $\mu$  and SD  $\sigma$  and let  $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ . A well-known theorem says that  $\{Z_n\}_{n=1}^{\infty}$  converges in distribution to a well-known distribution. What is the theorem and what is the distribution?
19. Let  $U \sim U(0, 1)$ . Now define the sequence of random variables  $X_1, \dots$  in terms of  $U$  by

$$X_n = \begin{cases} 1 & \text{if } U \leq n^{-1} \\ 0 & \text{otherwise.} \end{cases}$$

- (a) What is the distribution of  $X_n$ ?
- (b) Find the limit, in distribution, of the  $X_n$ 's.
- (c) Show that the  $X_n$ 's converge to that limit in probability.
- (d) Show that the  $X_n$ 's converge to that limit almost surely.
- (e) Find the sequence of numbers  $\mathbb{E}X_1, \mathbb{E}X_2, \dots$ .
- (f) Does the sequence  $\mathbb{E}X_n$  converge to  $\mathbb{E}X$ ?
20. This exercise is similar to Exercise 19 but with a subtle difference. Let  $U \sim U(0, 1)$ . Now define the sequence of constants  $c_0 = 0, c_1 = 1$  and, in general,  $c_n = c_{n-1} + 1/n$ . In defining the  $c_i$ 's, addition is carried out modulo 1; so  $c_2 = (1+1/2) \bmod 1 = 1/2$ , etc. Now define the sequence of random variables  $X_1, \dots$  in terms of  $U$  by

$$X_n = \begin{cases} 1 & \text{if } X_n \in [c_{n-1}, c_n] \\ 0 & \text{otherwise.} \end{cases}$$

where intervals are understood to wrap around the unit interval. For example,  $[c_3, c_4] = [5/6, 13/12] = [5/6, 1/12]$  is understood to be the union  $[5/6, 1] \cup [0, 1/12]$ . (It may help to draw a picture.)

- (a) What is the distribution of  $X_n$ ?
- (b) Find the limit, in distribution, of the  $X_n$ 's.

- (c) Find the limit, in probability, of the  $X_n$ 's.
- (d) Show that the  $X_n$ 's do not converge to that limit almost surely.
21. (a) Prove part 2 of Slutsky's theorem (8.7).
- (b) Prove part 3 of Slutsky's theorem (8.7).
22. Let  $X_n \sim \text{Bin}(n, \theta)$  and let  $\hat{\theta}_n = X_n/n$ . Use the  $\delta$ -method to find the asymptotic distribution of the log-odds,  $\log(\theta/(1 - \theta))$ .
23. In Figure 8.6, show that the ratio of asymptotic SD's,  $\phi\left(\frac{a-\mu}{\sigma}\right) / \sqrt{\theta(1-\theta)}$ , goes to infinity as  $\theta$  goes to 0 and also as  $\theta$  goes to 1.
24. Starting from Theorem 8.10, show that if  $\eta_n = h(\delta_n)$  is a sequence of estimators satisfying  $\sqrt{n}(\eta_n - h(\theta)) \rightarrow N(0, \text{SD}(h(\theta)))$ , then  $\text{SD}(h(\theta)) \geq h'(\theta) / \sqrt{I(\theta)}$ .

## BIBLIOGRAPHY

*Consumer Reports*, June:366–367, 1986.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York, 2nd edition, 1984.

D. F. Andrews and A. M. Herzberg. *Data*. Springer-Verlag, New York, 1985.

H. Bateman. On the probability distribution of  $\alpha$  particles. *Philosophical Magazine Series 6*, 20:704–705, 1910.

Richard J. Bolton and David J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17:235–255, 1992.

Paul Brodeur. Annals of radiation, the cancer at Slater school. *The New Yorker*, Dec. 7, 1992.

Lawrence D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 9. Institute of Mathematical Statistics, Hayward, CA, 1986. ISBN 0-940600-10-2.

Jason C. Buchan, Susan C. Alberts, Joan B. Silk, and Jeanne Altmann. True paternal care in a multi-male primate society. *Nature*, 425:179–181, 2003.

D. P. Byar. The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: Comparisons of placebo, pyridoxine, and topical thiotepa. In M. Pavone-Macaluso, P. H. Smith, and F. Edsmyn, editors, *Bladder Tumors and Other Topics in Urological Oncology*, pages 363–370. Plenum, New York, 1980.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, second edition, 2002.

- Lorraine Denby and Daryl Pregibon. An example of the use of graphics in regression. *The American Statistician*, 41:33–38, 1987.
- A. J. Dobson. *An Introduction to Statistical Modelling*. Chapman and Hall, London, 1983.
- D. Freedman, R. Pisani, and R. Purves. *Statistics*. W. W. Norton and Company, New York, 4th edition, 1998.
- Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, Boca Raton, 2nd edition, 2004.
- S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- John Hassall. *The Old Nursery Stories and Rhymes*. Blackie and Son Limited, London, 1909.
- W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.
- S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.
- Solomon Kullback. *Information Theory and Statistics*. Dover Publications, Inc., 1968.
- Shannon LaDeau and James Clark. Rising co<sub>2</sub> levels and the fecundity of forest trees. *Science*, 292(5514):95–98, 2001.
- Michael Lavine. What is Bayesian statistics and why everything else is wrong. *The Journal of Undergraduate Mathematics and Its Applications*, 20:165–174, 1999.
- Michael Lavine, Brian Beckage, and James S. Clark. Statistical modelling of seedling mortality. *Journal of Agricultural, Biological and Environmental Statistics*, 7:21–41, 2002.
- E. L. Lehmann. *Theory of Point Estimation*. John Wiley, New York, 1983.
- Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York, 2004.
- Jean-Michel Marin and Christian P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag, New York, 2007.

- N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21: 1087–1092, 1953.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL [HTTP://WWW.R-PROJECT.ORG](http://www.R-project.org). ISBN 3-900051-07-0.
- Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 1997.
- E. Rutherford and H. Geiger. The probability variations in the distribution of  $\alpha$  particles. *Philosophical Magazine Series 6*, 20:698–704, 1910.
- Mark J. Schervish. *Theory of Statistics*. Springer-Verlag, New York, 1995.
- Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley, New York, 1980.
- T.S. Tsou and R.M. Royall. Robust likelihoods. *Journal of the American Statistical Association*, 90:316–320, 1995.
- Jessica Utts. Replication and meta-analysis in parapsychology. *Statistical Science*, 4: 363–403, 1991.
- W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.
- L. J. Wei, D. Y. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84:1065–1073, 1989.
- Sanford Weisberg. *Applied Linear Regression*. John Wiley & Sons, New York, second edition, 1985.
- A.S. Yang. Seasonality, division of labor, and dynamics of colony-level nutrient storage in the ant *Pheidole morrisii*. *Insectes Sociaux*, 53:456–452, 2006.

# INDEX

- $\alpha$  particle, 291
- autocorrelation, 391
- autoregression, 397
- bandwidth, 105
- bias, 418
- case, 213
- cdf, *see* cumulative distribution function
- Central Limit Theorem, 79
- change of variables, 12
- characteristic functions, 271
- Chebychev's Inequality, 78
- chi-squared distribution, 308
- consistency, 417
- Convergence
  - almost surely, 433
  - in distribution, 431
  - in probability, 433
- coplots, 125
- correlation, 52
- covariance, 50
- covariance matrix, 263
- covariate, 213
- cross tabulation, 115
- cumulative distribution function, 267
- cumulative hazard function, 408
- DASL, *see* Data and Story Library, *see also* Data and Story Library, 201
- Data and Story Library, 103, 139
- density
  - probability, 261
- density estimation, 104
- dependence, 53
- distribution, 2
- Distributions
  - Bernoulli, 276
  - Beta, 308
  - Binomial, 14, 275
  - Cauchy, 331
  - Exponential, 20
  - Gamma, 301
  - inverse Gamma, 338
  - Multinomial, 285
  - Negative binomial, 280
  - Normal, 22, 311
  - Poisson, 17, 287
  - standard multivariate Normal, 316
  - standard Normal, 29
  - Uniform, 300
- errors, 216
- estimate, 151
- expected value, 32
- explanatory variable, 213

- fitted values, 221, 245  
fitting, 221  
floor, 418  
formula, 222  
  
gamma function, 301  
Gaussian density, 311  
generalized moment, 39  
genotype, 285  
  
half-life, 305  
histogram, 103  
  
independence, 53  
    joint, 262  
    mutual, 262  
indicator function, 55  
indicator variable, 55  
  
Jacobian, 264  
  
Kaplan-Meier estimate, 407  
  
Laplace transform, 270  
Law of Large Numbers, 78  
likelihood function, 131  
likelihood set, 153  
linear model, 216  
linear predictor, 237  
location parameter, 315  
logistic regression, 237  
logit, 237  
  
marginal likelihood, 138  
Markov chain Monte Carlo, 342  
maximum likelihood estimate, 151  
mean, 32  
mean squared error, 419  
median, 94  
Mediterranean tongue, 30  
mgf, *see* moment generating function  
  
minimal sufficient, 417  
moment, 39  
moment generating function, 270  
mosaic plot, 115  
multinomial coefficient, 286  
multivariate  
    change of variables, 264  
  
order statistic, 95, 417  
outer product, 320  
  
parameter, 14, 131, 275  
parametric family, 14, 131, 275  
partial autocorrelation, 397  
pdf, *see* probability density, *see also* probability density  
physics, 8  
Poisson process, 307  
predicted values, 245  
probability  
    continuous, 1, 7  
    density, 7  
    discrete, 1, 6  
proportional hazards model, 408  
  
QQ plots, 110  
quantile, 94  
  
R commands  
    !, 58  
    ==, 4  
    [[ ]], *see* subscript  
    [], *see* subscript  
    #, 4  
    %\*%, 345  
    %o%, 320  
    ~, 222  
    abline, 145  
    acf, 391  
    apply, 62

ar, 397  
array, 66  
arrows, 37  
as.factor, 242  
assignment, 4, 6  
 $+<-+$ , 4  
boxplot, 60, 109  
c, 9  
cbind, 75  
contour, 145  
coplot, 125  
cor, 53  
cov, 52  
crossprod, 345  
cumsum, 426  
data.frame, 242  
dbinom, 17, 278  
density, 12  
dexp, 22  
diff, 313, 401  
dim, 72  
dimnames, 298  
dmultinom, 287  
dnbinom, 283  
dnorm, 23  
dpois, 20  
expression, 133  
filter, 398  
fitted, 247  
for, 5  
glm, 241  
hist, 12  
if, 6  
is.na, 145  
legend, 22  
length, 73  
lines, 12  
list, 128  
lm, 222  
log10, 149  
lowess, 205  
matplot, 22  
matrix, 22, 60  
mean, 12  
median, 94  
mosaicplot, 115  
names, 72  
pacf, 397  
pairs, 50  
par, 10  
paste, 22  
pbinom, 278  
plot, 9, 20  
plot.ecdf, 97  
plot.ts, 391  
pnbinom, 283  
print, 5  
qbinom, 278  
qnbinom, 283  
qqnorm, 112  
quantile, 95  
rbinom, 39, 278  
read.table, 72  
rep, 5, 6  
rmultinom, 287  
rnbinom, 283  
rnorm, 26  
round, 426  
sample, 3  
scan, 128  
segments, 268  
seq, 20  
solve, 345  
sqrt, 426  
stripchart, 107  
subscript, 72, 73

- sum, 4, 5
- supsmu, 205
- Surv, 407
- survfit, 407
- t, 345
- tapply, 75
- text, 39
- unique, 103
- var, 12
- while, 58
- xyplot, 377
- R data sets
  - airquality, 199
  - attenu, 199
  - beaver1, 391, 393
  - co2, 63, 391
  - discoveries, 10
  - EuStockMarkets, 391
  - faithful, 119, 199
  - iris, 50
  - ldeaths, 391
  - lynx, 391
  - mtcars, 199, 226, 245
  - PlantGrowth, 211
  - presidents, 391, 393
  - Seatbelts, 391
  - sunspot.month, 391
  - Tooth Growth, 99, 181
  - UCBAdmissions, 112
- R datasets
  - Orthodont, 376
- R packages
  - lattice, 377
  - nlme, 376
- random vector, 261
- regression function, 205
- regressor, 213
- residual, 205
- residual plots, 228
- response variable, 213
- sampling distribution, 156
- scale parameter, 315
- scatterplot smoothers, 202
- standard deviation, 34
- standard error, 160
- standard Normal distribution, 315
- standard units, 26
- stationary distribution, 353
- StatLib, 72
- Strong Law of Large Numbers, 78
- sufficient statistic, 414
- trace plot, 357
- variance, 34
- Weak Law of Large Numbers, 78

## INDEX OF EXAMPLES

- 1970 draft lottery, 204
- baboons, 187
- bladder cancer, 359
- CEO salary, 141
- craps, 2, 3, 15, 49, 58, 59, 157
- FACE, 64, 144, 236
- hot dogs, 105, 210, 216, 225
- Ice Cream Consumption, 220
- neurobiology, 127, 298
- O-rings, 236
- ocean temperatures, 29, 316
- quiz scores, 109, 148
- Rutherford and Geiger, 296
- seedlings, 18, 43, 49, 56, 136, 148, 167,  
208, 245
- Slater school, 136, 140, 193