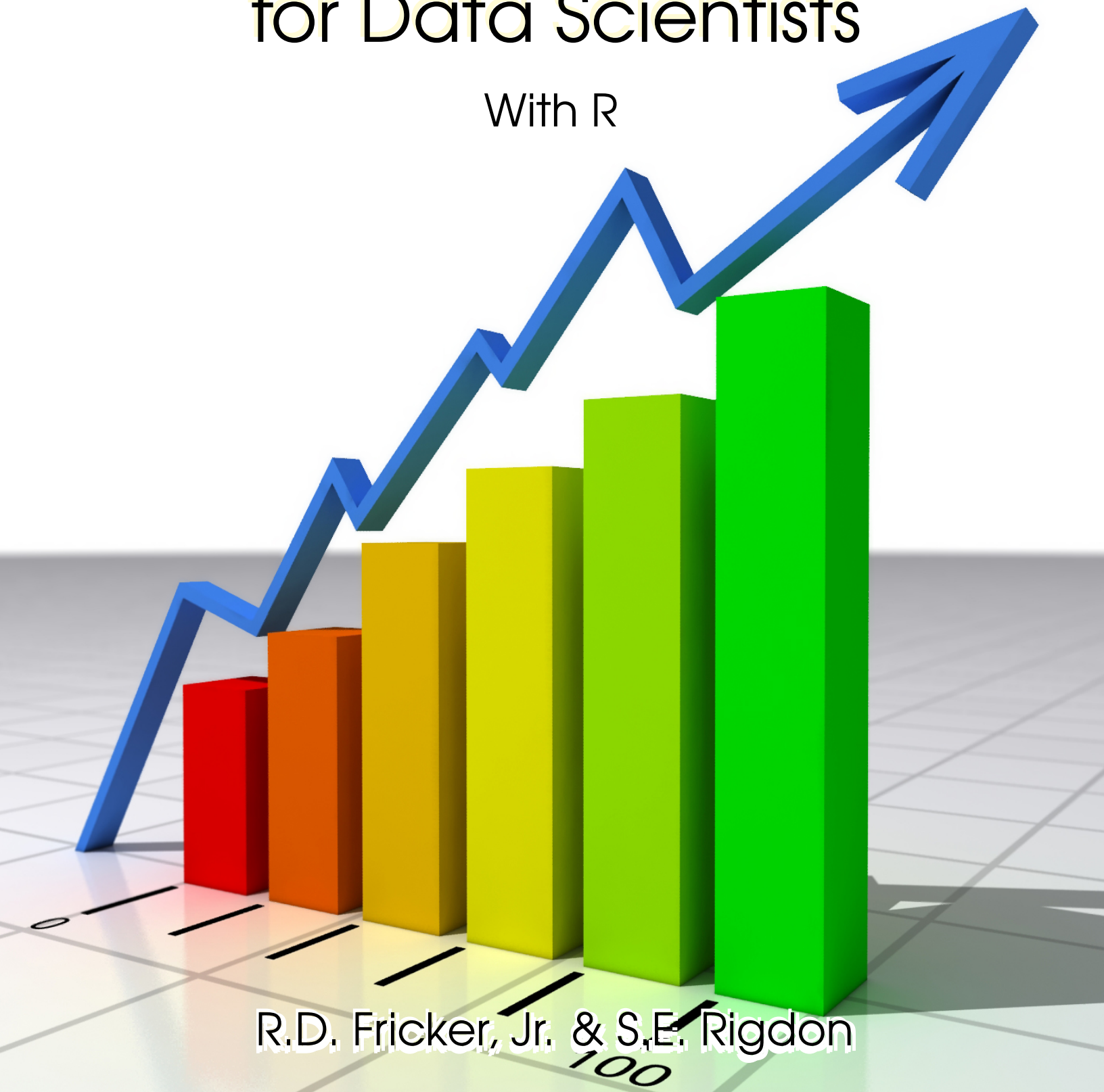
Introduction to Probability & Statistics for Data Scientists

With R



R.D. Fricker, Jr. & S.E. Rigdon

First printing, January 2020

DEDICATIONS

Ron Fricker

To my spouse, Christine:
Tu ventus sub alis meis es.

And to my first statistics professor, Randy Spoeri:
You introduced me to the subject and made it fun.

❊

Steve Rigdon

To be written

❊

Descriptive Statistics & Data Science

1 Introduction . . . 3

1.1 Data Science & Statistics 4

1.1.1 Why Study Probability and Statistics? . . . 5

1.2 More On Statistics 6

1.2.1 Populations and Samples . . . 8

1.2.2 Descriptive versus Inferential Statistics . . . 9

1.2.3 How to Study Statistics . . . 10

1.3 An Introduction to R 10

1.3.1 Getting Started . . . 11

1.3.2 Some Important R Paradigms . . . 16

1.3.3 Useful References and Getting Help . . . 19

1.3.4 Extending R: Installing Packages . . . 21

1.3.5 A Few Final Notes . . . 22

1.4 Chapter Summary 25

1.4.1 Problems . . . 26

2 Descriptive Statistics . . . 29

2.1 Introduction 29

2.1.1 Types of Data . . . 30

2.1.2 Example Data: U.S. Domestic Flights from 1987 to 2008 . . . 31

2.2 Cross-sectional Data **32**
2.2.1 Measures of Location . . . 34
2.2.2 Measures of Variation . . . 39
2.2.3 Measures of How Two Variables Co-vary . . . 42
2.2.4 Other Summary Statistics . . . 48
2.3 Longitudinal Data **51**
2.3.1 Statistics for Repeating Cross-sections of Data . . . 52
2.3.2 Statistics for Moving Windows of Data . . . 53
2.4 Tabular Summaries of Data **56**
2.5 Chapter Summary **59**
2.5.1 Problems . . . 60

3 Data Visualization . . . 63
3.1 Introduction **63**
3.2 Traditional Statistical Graphics **64**
3.2.1 Bar Charts . . . 64
3.2.2 Histograms . . . 69
3.2.3 Lattice (or Trellis) Plots . . . 73
3.2.4 Box Plots . . . 75
3.2.5 Scatterplots . . . 79
3.3 Graphics for Longitudinal Data **85**
3.3.1 Time Series Plots . . . 85
3.3.2 Repeated Cross-sectional Plots . . . 86
3.3.3 Autocorrelation Plots . . . 87
3.4 Other Types of Data Visualization **89**
3.4.1 Text Visualization . . . 90
3.4.2 Survey Data Visualization . . . 90
3.4.3 Geo-spatial Visualization . . . 91
3.4.4 Network Visualization . . . 93
3.5 Chapter Summary **97**
3.5.1 Problems . . . 98

Probability

4 Basic Probability . . . 103
4.1 Introduction **103**
4.2 Events and Sample Spaces **104**
4.2.1 Probability Axioms . . . 106
4.2.2 Union of Events . . . 106
4.2.3 Intersection of Independent Events . . . 109
4.2.4 Complementary Events . . . 111

4.2.5 Conditional Probability . . . 112
4.3 Calculating Probabilities 115
4.3.1 Sample Point Method . . . 116
4.3.2 Counting Sample Points . . . 117
4.3.3 Combining Events . . . 124
4.4 Bringing It All Together 124
4.4.1 Law of Total Probability . . . 124
4.4.2 Bayes' Theorem . . . 128
4.5 Chapter Summary 131
4.5.1 Problems . . . 133

5 Random Variables . . . 139
5.1 Introduction 139
5.2 Discrete Random Variables 140
5.2.1 Probability Mass Function . . . 140
5.2.2 Cumulative Distribution Function . . . 144
5.2.3 Expected Value . . . 147
5.2.4 Variance and Standard Deviation . . . 150
5.3 Continuous Random Variables 153
5.3.1 Probability Density Function . . . 153
5.3.2 Cumulative Distribution Function . . . 159
5.3.3 Expected Value . . . 161
5.3.4 Variance and Standard Deviation . . . 163
5.4 Expected Value Properties 164
5.5 Variance Properties 166
5.6 Chapter Summary 168
5.6.1 Problems . . . 169

6 Discrete Distributions . . . 173
6.1 Introduction 173
6.2 Binomial Distribution 174
6.3 Geometric Distribution 184
6.4 Negative Binomial 190
6.5 Hypergeometric Distribution 194
6.6 Poisson 200
6.7 Chapter Summary 206
6.7.1 Problems . . . 208

7 Continuous Distributions 211
7.1 Introduction 211
7.2 Uniform 212
7.3 Exponential 216
7.4 Normal 225
7.4.1 Standardizing 230
7.4.2 Bivariate and Multivariate Normal Distributions 233
7.5 Gamma 235
7.6 Distributions Related to the Normal 238
7.6.1 t Distribution 238
7.6.2 Chi-squared (χ^2) 241
7.6.3 F Distribution 243
7.7 Beta 244
7.8 Quantile-Quantile Plots 247
7.9 Chapter Summary 252
7.9.1 Problems 255

Classical Statistical Inference

8 About Data & Data Collection 259
8.1 Introduction 259
8.2 Data and the Scientific Method 260
8.2.1 Experimental vs. Observational Data 264
8.2.2 Convenience vs. Probability Sampling 266
8.3 Important Concepts 267
8.3.1 Data = Signal + Noise 267
8.3.2 Accuracy vs. Precision 267
8.4 Sampling Data 269
8.4.1 Types of Sampling 270
8.4.2 Sources of Bias 270
8.5 Types of Error 270
8.6 Historical Gaffes in Data Collection 272
8.7 Ethics and Statistical Practice 274
8.8 Chapter Summary 275
8.8.1 Problems 276

9 Sampling Distributions 277
9.1 Introduction 277
9.2 Linear Combinations of Random Variables 279
9.3 Sampling Distributions for Sums and Means 284
9.4 Sampling Distribution for the Sample Variance 291
9.5 The Central Limit Theorem 293
9.6 Normal Approximation to the Binomial 293
9.7 Tchebysheff's Theorem and the Law of Large Numbers 298
9.8 Chapter Summary 303

10 Point Estimation 305
10.1 Introduction and Intuitive Estimators 305
10.2 Estimation Criteria 307
10.2.1 Unbiased Estimators 307
10.2.2 Consistent Estimators 309
10.3 Method of Moments 312
10.4 Maximum Likelihood 317
10.5 Approximating MLEs 323
10.6 Sufficiency 327
10.7 Chapter Summary 331

11 Confidence Intervals 335
11.1 Introduction 335
11.2 Basic Properties 336
11.3 Large Sample Confidence Intervals 341
11.4 Small Sample Confidence Intervals 346
11.5 Confidence Intervals for Differences 352
11.5.1 Confidence Intervals for Differences of Proportions 352
11.5.2 Confidence Intervals for Differences in Means 355
11.5.3 Confidence Interval for Paired Data 360
11.6 Determining the Sample Size 362
11.7 Confidence Intervals from Complex Survey Data 368
11.7.1 Sampling from a Finite Population 368
11.7.2 Stratified Random Samples 371
11.7.3 Cluster Sampling 376
11.7.4 Secondary Data Sources 378
11.7.5 Software for Analyzing Data from Complex Surveys 381

11.8 Chapter Summary 383

12 Hypothesis Testing 385
12.1 Introduction 385
12.2 Elements of a Statistical Test 387
12.3 Power 392
12.4 P-values 394
12.5 Testing the Mean: Variance Known 396
12.5.1 Hypothesis Tests for the Mean from a Population with Known Variance 396
12.5.2 Power 398
12.6 Testing the Mean: Variance Unknown 404
12.7 Testing the Proportion 411
12.8 Testing the Variance 411
12.9 Sample Size Determination 411
12.10 Chapter Summary 411

13 Hypothesis Tests for Two or More Samples 413
13.1 Introduction 413
13.2 Testing Independent Samples 413
13.3 Testing Paired Samples 413
13.4 Confidence Intervals for Two Samples 413
13.5 Single-Factor ANOVA 413
13.6 Two-Factor ANOVA 413
13.7 Multi-Factor ANOVA 413
13.8 Introduction to Design of Experiments 413
13.9 Chapter Summary 413

14 Hypothesis Tests for Discrete Data 415
14.1 Introduction 415
14.2 Contingency Tables 415
14.3 Goodness-of-Fit Tests 415
14.4 Chi-square Tests 415
14.5 Fisher's Exact Test 415
14.6 Chapter Summary 415

15 Regression 417
15.1 Introduction 417
15.2 Correlation 417
15.3 The Regression Model 417
15.4 Inference on the Intercept (β_0) and Slope (β_1) Parameters 417
15.5 Prediction Intervals for Future Observations 417
15.6 Multiple Regression Model 417
15.7 Model Checking and Validation 417
15.8 Polynomial and Nonlinear Regression 417
15.9 Chapter Summary 417

Bayesian and Other Computer Intensive Methods

16 Bayesian Methods 421
16.1 Introduction 421
16.2 Bayes Theorem 422
16.3 The Bayesian Paradigm 429
16.4 Three Paradoxes 433
16.5 Conjugate Priors 437
16.6 Noninformative Priors 449
16.7 Simulation Methods 451
16.7.1 Metropolis-Hastings Algorithm 454
16.7.2 The Gibbs Sampling Algorithm 458
16.8 Hierarchical Bayes Models 460

17 Time Series Models 467
17.1 Introduction 467
17.2 Smoothing Models 467
17.3 Regression-based Models 467
17.4 ARMA and ARIMA Models 467
17.5 State-Space Models 467
17.6 Chapter Summary 467

18 The Jackknife and Bootstrap 469
18.1 Introduction 469

18.2 The 'Plug In' Principle 469
18.3 Bootstrap Estimate of the Standard Error 469
18.4 Bootstrap Confidence Intervals 469
18.5 Bootstrap Hypothesis Testing 469
18.6 The Jackknife and the Standard Error 469
18.7 Chapter Summary 469

Advanced Topics in Inference & Data Science

19 Generalized Linear Models and Regression Trees 473
19.1 Introduction 473
19.2 Logistic Regression 473
19.3 Probit and Extreme Value Models 473
19.4 Generalized Linear Models 473
19.5 Poisson Models 473
19.6 Regression Trees 473
19.7 Chapter Summary 473

20 Cross-Validation and Estimates of Prediction Error 475
20.1 Introduction 475
20.2 Prediction Rules 475
20.3 Cross-Validation 475
20.4 Covariance Penalties 475
20.5 Model Training and Validation 475
20.6 Chapter Summary 475

21 Large-Scale Hypothesis Testing and the False Discovery Rate . . . 477
21.1 Introduction 477
21.2 Multiple Testing 477
21.2.1 Issues 477
21.2.2 Bonferroni Corrections 477
21.2.3 Tukey's HSD 477
21.3 Large-Scale Testing 477
21.4 The False Discovery Rate (FDR) 477
21.5 Controlling the FDR 477

21.6 Chapter Summary 477

A More About R 479
A.1 Introduction 479
A.2 Reading Data into R 479
A.3 Managing Your Workspace 479
A.4 Writing Scripts 479
A.5 Writing Functions 479

Index 485

Preface

Data science is a new field that has arisen to exploit the proliferation of data in the modern world. Statistics dates back to the mid-18th century, where the field began as the systematic collection of population and economic data by nations. The modern practice of statistics – which includes the collection, summarization, and analysis of data – dates to the early 20th century. Today statistical methods are widely used by governments, businesses and other organizations, as well as by all scientific disciplines.

In many ways, data science is really just the next step in the development of the discipline of statistics. Traditional statistics arose in an era in which data was hard (and thus expensive) to collect and so traditional statistical methods were created to extract the most information possible from data. Today, with the proliferation of computers, sensors, and the internet, some types of data are cheap and plentiful. This does not mean that the traditional statistical methods are now obsolete – they still have much to contribute to today's data-rich environment – but they do require appropriate application.

This book is for data scientists who want to improve how they work with, analyze, and extract information from data. It is intended for current and future data scientists alike, and for anyone interested in deriving information from data. It requires some mathematical sophistication on the part of the reader, as well as comfort using computers and statistical software, but for data scientists these prerequisites should be a given.

It has been said that a data scientist must have a better grasp of statistics than the average computer scientist and a better grasp of programming than the average statistician. This book will give the data scientist a basic grasp of statistics. It should be the first step on a longer journey to master statistical methods, a field which is both broad and deep. As the Chinese philosopher Laozi said, "A journey of a thousand miles begins with a single step."

Statistics is often viewed by students with either trepidation or distaste. This is unfortunate because applied statistics, which is nothing more than using data to find the answer a question or solve a scientific mystery, can be both interesting and a lot of fun. As Francis Galton said in 1889,

> *Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of Man.*

The focus of this book is how to appropriately apply statistical methods, both simple and sophisticated, to 21st century data and problems. Because differences in assumptions and methods can produce divergent results, data scientists must be statistically savvy enough to understand, interpret, and reconcile ensuing differences in conclusions. Only then can they separate fact from opinion, objective analysis from biased agenda, good data science from bad.

R.D. Fricker, Jr. & S.E. Rigdon
September 2017

Acknowledgements

This book was typeset in LaTeX using a modified version of The Legrand Orange Book template originally created by Mathias Legrand and modified by Vel and the author.

Bayesian and Other Computer Intensive Methods

1 — Bayesian Methods

1.1 Introduction

Up to this point is we have been talking about what are often called *frequentist* methods, because a statistical method is based on properties of its long run relative frequency. With this approach, probability of an event is defined as the proportion of times the event occurs in the long run. Parameters, that is values that characterize a distribution, such as the mean and variance of a normal distribution, are considered *fixed* but *unknown* values. In the frequentist paradigm parameters are fixed, not random.

The Bayesian approach departs from the frequentist approach at the very foundation: the definition of probability. To Bayesians, probability represents degree of belief. Parameters are treated as random variables having some distribution, because they are uncertain, and uncertain things are expressed in probabilistic terms. In essence, uncertainty and randomness are the same in the Bayesian approach. This is not the case for the frequentist approach.

In the frequentist approach, we often collect data from some assumed distribution. We then use this data to find point estimates and confidence intervals for the unknown parameters. These confidence intervals have a rather subtle interpretation. For example a 95% confidence interval for a parameter θ is an interval that has probability 0.95 of containing the true value of θ when viewed before the data were taken. In other words, before we take any data, but after we have chosen a rule for computing the confidence interval, the probability that the resulting interval will include the true value is 0.95. This seems odd to many users of statistics who point out that we have already collected data, computed the relevant statistics, and computed the endpoints of the confidence interval, which we will denote C.

What to make of this computed interval? We cannot say that $P(\theta \in C) = 0.95$, that is, the probability is 0.95 that the interval contains the true value. In the above probability, there is nothing random. θ is not random (at least not in the frequentist approach) and C is not random, since it has been computed from the observed data. (Once data are taken, C will be an interval with fixed endpoints, for example $C = (65, 67.)$ We can only say that the probability that an interval *calculated in this way* will include the true value of θ is 0.95 when viewed before collecting data. If we were

repeatedly computing confidence intervals, 95% of the time we would get an interval that contains the true value of the parameter, and 5% of the time we would miss it. Why would we believe, for example, that our computed interval of (65,67) contains the true value of θ? 95% of the time when we compute such intervals they contain θ, but whether this one does, or doesn't, we do not know and will never know. In the frequentist paradigm, confidence intervals are probability statements about the *method* that produced the confidence interval, and its long run properties.

The Bayesian approach, which treats unknown parameters as random variables, takes a more direct approach. Intervals are calculated and the interpretations are much more along the lines that many practitioners would like. They are direct probabilistic statements that a parameter is contained in the interval, conditioned on the observed data. This directness, however, comes at a price. Our interpretation of probability must change to the subjective, or degree of belief, interpretation, and we must specify what we believe about a parameter before we observe any data.

1.2 Bayes Theorem

We saw Bayes Theorem in section (cite), but we summarize the result here and explain how it forms the basis for a new system of statistical inference.

Suppose that the set of all possible outcomes for some random phenomenon, say taking a random sample from a population, is equal to the set S. Suppose also, that the set S can be partitioned into sets $A_1, A_2, \ldots, A_k$ where these sets are *mutually exclusive* and *exhaustive*. Mutually exclusive means that if we take any two distinct sets, their intersection is the empty set. Exhaustive means that if we union up all of the A_i's we end up with the entire sample space S; that is

$$\bigcup_{i=1}^{k} A_i = S.$$

This means that every possible outcome is in one, and only one, of the sets $A_1, A_2, \ldots, A_k$.

Now let B be any set that is contained in the sample space S. Then every element in B will be in one and only one of the sets

$$A_1 \cap B, \;\; A_2 \cap B, \;\; \ldots, \;\; A_k \cap B.$$

Since these events are mutually exclusive, we can write

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_k \cap B).$$

We can apply the laws of conditional probability to each term on the right side to obtain the Law of Total Probability:

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_k)P(B|A_k). \tag{1.1}$$

This leaves us just short of Bayes Theorem, which again is obtained from the definition of conditional probability. Bayes Theorem says

$$\begin{aligned} P(B|A_j) &= \frac{P(A_j)P(B|A_j)}{P(B)} \\ &= \frac{P(A_j)P(B|A_j)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_k)P(B|A_k)} \\ &= \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)}. \end{aligned} \tag{1.2}$$

Notice how Bayes Theorem *turns* conditional probabilities around. On the right side we have $P(A_i|B)$ and on the right we have $P(B|A_i)$. The Bayesian approach assumes $f(x|\theta)$ as a model for the distribution of the observation x (which is often a vector). What Bayes theorem tells us is how to compute the distribution of the parameter θ given the data, effectively swapping the events in the conditional probability.

■ **Example 1.1 — Disease Testing.** Tests for a disease are never perfect. You could have the disease, but test negative; you could also be disease free and still test positive. The probability of testing positive given that you have the disease is called the *sensitivity*, and the probability that you test negative given that you don't have the disease is called the *specificity*. Suppose that for a particular test, the sensitivity is 95% and the specificity is 90%. It is known that 1% of the population has the disease. Find the following:

1. the probability that someone selected at random tests positive, and
2. the probability that someone who tests positive actually has the disease.

Solution First, notice that the sensitivity and specificity are defined as conditional probabilities. Let D denote the event that someone has the disease and ND denote the event that someone does not have the disease; also, let "+" denote the event that someone tests positive and "−" denote the event that someone tests negative. Then the sensitivity and specificity are

$$\text{sensitivity} = P(+|D)$$

and

$$\text{specificity} = P(-|ND).$$

The probability of having the disease is an *unconditional* probability, and is called the *base rate*. The first part is answered by applying the Law of Total Probability. Here A_1 is the event that someone has the disease and A_2 event that the person does not have the disease. These events are clearly mutually exclusive and exhaustive. B is the event that someone tests positive. Thus

$$\begin{aligned} P(+) &= P\left((+ \text{ and } D) \text{ or } (+ \text{ and } ND)\right) \\ &= P(+ \text{ and } D) \ + \ P(+ \text{ and } ND) \\ &= P(D)P(+|D) \ + \ P(ND)P(+|ND) \end{aligned} \tag{1.3}$$

Remember that the logical "or" means the set union and the logical "and" means the set intersection. Each of the terms in this last expression is either a given in the problem, or can be obtained using the complement rule. For example,

$$P(ND) = 1 - P(D)$$

and

$$P(+|ND) = 1 - P(-|ND) = 1 - \text{ specificity}.$$

With this, everything in equation (16.3) is known. While the calculations could be done easily on a calculator, we show how this could be done in R.

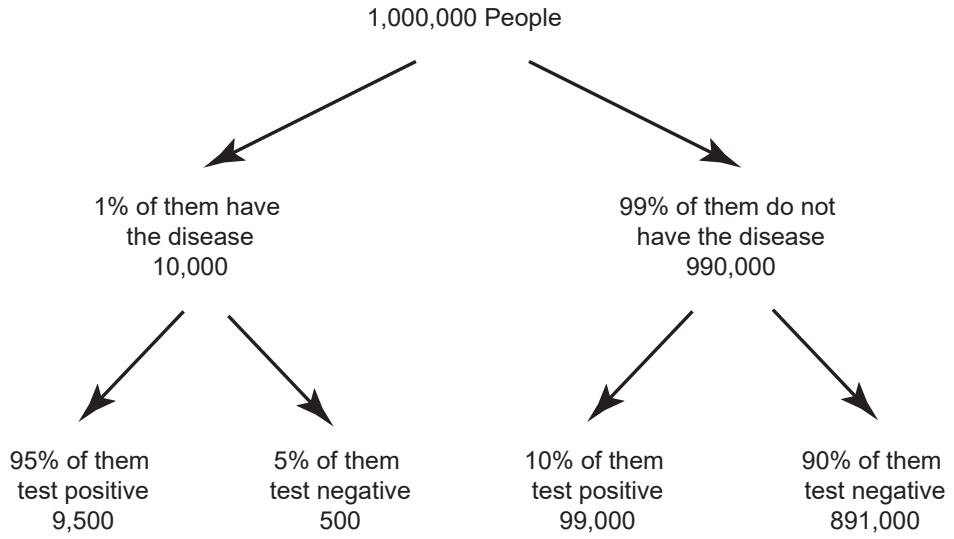


Figure 1.1: cap

```
sensitivity = 0.95
specificity = 0.90
baseRate = 0.01
probPositive = baseRate*sensitivity + (1-baseRate)*(1-specificity)
print(probPositive)
```

The result is 0.1085. Much of the work in answering the second part is already done once we have the probability of testing positive. Using Bayes theorem, we get

$$P(D|+) = \frac{P(D)P(+|D)}{P(+)} = \frac{0.0095}{0.1085} \approx 0.08756.$$

■

A few things about the solution are worth noting. First, about 11% of the population will test positive for the disease, even though only 1% have the disease. Second, if someone tests positive for the disease, the probability of actually having the disease is only 0.08756. Think about the importance of this. Fewer than one person in ten who tests positive will have the disease. The value of $P(D|+)$ is called the *positive predictive value* (PPV) and the value of $P(ND|-)$ is called the *negative predictive value* (NPV). The PPV is very low even though the sensitivity and specificity are fairly high (0.95 and 0.90, respectively). This can be explained by the low base rate and the moderate probability (0.10) of those without the disease testing positive. Figure 16.1 shows what would be expected in a population of 1,000,000 people. The tree first splits individuals on whether they do or don't have the disease. The second splits individuals on whether they test positive or negative. Thus, about $9,500 + 99,000 = 108,500$ people will test positive, and among these, only 9,500 will have the disease, a fraction of $9,500/108,500 = 0.08756$. The tree clearly shows why the PPV is so low: the number of false positives (99,000) is much greater than the number of true positives (9,500).

The second thing to note about this example is how it relates to the Bayesian approach to statistical inference. Think of the base rate as the prior probability of a person having the disease.

If someone is selected at random from the population, we would, knowing nothing else, put the probability of having the disease at 0.01. Think of the result of the test for the disease as *data*. We use Bayes theorem to transform our assessment of how likely it is that a person has the disease as we move from *before* observing data to *after* observing data. We use the respective terms *prior* and *posterior* to denote our assessment before and after observing data. Note that if a person tests positive, our belief that a person has the disease has gone from 0.01 (the base rate) to 0.08756 (the PPV). Had the test been negative, the probability would have gone from 0.01 to 0.000561; details are left to the reader.

■ **Example 1.2** Suppose that someone has three coins. One coin has two tails, so that it will end up tails every time it is tossed. The second coin is fair, meaning that the probability of being heads is 0.5 (same for tails). The third coin is a two-headed coin, so the it will always come up heads. One coin is selected at random, where each coin is equally likely, and tossed twice. We do not get to observe which coin was selected; we only see the results of the two tosses. Find the following:

1. the probability that both tosses result in heads,
2. the probability that the coin is the two-tailed coin given that both tosses result in heads,
3. the probability that the coin is the fair one given that both tosses result in heads,
4. the probability that the coin is the two-headed coin given that both tosses result in heads.

Solution To make the notation simpler, and to make this example a stepping stone to understanding the Bayesian approach, let θ denote the "head" probability for the selected coin. The possible values for θ are 0, $\frac{1}{2}$, and 1, and each of these occurs with probability $\frac{1}{3}$. We can summarize this prior distribution for θ in the following table:

θ	0	$\frac{1}{2}$	1
$\pi(\theta)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Conditioned on the value of θ the number of heads on two tosses then has a binomial distribution with $n = 2$ and probability of success θ. This is a conditional probability distribution, so we write $X|\theta \sim \text{BIN}(2,\theta)$. To find the probability of getting exactly two heads, we apply the law of total probability:

$$\begin{aligned} P(X=2) &= P(\theta=0)P(X=2|\theta=0)+P(\theta=\frac{1}{2})P(X=2|\theta=\frac{1}{2}+P(\theta=1)P(X=2|\theta=1) \\ &= \frac{1}{3}\times 0+\frac{1}{3}\times\left(\frac{1}{2}\right)^2+\frac{1}{3}\times 1 \\ &= \frac{5}{12}. \end{aligned}$$

The next three parts are obtained by applying Bayes theorem:

$$\begin{aligned} P(\theta=0|X=2) &= \frac{\pi(0)\ \Pr(X=2|\theta=0)}{\Pr(X=2)} = \frac{\pi(0)\times 0}{5/12} = 0 \\ P(\theta=\frac{1}{2}|X=2) &= \frac{\pi(\frac{1}{2})\ \Pr(X=2|\theta=\frac{1}{2})}{\Pr(X=2)} = \frac{\pi(\frac{1}{2})\times\frac{1}{4}}{5/12} = \frac{1}{5} \\ P(\theta=1|X=2) &= \frac{\pi(1)\ \Pr(X=2|\theta=1)}{\Pr(X=2)} = \frac{\pi(1)\times 1}{5/12} = \frac{4}{5} \end{aligned}$$

Given that $X = 2$ heads were observed, the posterior distribution for the parameter θ is

θ	0	$\frac{1}{2}$	1
$\pi(\theta\|X=2)$	0	$\frac{1}{5}$	$\frac{4}{5}$.

■

Again, a few important points should be made regarding this example. First, the posterior distribution is a valid (discrete) probability distribution, since all probabilities are nonnegative and sum to 1. Second, the probability that $\theta = 0$ is equal to 0, because if $\theta = 0$ then a head has probability 0 of occurring. Thus on two tosses of the coin we would *always* observe $X = 0$ heads, so getting $X = 2$ is impossible Third, if we observe both heads, we might believe it more likely that $\theta = 1$, that is, we have the two-headed coin. We might reason that if $\theta = \frac{1}{2}$, that is we selected the fair coin, then two heads is somewhat unlikely. Bayes theorem tells us exactly how we should change our minds about the value of θ from before data are take to after, when we observe $X = 2$ heads.

■ **Example 1.3** Expand the previous experiment to one where the probability of getting a head is any number in

$$\frac{0}{100}, \frac{1}{100}, \frac{2}{100}, \cdots, \frac{100}{100},$$

where each of these 101 numbers is equally likely. As in the previous example once a coin is selected, it is tossed twice. This might arise in a situation where we are uncertain of the probability of some event, and we have the outcomes of two trials on which to estimate this probability. (This is a rather trivial problem, but in just a bit we will give some more realistic numbers.) Before taking any data, any number (rounded to a multiple of $\frac{1}{100}$) between 0 and 1 is assumed to be equally likely. We then assign each of these 101 numbers the same prior probability, namely $\frac{1}{101}$. Apply the same reasoning as in the previous example to find the posterior probability for θ.

Solution In principle, the approach is the same, just messier. We could run the following code in R to obtain and plot the posterior:

```
n = 2
x = 2
theta = (0:100)/100
prior = rep(1/101,101)
likelihood = dbinom(x,n,theta)
evidence = sum(prior*likelihood)
posterior = prior*likelihood/evidence
plot( theta , posterior , xlab="" , ylab="Posterior Probability" , col="blue" )
```

The plot produced is shown in Figure 16.2. ■

Note that the most likely value of θ after observing two successes on two trials is $\theta = 1$, but even the most likely value is still rather unlikely. The posterior probability is approximately 0.03, which is small, but still the largest posterior probability.

To see what happens when the numbers are larger, consider the next example.

■ **Example 1.4** Suppose we had $X = 36$ successes out of $n = 100$ trials. This could occur if we ran a clinical trial to test whether a medication was helpful in relieving pain. (This is a rather bad design, because there is no control group, but we will address this issue later.) Find the posterior probability distribution for the probability θ.

Solution Apply a similar code to what was used in the previous example:

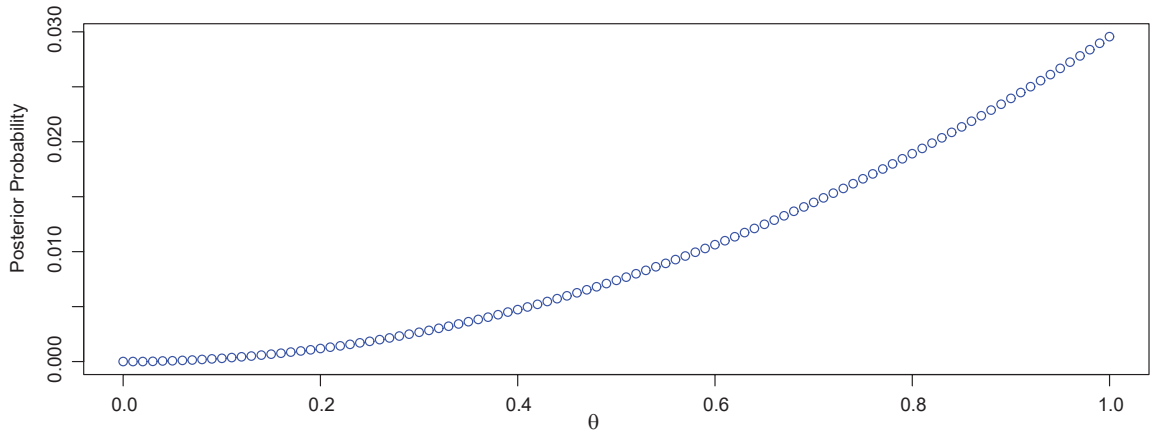


Figure 1.2: The (discrete) posterior distribution for the probability of success given $X = 2$ successes on $n = 2$ trials.

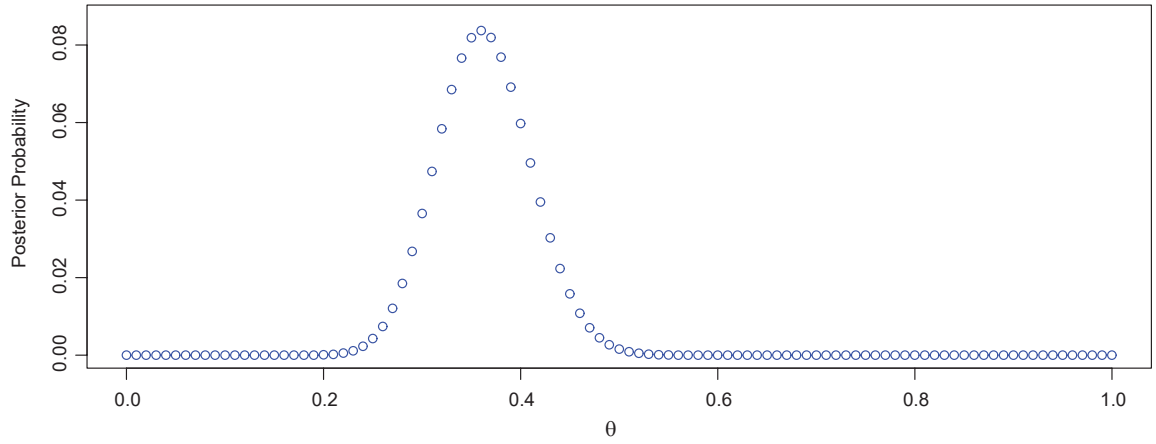


Figure 1.3: The (discrete) posterior distribution for the probability of success given $X = 2$ successes on $n = 2$ trials.

```
x = 36
n = 100
theta = (0:100)/100
prior = rep(1/101,101)
likelihood = dbinom(x,n,theta)
evidence = sum(prior*likelihood)
posterior = prior*likelihood/evidence
plot( theta , posterior , xlab="" , ylab="Posterior Probability" , col="blue" )
```

The resulting plot of the posterior is shown in Figure 16.3. From this plot we can see that values between about 0.25 and 0.45 have a fairly high posterior probability, but values outside of this interval have posterior probability near zero. ∎

The major takeaway from this example, is how we are interpreting probability. We can talk, for

example, about the posterior probability of θ (which is equal to the probability of a success) being equal to 0.35 (which is about 0.08) or 0.70 (which is virtually 0.00). Such statements cannot be made in the frequentist or classical paradigm because parameters are considered fixed (not random) quantities. Therefore no probabilistic statements can be made about parameters, either before or after data are collected.

The second issue that one might raise about the previous example is the discreteness imposed on the parameter θ. Is it really reasonable to say that θ could be equal to 0.35, or 0.36, but θ could not be 0.355? It seems more reasonable to say that θ could be *any* value between 0 and 1. This means that we would put a *continuous* prior distribution on θ; we would then observe a *discrete* random variable with conditional distribution $X \sim \text{BIN}(100, \theta)$. The posterior would then be computed from the continuous analogue of Bayes theorem.

Consider the general case of observing X successes on n trials where the probability of success on any trial is θ. The conditional distribution of $X|\theta$ is then

$$X|\theta \sim \text{BIN}(n, \theta)$$

Let $\pi(\theta)$ denote the prior distribution, and let

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n$$

For continuous parameters, Bayes theorem becomes

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int_0^1 \pi(\theta) f(x|\theta)\, d\theta}. \tag{1.4}$$

Notice how the sum in the denominator of Bayes theorem becomes an integral (the limit of sums) when we move to a continuous prior.

■ **Example 1.5** Suppose we observed $x = 36$ successes on $n = 100$ trials as in the previous example. If we were to take the prior on θ to be uniform on $[0, 1]$, then the posterior would be

$$\begin{aligned} \pi(\theta|x) &= \frac{1 \binom{100}{36} \theta^3 6 (1-\theta)^{100-36}}{\int_0^1 \pi(\theta) f(x|\theta)\, d\theta} \\ &= c(x)\, \theta^{36} (1-\theta)^{100-36}, \qquad 0 \le \theta \le 1. \end{aligned}$$

In the next section, we explain why the posterior distribution is the BETA(37,65) distribution. We can use the following R code to plot this posterior.

```
x = 36
n = 100
theta = (0:200)/200
prior = 1
posterior = dbeta( theta , 37 , 65 )
plot( theta , posterior , type="l" , xlab="" , lwd=2 ,
      ylab="Posterior Probability" , col="blue" )
```

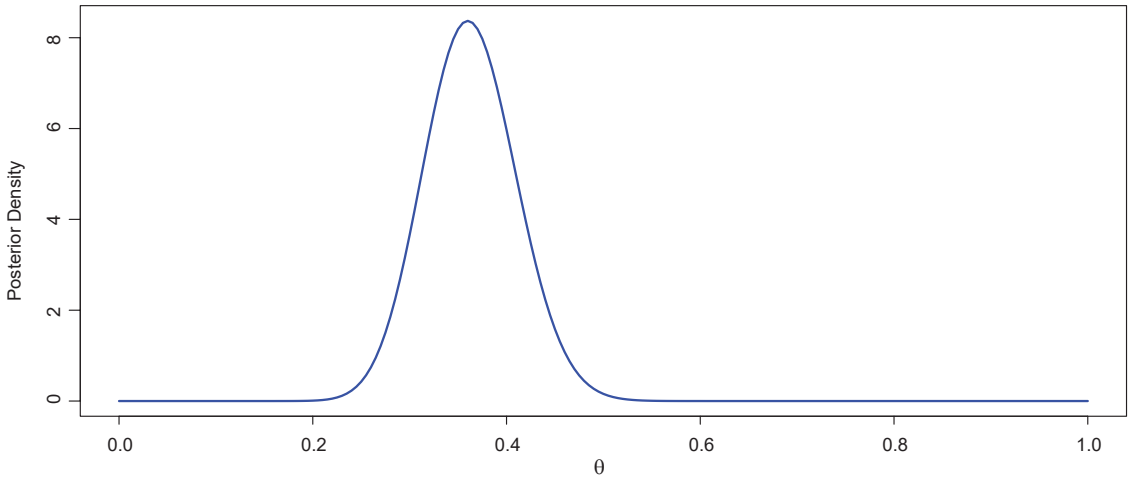


Figure 1.4: The (continuous) posterior distribution for the probability of success given $X = 2$ successes on $n = 2$ trials.

The plot of the posterior is shown in Figure 16.4. ∎

Notice how similar this plot is to the plot in Figure 16.3. There is little practical difference between assuming a discrete uniform prior on $0, 0.01, 0.02, \ldots, 0.99, 1.00$ and a continuous prior on the interval $[0, 1]$. One important difference, however is the vertical axis. In the former case, the vertical axis is the posterior probability of θ being equal to a particular value; here the sum of the posterior probabilities must be one. In the latter, it is the probability density, and it must integrate to one.

1.3 The Bayesian Paradigm

In the Bayesian paradigm, unknown parameters are treated as *uncertain*, and Bayesians treat uncertainty in probabilistic terms. In other words, if we don't know the value of a parameter, we express our knowledge (or ignorance) in a probability distribution. Suppose θ is an unknown parameter (which could be a vector) and x is the data (usually a vector) whose distribution depends on the parameters. The Bayesian approach to statistics is, in principle, simple:

- assess what we believe about θ *before* observing any data, which is done through a probability distribution $\pi(\theta)$ called the *prior distribution*;
- apply Bayes theorem to obtain the conditional distribution of θ given the data x:

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int \pi(\theta) f(x|\theta)\, d\theta}$$

  which is called the *posterior distribution*.

This two-step process is often called "turning the Bayesian crank." While in principle, the process is straightforward, the computation of the denominator in Bayes theorem is usually a problem except in simple cases. Monte Carlo simulation can be used to learn about the posterior in more complicated cases. We discuss this in Section 16.7.

The integral in the denominator is the integral over the entire θ space. For a scalar value, this will be an interval of the real line. For example, in the last example θ represented the unknown

probability of a success, which must be in $[0,1]$. In other cases it might be $[0,\infty)$ or $(-\infty,\infty)$. If θ is a vector, then the integral will be a multiple integral whose dimension is equal to the dimension of θ. We will ordinarily write the integral with a single integral sign, whether it is a single or multiple integral, with the understanding that its dimension matches the dimension of θ.

The posterior distribution reflects what we believe about the parameter θ given the observed data. Notice how this is the reverse of the likelihood function $f(x|\theta)$. (Recall that Bayes theorem is all about swapping the events in a conditional probability.)

If we wanted a point estimate for θ, we could take some measure of the center of the posterior. Often the posterior mean

$$\hat{\theta} = E(\theta|\boldsymbol{x}) = \frac{\int \theta\pi(\theta)f(\boldsymbol{x}|\theta)\,d\theta}{\int \pi(\theta)f(x|\theta)\,d\theta} \tag{1.5}$$

is used as a point estimate. Sometimes the posterior median or mode is used as a point estimate for θ. If an interval estimate of a parameter is desired, then we could determine an interval I for the case of a scalar parameter, or a region R for the case of a vector parameter, with the property that

$$P(\theta \in I|\boldsymbol{x}) = 1-\alpha. \tag{1.6}$$

Such an interval is called a *credible interval*.

By contrast, the classical approach determines a random interval $I(\boldsymbol{X}) = (L,U)$ with the property that

$$\Pr(\theta \in I(\boldsymbol{X})) = \Pr(L<\theta<U) = 1-\alpha. \tag{1.7}$$

These two formulations look similar, but they differ in a fundamental way. In (16.6), θ is the random variable, and $\boldsymbol{x}$ is considered fixed, whereas in (16.7) the interval I is a random variable, depending on the random vector X, and θ is considered fixed.

■ **Example 1.6** For the data ($n=100,\ x=36$) from the previous example, find the posterior mean, median, and mode. Assume a (continuous) uniform prior over the interval $[0,1]$. Also, find a 95% credible interval for θ.

Solution: Recall a few properties of the beta distribution. The PDF is

$$f(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1}, \qquad 0<\theta<1. \tag{1.8}$$

The mean of the BETA(a,b) distribution is

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

and the mode is

$$\text{mode} = \frac{a-1}{a+b-2}.$$

Given a uniform prior distribution for θ, which is a BETA$(1,1)$ distribution, we find that the posterior distribution is

$$p(\theta|x) = \frac{\pi(\theta)\times\binom{100}{36}\theta^{36}(1-\theta)^{100-36}}{\displaystyle\int_0^1 \pi(\theta)\times\binom{100}{36}\theta^{36}(1-\theta)^{100-36}d\theta}.$$

Note that the denominator is a constant, say $1/c$. In general the constant will depend on the data but not the parameter. The posterior is therefore

$$p(\theta|x) = 1 \times c \times \binom{100}{36} \theta^{36}(1-\theta)^{100-36} = c_1 \theta^{37-1}(1-\theta)^{65-1}, \qquad 0 < \theta < 1.$$

We recognize this as a beta distribution with parameters $a = 37$ and $b = 65$.

There is no simple formula for the median, but it can easily be obtained by using the beta quantile function `qbeta` in R.

Since $\theta \mid (x = 36) \sim \text{BETA}(37, 65)$, the posterior mean is

$$\hat{\theta}_{\text{mean}} = E(\theta|x = 36) = \frac{37}{37+65} \approx 0.363$$

and the mode is

$$\hat{\theta}_{\text{mode}} = \frac{37-1}{37+65+2} = \frac{36}{100} = 0.36$$

The median can be found using the `qbeta` function in R. For example,

```
median = qbeta( 0.5 , 37 , 65 )
```

This yields

$$\hat{\theta}_{\text{median}} = 0.362.$$

You may be wondering why all three estimates are greater than the usual point estimate of $\frac{36}{100} = 0.36$. The prior puts uniform prior across the entire interval from 0 to 1. Thus half of the prior probability is for $\theta > \frac{1}{2}$ and half for $\theta < \frac{1}{2}$. Thus, the posterior mean will be pulled slightly toward the prior mean, which is $\frac{1}{2}$. In this case, the point estimates are pulled very slightly from 0.36 toward 0.50.

To get a credible interval for θ, we find the 2.5 and 97.5 percentiles of the posterior distribution using R's `qbeta` function:

```
lowerLimit = qbeta( 0.025 , 37 , 65 )
upperLimit = qbeta( 0.975 , 37 , 65 )
```

This yields the credible interval $(0.273, 0.458)$. The two shaded tails in Figure 16.5.

■

To summarize, here are the important quantities, terminology, and notation for Bayesian statistics.

- $\pi(\theta)$ is the *prior* distribution, which reflects what we believe about the parameter θ *prior* observing any data.
- $f(\boldsymbol{x}|\theta)$ is the PDF (or probability mass function if the observed data are discrete) of the data $\boldsymbol{x}$ given the parameter θ. In classical statistics, the likelihood is often written as $L(\theta|\boldsymbol{x})$, or simply $L(\theta)$. Thus, the likelihood function is functionally the same as the point PDF; that is, $L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$. The difference, to the extent there is one, is that in the joint PDF $f(\boldsymbol{x}|\theta)$, we think of the parameter θ as *fixed* while the data $\boldsymbol{x}$ is the variable. Conversely, in the likelihood function $L(\theta|\boldsymbol{x})$ we think of the data $\boldsymbol{x}$ as fixed, and the parameter θ as the variable. In other words, the difference is a matter of perspective. Bayesians often ignore this difference and write $f(\boldsymbol{x}|\theta)$.

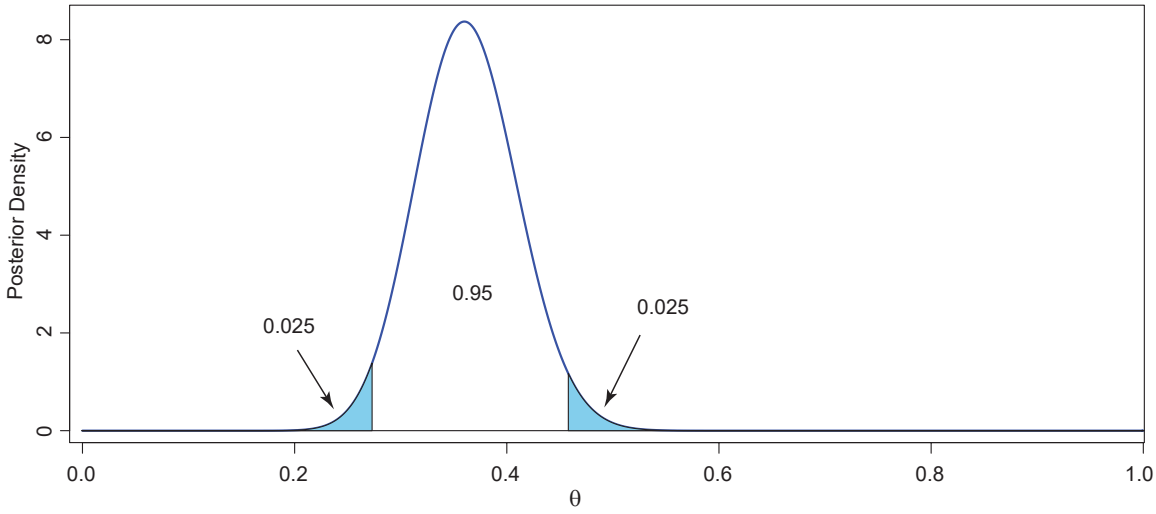


Figure 1.5: The posterior distribution for the probability of success given $X = 2$ successes on $n = 2$ trials, showing the equal-tail credible interval.

- $\pi(\theta|\boldsymbol{x})$ is the *posterior distribution*, which says what we believe about θ given the observed data $\boldsymbol{x}$. Bayes theorem tells us how to change our minds, going from the prior to the posterior once we have observed data. Bayes theorem says

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta) f(\boldsymbol{x}|\theta)}{g(x)}.$$

- The denominator in Bayes theorem, $g(\boldsymbol{x})$ is called the *evidence*. It is the unconditional probability density of observing the data $\boldsymbol{x}$. The evidence is obtained by integrating out the parameter θ in the joint PDF of $\boldsymbol{x}$ and θ:

$$g(\boldsymbol{x}) = \int h(\boldsymbol{x}, \theta)\, d\theta = \int \pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta$$

Note that Bayes theorem guarantees that the posterior integrates to one, as it must if it is to be a valid PDF, because

$$\int \pi(\theta|\boldsymbol{x})\, d\theta = \int \frac{\pi(\theta) f(\boldsymbol{x}|\theta)}{\int \pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta} = \frac{\int \pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta}{\int \pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta} = 1$$

Sometimes, mostly for simple problems, we can determine the family to whoch the posterior distribution belongs, without having to determine $g(x)$. For example, if $\theta \sim$UNIF$(0,1)$ and $X|\theta \sim$BIN(n,θ), then the posterior distribution is

$$\pi(\theta|\boldsymbol{x}) = \frac{1\binom{n}{x}\theta^x(1-\theta)^{n-x}}{g(\boldsymbol{x})} = c(\boldsymbol{x})\theta^{x+1-1}(1-\theta)^{n+1-x-1}, \qquad 0 \le \theta \le 1$$

where $c(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ only. It does not depend on θ; this is important because it sometimes allows us to see the family of distributions from just the part that involves θ. The part of the posterior

above that depends on just the parameter θ is called the *kernel* of the density. (Note that the kernel is not unique, so it might be better to call this "a" kernel, rather than "the" kernel.) In the above case the kernel is

$$\theta^{x+1-1}(1-\theta)^{n+1-x-1}$$

In order to make this integrate to one, the constant must be

$$\left(\int_0^1 \theta^{x+1-1}(1-\theta)^{n+1-x-1}\,d\theta\right)^{-1} = \left(\frac{\Gamma(x+1)\Gamma(n+1-x)}{\Gamma(n+2)}\right)^{-1} = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n+1-x)}.$$

The posterior PDF is therefore

$$p(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n+1-x)}\theta^{(x+1)-1}(1-\theta)^{(n+1)-x-1}, \qquad 0<\theta<1.$$

This makes it clear that the posterior must be the BETA$(x+1, n+1-x)$ distribution. Often, however, we can tell what the posterior is by looking at just the kernel. Here we would notice that $\theta^{x+1-1}(1-\theta)^{n+1-x-1}$ is the kernel of the BETA$(x+1, n+1-x)$ distribution, from which we would conclude that $\theta|\boldsymbol{x} \sim$BETA$(x+1, n+1-x)$.

1.4 Three Paradoxes

In this section we describe some paradoxes involving Bayesian statistics, or the connection between Bayesian and classical, or frequentist methods. We give one example from Bayesian statistics, and two when frequentist methods applied to a problem. We give both Bayesian and frequentist solutions and indicate the inconsistencies with the frequentist approach.

The first paradox involves the nature of what is often called a noninformative prior, that is a prior that purportedly admits total ignorance about a parameter.

■ **Example 1.7** Suppose that $X|\theta \sim$BIN(n,θ) where θ is an unknown parameter. If, *a priori*, we know absolutely nothing about θ we could place a UNIF$(0,1)$ prior on θ. Does this mean that we know nothing about the log odds $\log(\theta/(1-\theta))$?

Solution Let $\phi = \log(\theta/(1-\theta))$. If the prior is UNIF(0,1), then the CDF for ϕ is

$$\begin{aligned} F(x) &= P(\phi < x) \\ &= P(\log(\theta/(1-\theta)) < x) \\ &= P\left(\theta < \frac{\exp(x)}{1+\exp(x)}\right) \\ &= \frac{\exp(x)}{1+\exp(x)}. \end{aligned}$$

The last line follows because the CDF of the UNIF(0,1) distribution is

$$F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x \ge 1. \end{cases}$$

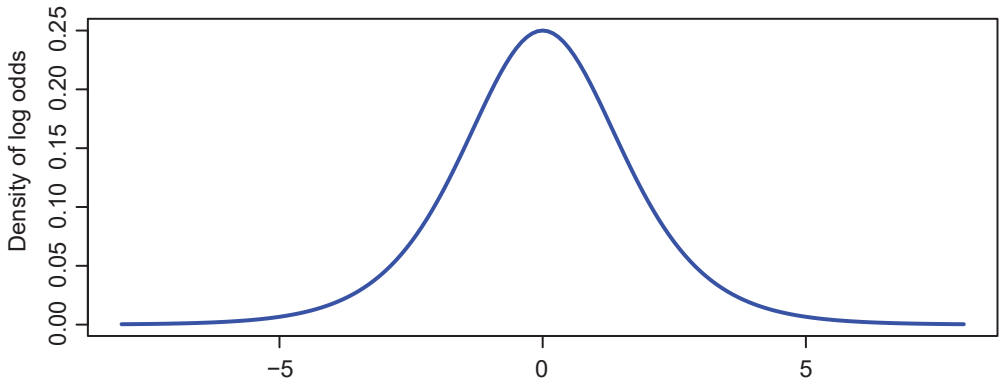


Figure 1.6: The prior distribution for the log odds $\log(\theta/(1-\theta))$ given a uniform distribution for θ.

Note that the quantity $\exp(x)/(1+\exp(x))$ is *always* between 0 and 1, so the middle condition in the CDF above always holds. The PDF of ϕ is thus

$$f_\phi(x) = F'(x) = \frac{\exp(x)}{(1+\exp(x))^2}.$$

The graph of the PDF for the log odds $\phi = \log(\theta/(1-\theta))$ is shown in Figure 16.7. We might think that if we were completely ignorant about the value of θ, we should be equally ignorant about the log odds of θ, but this is not the case. The PDF shows that values of $\phi = \log(\theta/(1-\theta))$ below -5 or above 5 are unlikely, and values near 0 are most likely. A uniform prior in one parameterizations does not imply a uniform prior in another parameterization. Care must be given to the choice of a parameterization in which we would like to supply a noninformative prior. We will discuss this more in section 16.6.

■

The second paradox involves how, or even whether, the design for data collection affects the inference made.

■ **Example 1.8** Suppose that we want to test whether a pain reliever is more effective than a placebo. We select seven people at random and give them either the pain reliever or a placebo, chosen at random. After a few days, the person is given the other; for example, if they received the placebo first, they would get the real pain killer the second time, and vice-versa. Each person is asked which drug worked better. Participants had to select one drug or the other; ties were not allowed. We would like to know whether there is evidence that the pain killer works better than a placebo. In other words, if θ is the probability that someone would select the real pain killer (over the placebo) we would like to know whether $\theta = \frac{1}{2}$ or whether it is larger than $\frac{1}{2}$, so we set up the null and alternative hypotheses as $H_0 : \theta = \frac{1}{2}$ and $H_a : \theta > \frac{1}{2}$. Notice that this problem is exactly equivalent to tossing a coin and testing whether it is fair, or whether the probability of a head is greater than $\frac{1}{2}$. The outcomes of the seven cases is

D, D, D, D, D, D, P

where D indicates the drug was preferred, and P indicates the placebo was preferred.

Solution The test statistic here would be $X =$ number of times the drug was preferred out of the seven trials. Under the null hypothesis, $X \sim \text{BIN}(7, \frac{1}{2})$. The P-value would then be

$$\begin{aligned} P\text{-Value} &= P\left(X = 6|\theta = \frac{1}{2}\right) + P\left(X = 7|\theta = \frac{1}{2}\right) \\ &= \binom{7}{6}\left(\frac{1}{2}\right)^6\left(1-\frac{1}{2}\right)^{7-6} + \binom{7}{7}\left(\frac{1}{2}\right)^7\left(1-\frac{1}{2}\right)^{7-7} \\ &= 0.0625 \end{aligned}$$

With a P-value exceeding the standard $\alpha = 0.05$ we would fail to reject $H_0 : \theta = \frac{1}{2}$. There is not enough evidence to say that the new pain killer is more effective than a placebo. ∎

∎ **Example 1.9** Suppose now that the experimental design from the previous example was a bit different than that described. Suppose that, instead of selecting seven people to participate, we continued experimenting, one person after another, until we got one person who preferred the placebo. The observed data were

D, D, D, D, D, D, P

exactly as before, but the experimental protocol was different. Test whether the drug is effective.

Solution Now, the test statistic is $Y =$ number of trials needed to get one person to respond "placebo." Thus, $Y|\theta$ has a geometric distribution with parameter θ. In this case "as extreme or more extreme" means requiring 7 or more trials to get one "placebo." The P-value is now

$$\begin{aligned} P\text{-Value} &= P\left(Y \geq 7|\theta = \frac{1}{2}\right) \\ &= 1 - P\left(Y \leq 6|\theta = \frac{1}{2}\right) \\ &= 0.0078125 \end{aligned}$$

This P-value is well below 0.05, so we would reject $H_0 : \theta = \frac{1}{2}$. There *is* evidence that drug is effective. ∎

So, if the plan is to test seven people, and we six Ds followed by one P, then we have no evidence that the drug is effective. But if the plan is to test until we get one P, and it takes seven trials for this to occur, then we do have evidence that the drug is effective. But the data are exactly the same in both cases. Shouldn't the data itself determine whether there is or isn't evidence of the drug working? This is the *paradox* (a seeming contradiction).

∎ **Example 1.10** Apply a Bayesian analysis to the problems in Examples 16.8 and 16.9.

Solution Let's take for θ a uniform prior across the interval $[0,1]$. For the case of a fixed number of trials, the likelihood is

$$f(x|\theta) = \binom{7}{x}\theta^x(1-\theta)^{7-x}, \qquad x = 0,1,\ldots,7;\ 0 \leq \theta \leq 1.$$

The posterior is therefore

$$\begin{aligned} \pi(\theta|x) &= c(x)1 \times \binom{7}{x}\theta^x(1-\theta)^{7-x} \\ &= c_1(x)\ \theta^{x+1-1}(1-\theta)^{7+1-x-1}, \qquad 0 < \theta < 1. \end{aligned}$$

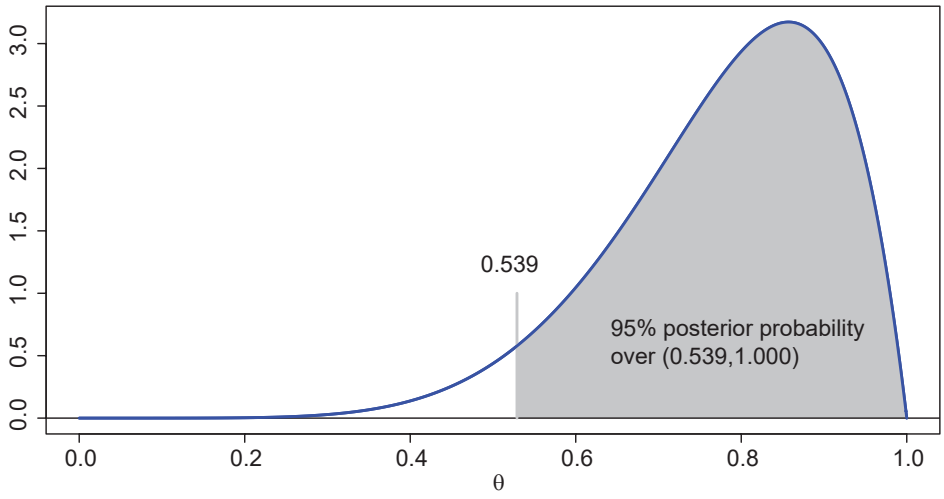


Figure 1.7: The posterior distribution for the probability of success.

Here we combined $c(x)$ and $\binom{7}{x}$ into the new constant $c_1(x)$. We can see that this is the kernel of the BETA$(x+1, 8-x)$ distribution, which for $x = 6$ is the BETA(7,2) distribution.

Now consider the second scenario where the experiment terminated after the first "placeblo." Here we observe $Y =$ number of trials needed to get one "placebo." Thus $Y \sim \text{GEOM}(\theta)$, so the likelihood function is

$$f(y|\theta) = (1-\theta)^{y-1}\theta, \qquad y = 1, 2, \ldots.$$

With a uniform prior on θ as before, we find that the posterior is

$$\pi(\theta|y=7) = c \times 1 \times (1-\theta)^{7-1}\theta.$$

We recognize this as the kernel of the BETA$(7,2)$ distribution. This is the same conclusion that we reached when we assumed a fixed number of trials. ∎

Thus, for the Bayesian approach, it doesn't matter which experimental plan (fixed number of trials or fixed number of "placebos") is selected. The posterior distribution is exactly the same. Note that in both cases the posterior distribution is shown in Figure 16.7. A one-sided credible interval is $(0.539, 1)$ which lies entirely to the right of 0.5, giving evidence that the drug is effective. The frequentist approach had to account for the plan for data collection, whereas the Bayesian approach involved the data only.

This type of example has broader implications about stopping an experiment. For example, if we are running a clinical trial hoping to find evidence of a positive effect, can we put the trial on hold while we do a preliminary analysis of the data? Does this affect the inference? From the frequentist paradigm, yes, this matters. If we want to test the null hypothesis of no effect, and if we allow ourselves two chances to reject H_0, then these must be taken into account when determining α. If we don't account for this and select a decision rule for a fixed α, applying it in the interim analysis, and in the final analysis, then our true α will be a bit higher than we would like, because we are allowing two, not one, opportunities to reject H_0. It is possible to come up with a scenario (see Efron and Hastie, 2016, pp. 31-32) where there was insufficient evidence to reject H_0 when we peeked

at the data, but (assuming we don't account for the two-stage test) there was enough evidence to reject at the planned end of the experiment. But, had we accounted for the two-stage nature of the test, the true α is inflated to the point where we could no longer reject H_0 at the planned end. Had we not peeked, the end result would be significant. Because we peeked at the data, the result is not significant. But why should the act of peeking at the data, which doesn't change the data at all, change the conclusion? In Bayesian statistics, only the data matters, so peeking at the data has no effect.

The third paradox is due to Edwards (1992) and is described in Efron and Hastie (2016, pp. 30-31). Here we adapt their presentation.

■ **Example 1.11** Suppose an inspector measures voltages on a batch of 6 items. The measurements are

$$90, 92, 99, 89, 91, 91$$

yielding a sample average of $\bar{x} = 92$. Assuming that the population of voltages is normally distributed with mean μ and known standard deviation $\sigma = 2$, find an unbiased estimate of μ.

Solution The sample mean $\overline{X}$ is an unbiased estimator of a population mean, so the estimate $\hat{\mu} = \bar{x} = 92$ is the desired estimate. ■

■ **Example 1.12** The inspector discovers that there is a problem with the voltmeter. Any reading of 100 or greater will be reported as 100. Find an unbiased estimate of μ.

Solution Now the population is not normally distributed, but rather it has a truncated normal distribution. The mean of a truncated normal (truncated on the left at $-\infty$, i.e., not truncated at all, and on the right at $b = 100$ has mean

$$E(X) = \mu - \sigma \frac{\phi(\beta)}{\Phi(\beta)}$$

where $\beta = (b-\mu)/\sigma = (100-\mu)/2$ and ϕ and Φ are respectively the PDF and CDF of the standard normal distribution. Finding an unbiased estimator for μ for the case of truncated data is beyond the scope of this book, but suffice it to say that $E(X)$ is slightly below μ, so the expected value of the sample average $\overline{X}$ is less than μ. Thus the sample mean is a biased estimate and it tends to be be on the low side of the true μ. ■

You may have noticed that none of the readings were 100 or greater, so had the voltmeter worked fine (and not truncated readings at the upper value of 100) nothing would have changed. Why would we change our estimate of μ because some of the observations *might have* exceeded 100, even though none of them did. As you might have guessed, with the Bayesian approach there would be no difference.

1.5 Conjugate Priors

In the section 16.2 we considered the problem of binomial data for which the probability of success had a uniform distribution over $(0,1)$. This suggests ignorance about the value of θ, but as the first example in section 16.4 indicates, it might be an informative prior in a different parameterization. Suppose we do have prior information about θ, for example, θ might be the probability of having a particular disease. A prior that puts most probability between 0 and 0.1 might be better than spreading

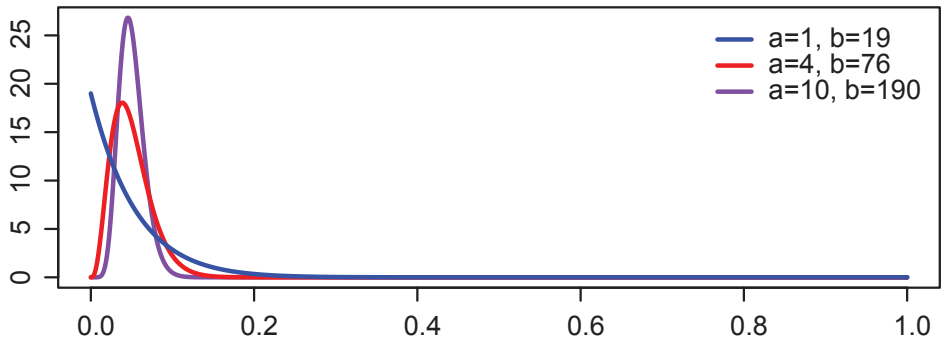


Figure 1.8: Three beta prior distributions. All have the same mean, $\mu = a/(a+b) = 0.05$, but the probability is more concentrated around the mean for larger values of $a+b$.

it over the whole interval $(0,1)$. In a case like this, we might search for a family of distributions whose support is the interval $(0,1)$ and is flexible to handle our belief about the parameter. The beta distribution, described in section 7.7, may be able to fit these requirements. The BETA(a,b) distribution has mean

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

and standard deviation

$$\sigma = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}} = \sqrt{\frac{a}{a+b}\left(1-\frac{a}{a+b}\right)\frac{1}{a+b+1}} = \sqrt{\frac{\mu(1-\mu)}{a+b+1}}.$$

We can select the mean by appropriately choosing a and b, but there are infinitely many ways to do this. For example, if we wanted the prior mean to be 0.05 (for example if we are sampling for a rather rare disease) we could select $a = 1$ and $b = 19$. Or we could select $a = 10$ and $b = 190$. The last part of the formula for the standard deviation suggests that for fixed values of $a/(a+b)$, the variance is proportional to $1/(a+b+1)$, so that larger values of a and b corresponds to smaller variability. Figure 16.8 illustrates this. It shows three priors, each having a mean of 0.05, but the standard deviations are 0.04756 (for $a = 1, b = 19$), 0.02422 (for ($a = 4, b = 76$), and 0.015373 (for $a = 10, b = 190$). Thus, the ratio $a/(a+b)$ controls the mean, and the magnitude of $a+b+1$ controls the scaling.

A family of distributions is said to be **rich** if it is able to accommodate a wide variety of prior beliefs. In this sense the beta distribution is rich because most forms of prior belief can be expressed (at least approximately) by a beta distribution with some a and b.

If we were to assume a beta prior distribution for the probability of success in a sequence of Bernoulli trials, then the posterior distribution will have a familiar form. To be specific, suppose we observe $X \sim \text{BIN}(n,\theta)$ where θ is unknown and has prior distribution $\theta \sim \text{BETA}(a,b)$. (Note that in a given problem, a and b would be *numbers* that express our belief about θ before observing any data; here we use arbitrary values a and b with the understanding that these will be known constants

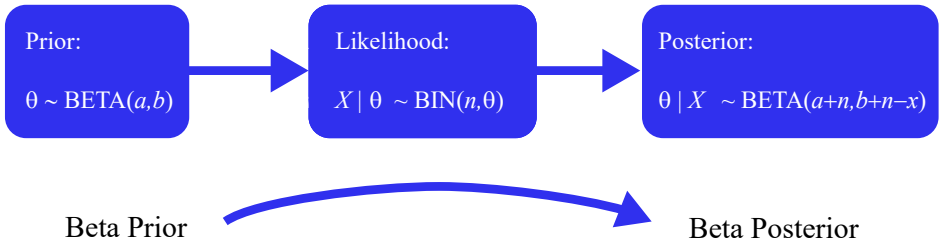


Figure 1.9: If we observe binomial data, a Beta prior distribution will lead to a Beta posterior distribution (but with different parameters). This means that the Beta distribution is a conjugate prior for the binomial.

in our problem.) The posterior is then

$$\begin{aligned}
\pi(\theta|x) &= c(x)\pi(\theta)\ell(x|\theta) \\
&= c(x)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\binom{n}{x}\theta^{x}(1-\theta)^{n-x} \\
&= c_1(x)\theta^{a+x-1}(1-\theta)^{b+n-x-1}, \\
&= \underbrace{c_1(x)}_{\text{depends on } x \text{ only}}\ \underbrace{\theta^{a+x-1}(1-\theta)^{b+n-x-1}}_{\text{depends on } \theta}, \qquad 0<\theta<1.
\end{aligned}$$

Notice that this posterior distribution factors into a product of one expression that depends on the data x only, and a piece that depends on both the data x and the parameter θ. We recognize from the latter that this must be a BETA distribution with parameters $a+x$ and $b+n-x$. Without doing any integration, we find that the normalizing constant must be

$$c(x) = \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}.$$

We call $c(x)$ a *constant* even though it depends of the data and the prior parameters a and b. Remember, that in a particular problem a and b will both be fixed numbers. Also, the data x is considered to be fixed when we compute the posterior distribution. Thus, the posterior distribution of θ given the data x is the BETA$(a+x, b+n-x)$ distribution. This idea is illustrated in Figure 16.5.

Definition 1.5.1 If, for a given likelihood function, the posterior distribution is in the same family as the prior, then the prior is said to be a *conjugate prior* for the given likelihood.

We saw previously that the beta is the conjugate prior for the binomial likelihood. There are a number of other of these conjugate priors.

■ **Example 1.13** Suppose we observe a sample $X_1, X_2, \ldots, X_n$ from the EXP(λ) distribution, which has PDF

$$f(x|\lambda) = \lambda \exp(-\lambda x), \qquad x > 0.$$

(Note that this is one of the two commonly used parameterizations of the exponential; the other uses the parameter $\theta = 1/\lambda$.) What is a conjugate prior for λ?

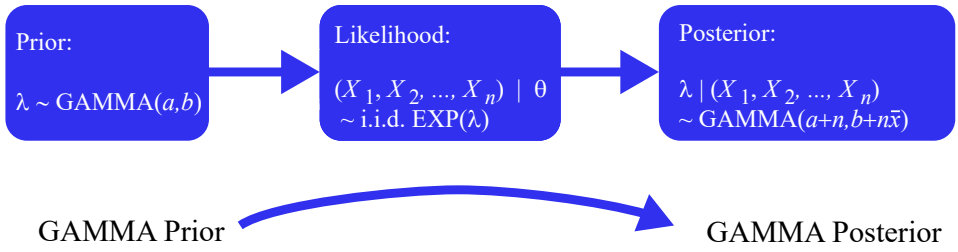


Figure 1.10: If we observe data from an exponential distribution, maybe lifetimes, then a Gamma prior distribution will lead to a Gamma posterior distribution (but with different parameters). This means that the Gammaa distribution is a conjugate prior for a sample from the exponential distribution.

Solution The likelihood function is

$$f(x_1,x_2,\ldots,x_n|\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

In order to determine a conjugate prior, we have to look at the likelihood as a function of λ; that is, we have to view this as

$$f(x_1,x_2,\ldots,x_n|\lambda) = \lambda^A \exp(-B\lambda)$$

where $A = n$ and $B = \sum_{i=1}^{n} x_i$. We then have to look for a prior that, when multiplied by this likelihood function, will yield a posterior in the same family. After a little searching, we would find that the GAM(a,b) distribution fills the bill.

If we take the prior to be

$$\pi(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda), \qquad \lambda > 0$$

then the posterior will be of the form

$$\begin{aligned} \pi(\lambda|x_1,x_2,\ldots,x_n) &= c\left(\lambda^{a-1}\exp(-b\lambda)\right)\left(\lambda^n \exp(-n\bar{x}\lambda)\right) \\ &= c\lambda^{a+n-1}\exp(-(b+n\bar{x})\lambda). \end{aligned}$$

This is obviously the GAM$(a+n, b+n\bar{x})$ distribution, so the constant c must be

$$c = \frac{(b+n\bar{x})^{a+n}}{\Gamma(a+n)}.$$

A GAM(a,b) prior for λ leads to a GAM$(a+n, b+n\bar{x})$ posterior. Since the prior and posterior are in the same family, the gamma distribution is a conjugate prior for the rate parameter λ of the exponential distribution.

■

The next example is a bit contrived, because we assume that the variance of a normal is known and the mean is not. Just as we used the z-test for the mean of a normal distribution with known

variance as a stepping stone to the t-test for the case when the variance was unknown, we use the Bayesian analysis of the known-variance case as a stepping stone to the unknown-variance case.

Bayesians usually prefer to work with the reciprocal of the variance, rather than the variance itself. This may seem a rather odd thing to do, but the reason should be clear after working through the next example. A skeptical reader is invited to work this example through using the variance.

Definition 1.5.2 The reciprocal of the variance of a normal distribution is called the *precision*. Usually, we use σ^2 to denote the variance, and τ to represent the corresponding precision; that is

$$\tau = \textit{ precision } = \frac{1}{\sigma^2}.$$

If there is a subscript on σ^2, we will use the same subscript on τ; for example, the precision for the $N(\mu_0, \sigma_0^2)$ distribution will be denoted τ_0.

■ **Example 1.14** Suppose that $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ where σ^2 is known. What is a conjugate prior for the parameter μ?

Solution Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]'$ denote the observed data. The likelihood function is then

$$L(\mu|\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \tag{1.9}$$

The question regarding a conjugate prior is really about what prior might we might combine with this likelihood in order to obtain a recognizable posterior. The right side of (16.9), when viewed as a function of μ (not $\boldsymbol{x}$) looks like it could be factored to be a normal distribution. The likelihood in (16.9) can be written in the form

$$L(\mu|\boldsymbol{x}) = c \ \exp\left(-\frac{1}{2\sigma^2}(n\mu^2 - 2\mu n\bar{x})\right). \tag{1.10}$$

Note that the constant c can depend on the data $\boldsymbol{x}$ and σ, because σ was assumed known. You are asked to derive this in Problem **??**. If we assume the $\mathrm{N}(\mu_0, \sigma_0^2)$ prior distribution for the mean μ, then our prior is of the form

$$\pi(\mu) = c \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right).$$

It is important to keep the notation straight: μ is the unknown mean of the normal distribution from which our data $x_1, x_2, \ldots, x_n$ comes from. Since we are taking a Bayesian approach, we must specify a prior distribution for all unknown parameters; in this case μ is the only unknown parameter. We assume the normal prior with mean μ_0 and standard deviation σ_0. In a real application, μ_0 and σ_0 will be fixed numbers.

In terms of precision, the prior and likelihood function are

$$\pi(\mu) = \exp\left(-\frac{\tau_0}{2}(\mu - \mu_0)^2\right)$$

and

$$L(\mu|\boldsymbol{x}) = c \ \exp\left(-\frac{\tau}{2}(n\mu^2 - 2\mu n\bar{x})\right).$$

The posterior distribution is therefore

$$\begin{aligned}
\pi(\mu|\boldsymbol{x}) &= c\pi(\mu)L(\mu|\boldsymbol{x}) \\
&= c\,\exp\left(-\frac{\tau_0}{2}(\mu-\mu_0)^2\right)\exp\left(-\frac{\tau}{2}(n\mu^2-2\mu n\bar{x})\right) \\
&= c\,\exp\left(-\frac{\tau_0}{2}(\mu^2-2\mu_0\mu+\mu_0)-\frac{\tau}{2}(n\mu^2-2n\bar{x}\mu)\right) \\
&= c\,\exp\left(-\frac{\tau_0+n\tau}{2}\mu^2+(\tau_0\mu_0+\tau n\bar{x})\mu\right) \\
&= c\,\exp\left(-\frac{\tau_0+n\tau}{2}\left(\mu^2-2\frac{\tau_0\mu_0+\tau n\bar{x}}{\tau_0+n\tau}\mu\right)\right).
\end{aligned}$$

Now comes the trick of completing the square. In order to make the exponent in this last expression be of the form μ minus something squared, we take half of the linear term, square it, and add it in the exponent. Because of the minus sign outside the parentheses, we are really subtracting this constant. Of course, if we subtract it out, we must also add it back in. This yields

$$\pi(\mu|\boldsymbol{x}) = c\,\exp\left(-\frac{\tau_0+n\tau}{2}\left(\mu^2-2\frac{\tau_0\mu_0+\tau n\bar{x}}{\tau_0+n\tau}\mu+\left(\frac{\tau_0\mu_0+\tau n\bar{x}}{\tau_0+n\tau}\right)^2\right)+\underbrace{\frac{\tau_0+n\tau}{2}\left(\frac{\tau_0\mu_0+\tau n\bar{x}}{\tau_0+n\tau}\right)^2}_{\text{adding the "square" back in}}\right).$$

Notice, however, that everything in the "complete the square" part (shown with the underbrace in the formula above) is *known*! (Remember, we've assumed that the variance, hence the precision τ is known.) Thus, the entire "square" part can be absorbed in the constant, which we now call $c(\boldsymbol{x})$. The posterior distribution is therefore

$$\pi(\mu|\boldsymbol{x}) = c_1(\boldsymbol{x})\,\exp\left(-\frac{\tau_0+n\tau}{2}\left(\mu-\left(\frac{\tau_0\mu_0+\tau n\bar{x}}{\tau_0+n\tau}\right)\right)^2\right).$$

This is the normal distribution with mean

$$\mu_1 = \frac{\tau_0\mu_0+\tau n\bar{x}}{\tau_0+n\tau}. \tag{1.11}$$

The precision is

$$\tau_1 = \tau_0+n\tau \tag{1.12}$$

and the variance is

$$\sigma_1^2 = \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}.$$

Thus, the normal distribution is the conjugate prior for the mean μ of the normal distribution. ∎

There are some interesting relationships to glean from (16.11) and (16.12). First, note that the posterior mean is a weighted average of the prior mean μ_0 and the sample mean $\bar{x}$; the weights are τ_0 (for the prior mean) and $n\tau$ for the sample mean. So the more data we have, the greater will be the weight on the sample mean. On the other hand, if the sample size is small, but the prior precision is

large (large precision means small prior variance) then the prior precision will dominate. The other feature to notice about (16.12) is that the posterior precision is just the sum of the prior precision and n times the known precision of the distribution. Recall that the variance of the sample mean $\overline{X}$ is σ^2/n, so the precision of $\overline{X}$ is $n\tau$. The posterior precision is thus the sum of the prior precision and the precision of the sample mean. This is why Bayesians like to work with the precision, rather than the variance.

The conjugate prior for the case where both the mean μ and variance $\phi = \sigma^2$ is derived in a similar manner, but it is considerably messier. We give the result and then show how in meets the conjugacy requirement. For starters, we note that we are now dealing with a joint density, so the family for the joint posterior must match the family for the joint prior. Also, to avoid confusion about whether σ^2 is itself a parameter, or whether it is the square of a parameter, σ, we use ϕ to denote the variance.

The conjugate prior is developed by first assuming that the variance has the scaled inverse χ^2 distribution, denoted $\text{SInv}\chi^2(\nu 0, \phi 0)$, with PDF

$$p(\phi) = \frac{(\nu_0/2)^{\nu_0/2}\phi_0^{\nu_0/2}}{\Gamma(\nu_0/2)}\phi^{-\nu_0/2-1}\exp\left(-\frac{\nu_0\phi_0}{2\phi}\right), \qquad \phi > 0 \tag{1.13}$$

The mean of the $\text{SInv}\chi^2(\nu 0, \phi 0)$ is

$$\mathbb{E}(\phi) = \frac{\nu_0}{\nu_0 - 2}\phi_0 \tag{1.14}$$

and the variance is

$$\mathbb{V}(\phi) = \frac{2\nu_0^2}{(\nu_0-2)^2(\nu_0-4)}\phi_0^2. \tag{1.15}$$

For the prior mean to exist, we must have $\nu_0 > 2$ and for the variance to exist, we must have $\nu_0 > 4$. We can incorporate prior information about ϕ by taking ϕ_0 to be near our prior estimate since the prior mean is $\frac{\nu_0}{\nu_0-2}\phi_0$ which is nearly ϕ_0 if ν_0 is large. The more confident we are in our prior estimate, the larger we should choose ν_0 because the prior variance goes to zero as $\nu_0 \to \infty$.

Given ϕ, we assign the normal prior $\text{N}(\mu_0, \phi/\kappa_0)$ to μ. Obviously, we would assign μ_0 to be our prior estimate of μ. If we have strong prior beliefs about μ we would choose κ_0 to be large, and if our prior beliefs are weak, we would choose a small κ_0. The parameter κ_0 can be thought of as the number of data points we believe our prior belief is worth. For example, suppose we believe that the mean μ of a population is about 100. When asked how strongly we feel about this, we might say that our belief is as strong as we observed 5 data values from which we obtained a mean of 100; here we would select $\mu_0 = 100$ and $\kappa_0 = 2$.

The joint prior obtained this way is called the normal-inverseχ^2 distribution, denoted $\text{NInv}\chi^2(\mu_0, \kappa_0, \nu_0, \phi_0)$. If we observe $X_1, X_2, \ldots, X_n$ from the $\text{N}(\mu, \phi)$ distribution, it can be shown that the posterior distribution is the $\text{N}(\mu_1, \kappa_1, \nu_1, \phi_1)$ where

$$\mu_1 = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\overline{X} \tag{1.16}$$

$$\kappa_1 = \kappa_0 + n \tag{1.17}$$

$$\nu_1 = \nu_0 + n \tag{1.18}$$

$$\phi_1 = \frac{\nu_0\phi_0}{\nu_0+n} + \frac{(n-1)}{\nu_0+n}S^2 + \frac{\kappa_0 n}{(\nu_0+n)(\kappa_0+n)}(\overline{X}-\mu_0)^2. \tag{1.19}$$

where $\overline{X}$ and S^2 are the sample mean and variance, respectively. Several rather intuitive features can be seen in equations (16.16) through (16.19). First, equation (16.16) shows that the posterior mean is a weighted average of the prior mean μ_0 and the sample mean $\overline{X}$. Larger values of the sample size n makes the sample mean $\overline{X}$ have a higher weight. Also, larger values of κ_0 give more weight to the prior mean μ_0. Since the prior parameters κ_0 and ν measure the strength of our prior beliefs, it shouldn't be surprising that our strength increases for larger sample sizes. This happens in a simple way, by adding n to each of κ_0 and ν_0. Equation (16.19) is the messiest to interpret. The posterior estimate of the variance ϕ is dependent on the prior estimate ϕ_0, the sample variance S^2 and them squared difference between the prior mean μ_0 and the sample mean $\overline{X}$. In some sense equations tell us how to update our beliefs about the parameters μ and $\phi = \sigma^2$ once we have observed the results of our sample in the sense that with prior

$$\text{NInv}\chi^2(\mu_0, \kappa_0, \nu_0, \phi_0)$$

we end up with the posterior

$$\text{NInv}\chi^2(\mu_1, \kappa_1, \nu_1, \phi_1)$$

where μ_1, κ_1, ν_1, and ϕ_1 are given in equations (16.16) through (16.19).

The marginal posterior distribution for μ is then

$$\mu \mid (\overline{X}, S^2) \sim t_{n_\nu}(\mu_n, \phi_n/\kappa_n) \tag{1.20}$$

where $\sim t_{n_\nu}(\mu_n, \phi_n/\kappa_n)$ is the t distribution with n_ν degrees of freedom, centered at μ_n with scale parameter ϕ_n/κ_n; in other words

$$\frac{\mu - \mu_n}{\sqrt{\phi_n/\kappa_n}} \mid (\overline{X}, S^2) \sim t_{\nu_n}. \tag{1.21}$$

We should point out a few important points here. First, the random variable on the left side of (16.21) is μ. Remember that in the Bayesian paradigm parameters are treated as random variables and the uncertainty is expressed in probabilistic terms. The values of μ_n and ϕ_n are functions of the data and the prior parameters, so once data are collected and recorded μ_n and ϕ_n are fixed numbers. Second, this t distribution is centered at μ_n, but this is not what we call the noncentral t distribution; this term is reserved for the distribution introduced in Chapter 12.

The marginal posterior for the population variance ϕ is

$$\phi \mid (\overline{X}, S^2) \sim \text{Inv}\chi^2(\nu_n, \phi_n). \tag{1.22}$$

Equivalently, we could say that

$$\frac{1}{\phi} \sim \chi^2(\nu_n, \phi_n). \tag{1.23}$$

The derivations of the results described here are not difficult, but are rather messy. An interested reader should consult Gelman et al. (2013, Chapter 3) or del Castillo and Colosimo (2006, Chapter 1).

■ **Example 1.15 — Estimating the Mean and Variance of a Normally Distributed Population.** Suppose that we wish to estimate the mean and variance of the body mass index (BMI) of a particular population. Before the study begins, we don't know whether to expect that the BMI is higher or lower than the country as a whole. For this case, develop reasonable priors for μ and $\phi = \sigma^2$. Then use these priors, along with the data

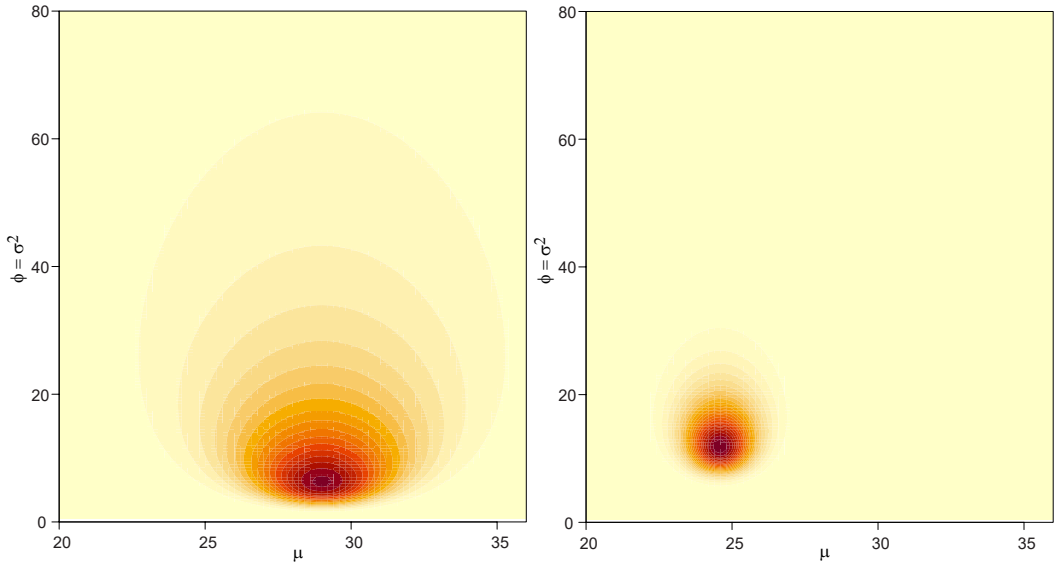


Figure 1.11: Joint NInvχ^2 prior (left) and posterior (right) for hypothetical data on BMI.

```
26, 32, 19, 23, 22, 25, 28, 19, 23, 27, 26, 22, 18, 26, 24
```

Solution: In the US, the mean BMI is about 29 with a standard deviation of approximately 5. For our prior for ϕ we choose $\phi_0 = 5^2$ and with little confidence in this initial estimate, we choose $\nu_0 = 1$.

Since we don't know whether our population will have a higher or lower mean BMI, we take as our prior parameters

$$\mu_0 = 29, \qquad \phi_0 = 5^2.$$

As for our uncertainty in our prior estimate of the mean μ, we might choose $\kappa_0 = 2$, suggesting little confidence in our initial estimate. Our joint prior distribution is therefore

$$(\mu, \phi) \sim \text{NInv}\chi^2(\mu_0 = 29, \kappa_0 = 2, \phi_0 = 5, \nu_0 = 1)$$

Figure 16.15 shows both the joint prior and the posterior for our hypothetical BMI data. Darker colors indicate higher values of the PDF. While the prior is mostly spread out over reasonable values of μ and $\phi = \sigma^2$, the posterior is concentrated near $\mu = 25$ and $\phi = 12$. The marginal posterior distributions are

$$\mu|\boldsymbol{x} \sim t_{\nu_n}(\mu_n, \phi_n/\kappa_n)$$

and

$$\phi|\boldsymbol{x} \sim \text{Inv}\chi^2(\nu_n, \phi_n)$$

If we take the posterior mean as the point estimate for parameters, we have

$$\hat{\mu} = \mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\overline{X}$$

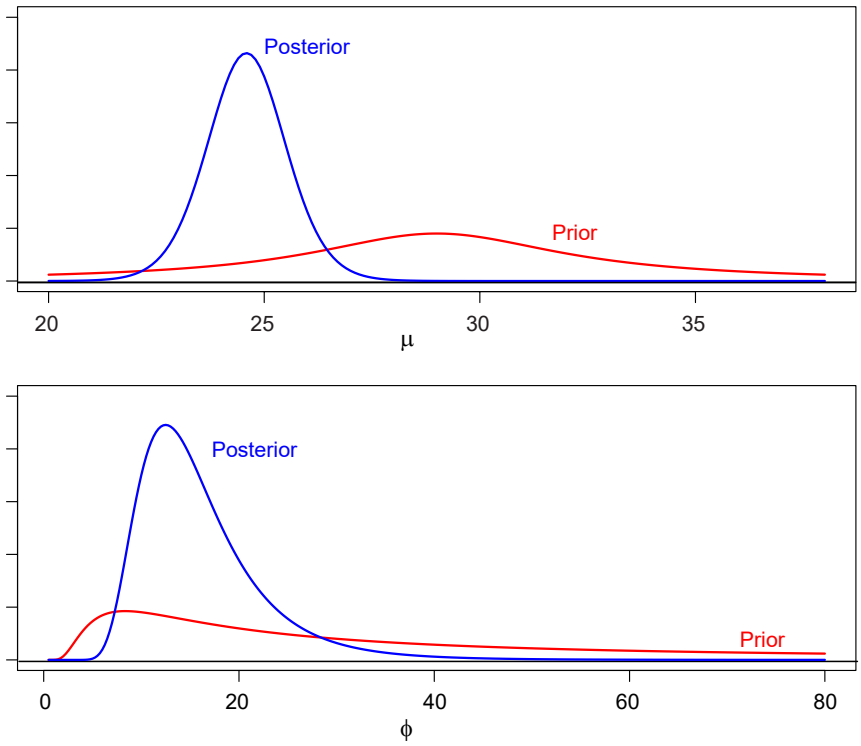


Figure 1.12: Marginal prior (in red) assuming a joint NInvχ^2 prior and the posterior (blue) for the hypothetical BMI data. The prior and posterior for μ are shown at the top, and the prior and posterior for ϕ are shown at the bottom.

and

$$\hat{\sigma}^2 = \hat{\phi} = \frac{\nu_n}{\nu_n - 2}\phi_n.$$

Often studies such as this focus on the population mean. The marginal prior distribution for μ is a $t_{\nu_0}(\mu_0, \phi_0/\kappa_0)$ distribution and the posterior is a $t_{\nu_1}(\mu_1, \phi_1/\kappa_1)$ distribution. Substituting the numbers for the parameters we find that

$$\begin{aligned} \text{Prior:} \quad & t_1(29, 5^2/2) \\ \text{Posterior:} \quad & t_{16}(24.59, 15.05/17). \end{aligned}$$

Figure 16.15 shows the prior and posterior PDF for the population mean μ.

■

In the next example we see that a larger sample size yields posteriors with much less variability.

■ **Example 1.16** Consider the data we studied previously on health data from 768 pregnant women with diabetes. Let's focus our interest on the mean and variance of the BMI. We will start with the same prior distributions for μ and ϕ as we did in the last example, namely, the

$$\text{NInv}\chi^2(\mu_0 = 29, \kappa_0 = 2, \phi_0 = 5, \nu_0 = 1) \tag{1.24}$$

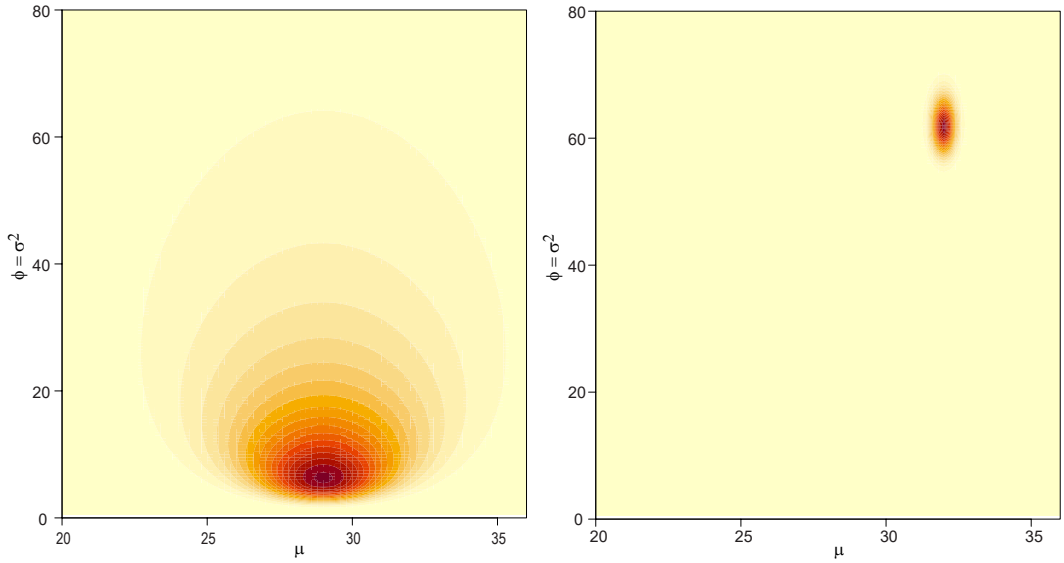


Figure 1.13: Joint NInvχ^2 prior (left) and posterior (right) for the data set on BMI for pregnant women.

distribution. The sample mean and sample variance are the sufficient statistics for this data set and these are

$$\begin{aligned} \bar{x} &= 31.99 \\ s^2 &= 62.16. \end{aligned}$$

The parameters of the normal-scaled inverse χ^2 distribution are

$$\begin{aligned} \mu_1 &= 31.98 \\ \kappa_1 &= 770 \\ \phi_1 &= 62.03 \\ \nu_1 &= 769 \end{aligned}$$

yielding posterior point estimates (the means of the posterior distribution)

$$\begin{aligned} \hat{\mu} &= \mu_1 = 31.98 \\ \hat{\sigma}^2 &= \hat{\phi} = \frac{\nu_n}{\nu_n - 2}\phi_n = 62.20. \end{aligned}$$

Figure 16.16 shows the joint prior and posterior for the population mean and variance. Note that the left side of Figure 16.16 is the same as the left side of Figure 16.15; this occurs because the prior was taken to be the same for both cases. The joint posterior distribution, shown on the right side of Figure 16.16 indicates a very small region of high posterior density, much smaller than in Figure 16.15. This is due to the much larger sample size ($n = 768$ here compared to $n = 15$ in the previous example).

■

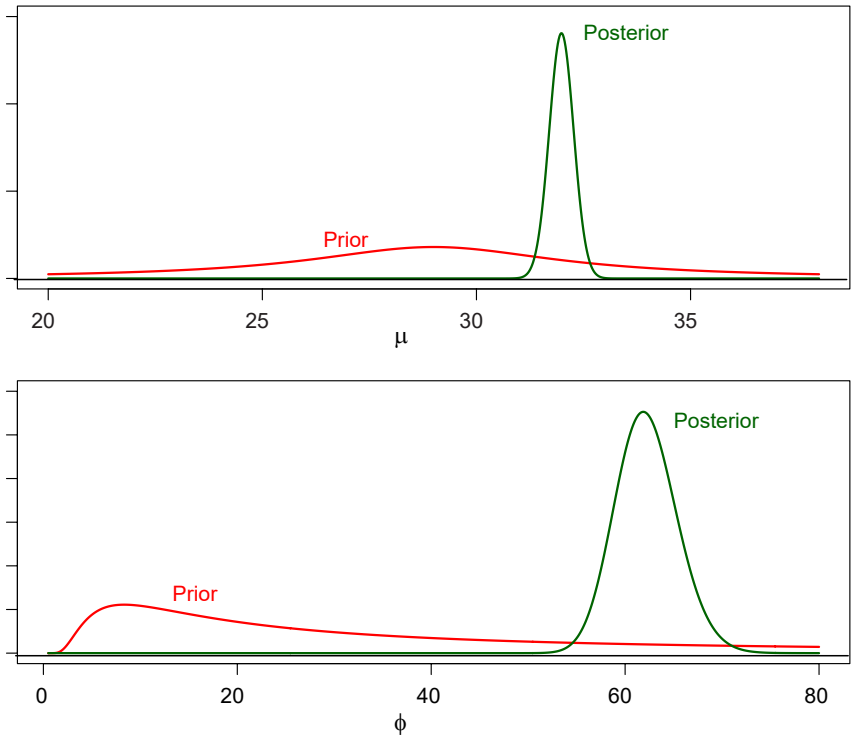


Figure 1.14: Marginal prior (in red) assuming a joint NInvχ^2 prior and the posterior (green) for the BMI data for pregnant women. The prior and posterior for μ are shown at the top, and the prior and posterior for ϕ are shown at the bottom.

1.6 Noninformative Priors

If we have little or no information about a parameter it seems that a prior that is uniform would reflect this. In example 16.7 we showed that a uniform prior on the probability θ of the binomial distribution leads to a nonuniform prior on the log odds, $\log \theta/(1-\theta)$. While we didn't show it, the converse also holds: a uniform prior on the log odds leads to a nonuniform prior on θ itself. This suggests that we should be careful in the choice of the parameterization and which parameterization gets the uniform prior.

The good news is that if we have a lot of data the posterior is dominated by the data and the prior has little influence. When the sample size is small or moderate, the prior does exert a reasonable amount of influence on the posterior.

In 1946 Harold Jeffreys determined the form of a prior that invariant under transformations. In other words, the probability of being in some region with one parameterization is the same as the probability in the transformed region with the transformed parameterization. Jeffreys showed that the prior

$$p(\theta) = c\sqrt{I(\theta)} \tag{1.25}$$

is invariant under transformations, where

$$I(\theta) = E\left(-\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right) \tag{1.26}$$

is the Fisher information. Check to see whether this was mentioned previously.

The definition given here assumes a single parameter, and as we have seen, most problems involve many parameters. Although there is a multivariate version of the definition of Jeffreys prior, the prior is often determined by taking the Jeffreys prior for each parameter, one at a time, by assuming the other parameters are fixed.

■ **Example 1.17 — Jeffreys prior for the binomial.** Suppose we observe the number of successes on n Bernoulli trials. Let X denote the number of successes; then $X \sim \text{BIN}(n,\theta)$ where θ is the unknown probability of a success on any one trial. Find the Jeffreys prior.

Solution: The likelihood function is

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0,1,\ldots,n;\ \ 0 \le \theta \le 1.$$

Taking the logarithm and then the first two derivatives yields

$$\begin{aligned}
\log f(x|\theta) &= \log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta) \\
\frac{\partial \log f}{\partial\theta} &= \frac{x}{\theta} + \frac{n-x}{1-\theta}(-1) \\
&= \frac{x}{\theta} - \frac{n-x}{1-\theta} \\
\frac{\partial^2 \log f}{\partial\theta^2} &= -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}.
\end{aligned}$$

The Fisher information is therefore

$$\begin{aligned} I(\theta) &= E\left(-\frac{\partial}{\partial\theta}\log f(X|\theta)\right) \\ &= \frac{1}{\theta^2}E(X|\theta)+\frac{1}{(1-\theta^2)}E(n-X|\theta) \\ &= \cdots \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

We've skipped a number of steps in the algebraic solution. Problem 16.2 asks you to fill in the required steps. The Jeffreys prior is therefore

$$p(\theta) = c\sqrt{I(\theta)} = c\sqrt{\frac{n}{\theta(1-\theta)}} = c_1\theta^{-1/2}(1-\theta)^{-1/2}. \tag{1.27}$$

This is the BETA$(\frac{1}{2},\frac{1}{2})$ distribution. ∎

In the last example, the Jeffreys prior is a proper PDF, in the sense that it integrates to 1, as all PDFs must. In the next example, we see that sometimes the Jeffreys prior is improper, in the sense that it does not integrate to anything finite. Often, though not always, an improper prior still leads to a proper posterior.

∎ **Example 1.18 — Jeffreys prior for the parameters of the normal distribution.** Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from the $\mathrm{N}(\mu,\sigma^2)$ distribution. Find the Jeffreys prior for μ and for $\phi = \sigma^2$.

Solution: We will obtain the Jeffreys prior for the two parameters one at a time, by assuming that the other is known. We begin with the mean parameter μ.

The likelihood function for a single observation is

$$f(x|\mu,\phi) = \frac{1}{\sqrt{2\pi}\sqrt{\phi}}\exp\left(-\frac{1}{2\phi}(x-\mu)^2\right)$$

Taking the logarithm, and the first two derivatives yields

$$\begin{aligned} \log f(x|\mu,\phi) &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\phi - \frac{1}{2\phi}(x-\mu)^2 \\ \frac{\partial \log f}{\partial\mu} &= 0+0-\frac{1}{2\phi}2(X-\mu)(-1) \;=\; \frac{x-\mu}{\phi} \\ \frac{\partial^2 \log f}{\partial\mu^2} &= -\frac{1}{\phi} \\ I(\theta) &= E\left(-\frac{\partial^2}{\partial\theta^2}\log f(X|\mu,\phi)\right) \;=\; -E\left(-\frac{1}{\phi}\right) \;=\; \frac{1}{\phi}. \end{aligned}$$

Jeffreys prior is therefore

$$p(\mu) = c\sqrt{I(\mu)} = \frac{c}{\phi} = c_1.$$

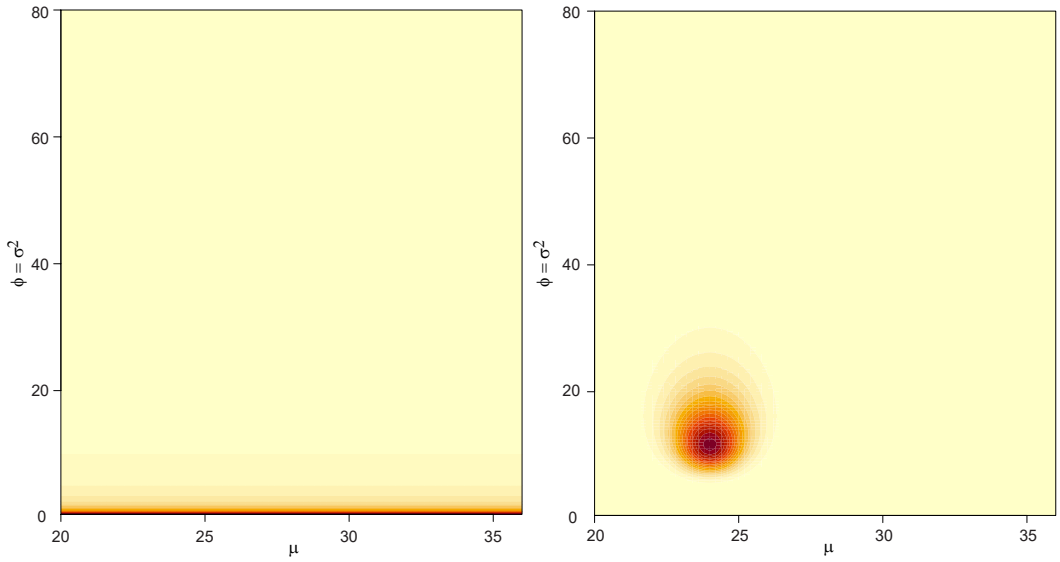


Figure 1.15: Joint Jeffreys prior (left) and posterior (right) for the hypothetical BMI data.

Note that the prior for μ is a constant. No matter how close to zero the constant is, the prior will always integrate to ∞. This is an example of an improper prior.

The partial derivatives of the PDF with respect to ϕ and the Fisher information are:

$$\begin{aligned}
\frac{\partial \log f}{\partial \phi} &= 0 - \frac{1}{2}\theta^{-1} + \frac{1}{2}\phi^{-2}(x-\mu)^2 \\
\frac{\partial^2 \log f}{\partial \phi^2} &= \frac{1}{2}\phi^{-2} - \phi^{-3}(x-\mu)^2 \\
I(\theta) &= E\left(-\frac{\partial^2}{\partial \phi^2}\log f(X|\mu,\phi)\right) = -\frac{1}{2}\phi^{-2} + \phi^{-3}E\left((X-\mu)^2\right) = \frac{1}{2}\phi^{-2}.
\end{aligned}$$

The Jeffreys prior for ϕ is therefore

$$p(\phi) = c\sqrt{\frac{1}{2}\phi^{-2}} = c_1\frac{1}{\phi}.$$

This is also an improper prior. ■

■ **Example 1.19** Rework Example 16.16 assuming the Jeffreys prior.
Solution:

■

■ **Example 1.20** Rework Example 16.15 assuming the Jeffreys prior.
Solution:

■

1.7 Simulation Methods

As we have seen in the previous section, Bayesian calculations become difficult even for simple two parameter problems like the problem of estimating the mean and variance of a normal distribution.

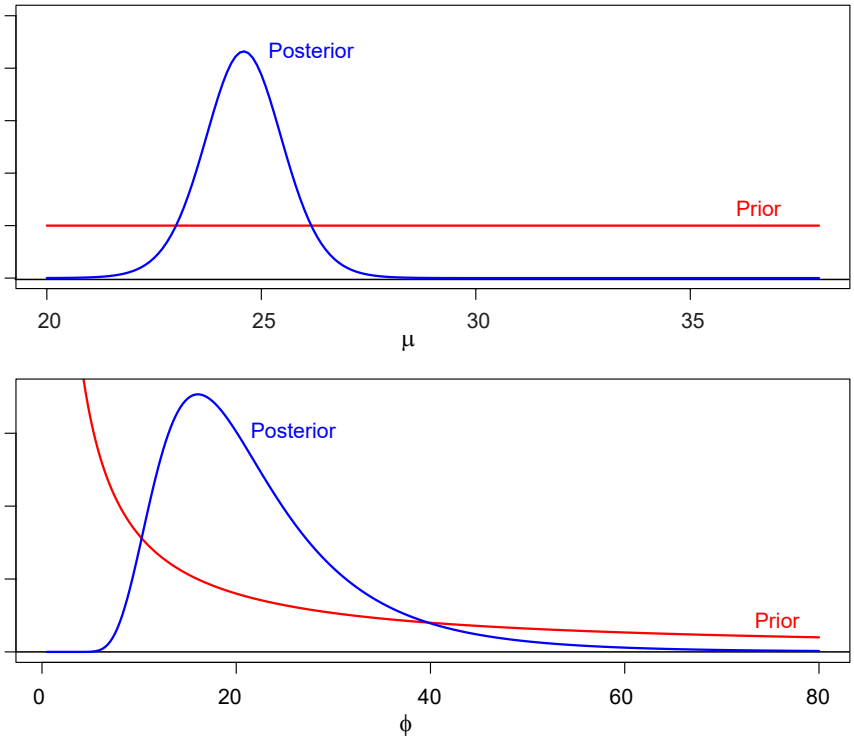


Figure 1.16: Marginal prior (in red) assuming the Jeffreys prior and the posterior (blue) for the hypothetical BMI data. The prior and posterior for μ are shown at the top, and the prior and posterior for ϕ are shown at the bottom.

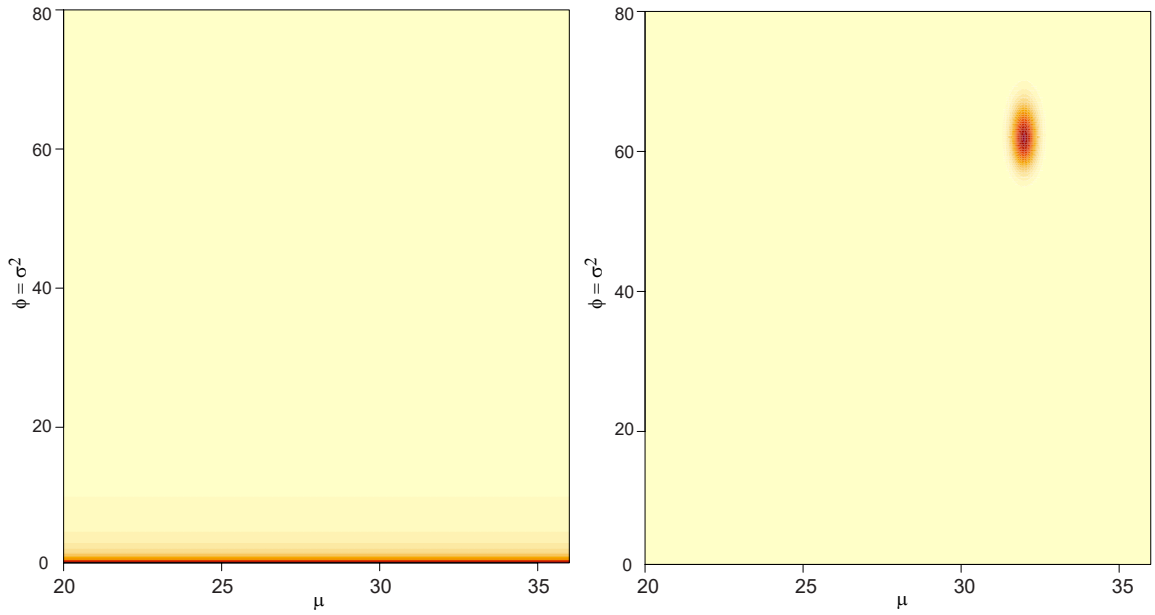


Figure 1.17: Marginal prior (left) and posterior (right) distributions for the mean μ and variance ϕ for the Pima diabetes data. The priors are Jeffreys' priors.

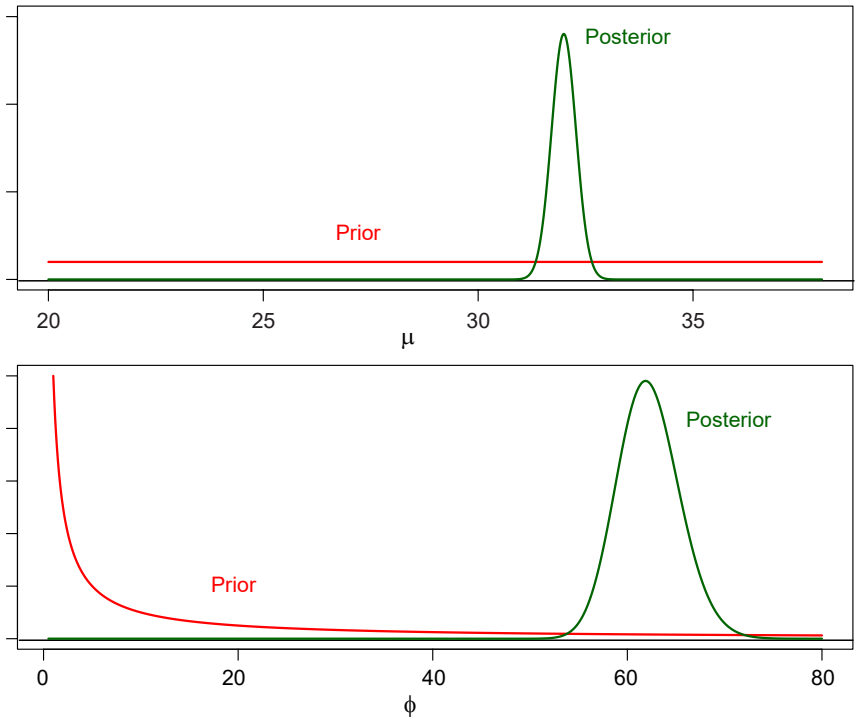


Figure 1.18: Marginal prior (left) and posterior (right) distributions for the mean μ and variance ϕ for the Pima diabetes data. The priors are Jeffreys' priors.

For higher dimensional problems, an exact solution for the posterior distribution, and the marginal posterior distributions, is often intractable. The problem usually lies in the denominator of Bayes Theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})}{\int p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}}.$$

This integral is a p-dimensional integral, where p is the number of unknown parameters in the vector $\boldsymbol{\theta}$.

For the case of an unknown mean and variance from a normal distribution with a NInvχ^2 prior distribution, the posterior has PDF

$$p(\mu,\phi|\boldsymbol{x}) = \frac{\phi^{-\nu_0/2-n/2-3/2}\exp\left(-\frac{1}{2\phi}\left[\nu_0\phi_0^2+\kappa_0(\mu-\mu_0)^2+(n-1)s^2+n(\bar{x}-\mu)^2\right]\right)}{\int_0^\infty\int_{-\infty}^\infty \phi^{-\nu_0/2-n/2-3/2}\exp\left(-\frac{1}{2\phi}\left[\nu_0\phi_0^2+\kappa_0(\mu-\mu_0)^2+(n-1)s^2+n(\bar{x}-\mu)^2\right]\right)\, d\mu\, d\phi}. \tag{1.28}$$

This is a formidable equation indeed! Fortunately, in this case it works out to be the NInvχ^2 distribution, whose marginals are known. In more complicated cases we'er just left with a complicated expression, and in dimensions higher than two it is difficult (impossible in some cases) to visualize what the posterior is telling us about the unknown parameters.

What if we could simulate random values from the joint posterior distribution? We could them plot them, either one parameter at a time if we're interested in a marginal posterior, or two at a time

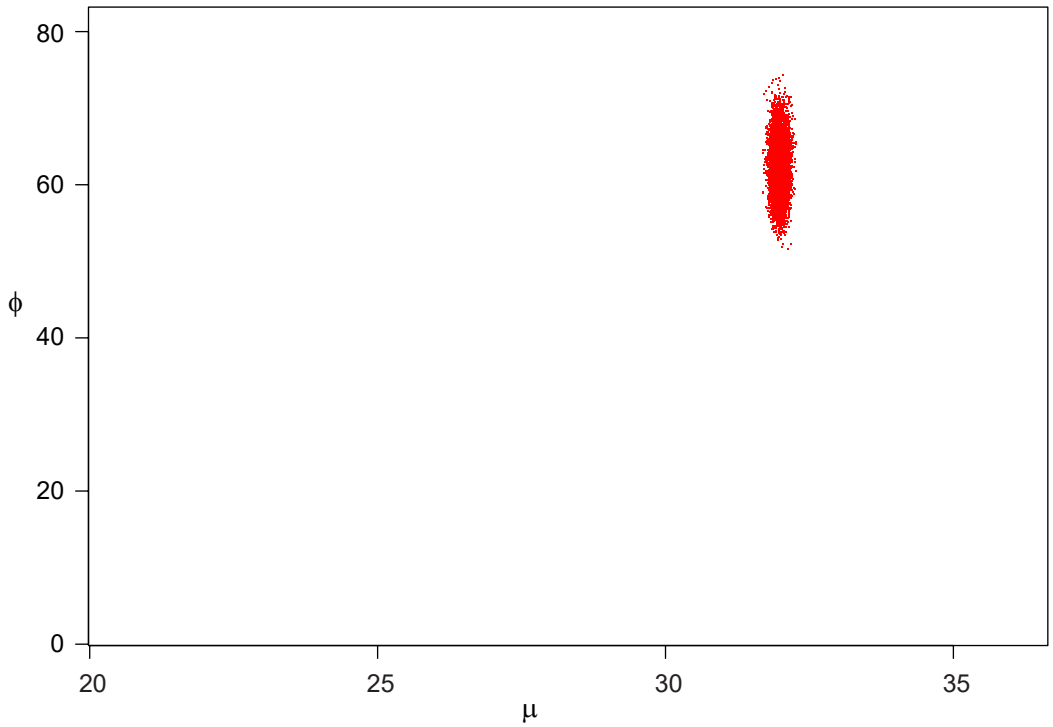


Figure 1.19: 10,000 simulations from the joint posterior distribution.

in a scatter plot if we're interested in a joint posterior. Figure 16.19 shows the scatterplot of 10,000 simulations from the posterior distribution of the mean and variance of the distribution of BMI in Example 16.16. Compare this with the contour plot of the posterior distribution given on the right side of Figure 16.16; these plots convey basically the same information.

Histograms for the marginals for μ and ϕ are shown in Figure 16.20. Compare these plots with the marginal distributions derived analytically and shown in Figure 16.16

Analogously, the marginal posteriors for μ and ϕ are shown in Figure 16.20. These convey much the same information as shown in Figure 16.16.

In the case we have discussed at length in this chapter, namely the problem of estimating the posterior distribution of the mean and variance of a normal distribution, the exact solution is messy but tractable. The point of this section is that the tractability ends as problems get progressively harder and in theses cases some sort of simulation methods are required. The question you should ask yourself is this: which gives me more useful information about the posterior, a messy formula such as (16.28 or the scatter plot of 10,000 simulated values from the posterior? Usually, the answer is the latter.

1.7.1 Metropolis-Hastings Algorithm

The algorithm we describe here, called the Metropolis-Hastings algorithm, dates back to 1953 when Metropolis, Rosenbluth, Rosenbluth, Teller and Teller developed a simulation method for a problem in physics. Nearly two decades later, in 1970, Hastings generalized the method. It remained an obscure method to statisticians until 1984 when brothers Donald Geman and Stuart Geman applied the algorithm to the problem of image restoration using Bayesian methods. This led to the technique

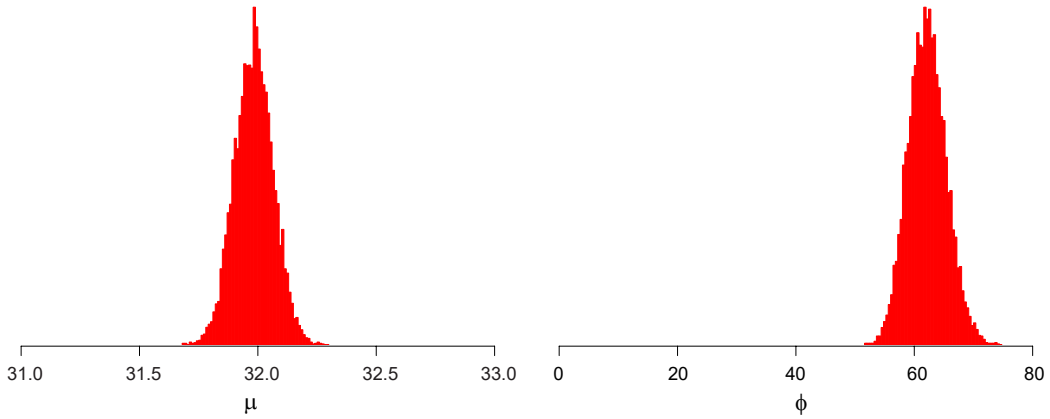


Figure 1.20: Histograms of the marginal posterior distributions for μ and ϕ based on 10,000 simulations.

generally known as Markov chain Monte Carlo (MCMC). This method opened up a new world of applications of Bayesian statistics. Problems that were intractable were immediately tackled using simulation methods based on simulation.

Algorithm. *Metropolis-Hastings Algorithm for Simulating from a Posterior Distribution*

Suppose we want to simulate from a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$

Step 1: Starting Value: *Begin with an initial guess for value of* $\boldsymbol{\theta}$ *that is near the middle of the posterior distribution. Call this value* $\boldsymbol{\theta}^{(0)}$*. This must be a value for which* $p(\boldsymbol{\theta}^{(0)}|\boldsymbol{x}) > 0$*. Set* $i = 1$*.*

Step 2a: Proposed Move: *Let* $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$ *be some conditional PDF that has the same support as* $p(\boldsymbol{\theta}|\boldsymbol{x})$*. (This PDF may depend on the previous value* $\boldsymbol{\theta}^{(i-1)}$ *but it could be independent of* $\boldsymbol{\theta}^{i-1}$*.) Here* g *is called the* proposal density*. Simulate one value from* $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$ *and call it* $\boldsymbol{\theta}^{(prop)}$*.*

Step 2b: Compute Acceptance Probability: *Compute*

$$\alpha = \frac{p(\boldsymbol{\theta}^{(prop)}|\boldsymbol{x})\ g(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^{(prop)})}{p(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{x})\ g(\boldsymbol{\theta}^{(prop)}|\boldsymbol{\theta}^{(i-1)})}. \tag{1.29}$$

Step 2c: Accept or Reject the Proposed Move to $\boldsymbol{\theta}^{(prop)}$***:***

If $\alpha \geq 1$*, then accept the move to* $\boldsymbol{\theta}^{(prop)}$*. If* $\alpha < 1$ *then accept the move to* $\boldsymbol{\theta}^{(prop)}$ *with probability* α*. This can be done with the* `sample` *command in R:*

```
sample( c("Accept","Reject") , size=1 , prob=c(alpha,1-alpha) )
```

If we accept the move we set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(prop)}$*. Otherwise we stay at the same place, namely* $\boldsymbol{\theta}^{(i-1)}$*, in which case we set* $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$*. In other words, we set*

$$\boldsymbol{\theta}^{(i)} = \begin{cases} \boldsymbol{\theta}^{(prop)} & \text{if the move is accepted} \\ \boldsymbol{\theta}^{(i-1)} & \text{if the move is rejected} \end{cases} \tag{1.30}$$

Step 3: Branch: *If the required number of simulations has been reached, then stop. Otherwise, set* $i \leftarrow i+1$ *and go to Step 2a.*

The successive values generated by this process are not independent, partly because consecutive values might (or might not) be the same. The Metropolis-Hastings algorithm is a realization of a Markov chain, that is a sequence of random variables with the property that the probability distribution of the random variable at time i depends only on the value at time $i-1$ and not on the value of the random variable at any time points before time $i-1$. Under the right circumstances, Markov chains converge to a steady state distribution where the unconditional distribution of the random variable at time i is the same regardless of i. (Note that this paragraph very briefly summarizes some deep theoretical notions. Interested readers are referred to books on Bayesian statistics for details that we have omitted here.)

If the proposal density is symmetric about the current value $\boldsymbol{\theta}^{(i-1)}$, then

$$g(\boldsymbol{\theta}^{(\text{prop})}|\boldsymbol{\theta}^{(i-1)}) = g(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^{(\text{prop})})$$

so these cancel in the calculation of α, which then becomes

$$\alpha = \frac{p(\boldsymbol{\theta}^{(\text{prop})}|\boldsymbol{x})}{p(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{x})} = \frac{p(\boldsymbol{\theta}^{(\text{prop})})\ f(\boldsymbol{x}|\boldsymbol{\theta}^{(\text{prop})})}{p(\boldsymbol{\theta}^{(i-1)})\ f(\boldsymbol{x}|\boldsymbol{\theta}^{(i-1)})}. \tag{1.31}$$

When the proposal density is symmetric like this, the algorithm is usually called the Metropolis Algorithm; when the proposal density is not symmetric, the algorithm is usually called the Metropolis-Hastings algorithm.

The theory says that the steady state distribution of the Markov chain defined by the Metropolis-Hastings algorithm is the posterior distribution. This fact suggests the following method for learning about a posterior distribution.

Algorithm. *Using the Metropolis-Hastings Algorithm to Learn about the Posterior*

Burn-in: *Apply the Metropolis-Hastings algorithm for enough iterations that we are convinced, for all practical purposes, that the steady state has been reached.*

Simulate from the Posterior: *Simulate a number of additional iterations from the Markov chain and use these as simulations from the desired posterior. Note, however, that these are not independent observations, so the information in say 1000 iterations is less that it would be if we had 1000 independent simulations.*

There are software engines that will perform the Metropolis-Hastings algorithm or its relative, Gibbs sampling (discussed below). We believe it is instructive to code the Metropolis-Hastings algorithm from scratch to see how it works. In the code below, we define three functions at the top: the beta PDF (excluding the $\Gamma(a+b)/\Gamma(a)\Gamma(b)$ constant), the likelihood (excluding the $\binom{n}{x}$ constant), and the function that proposes the next value in the Markov chain. The proposal distribution returns a normally distributed random variable with mean equal to the current value of θ and standard deviation of 0.1. Note that if this value is outside of the interval $[0,1]$ the proposal is automatically rejected because the posterior distribution is equal to 0 outside this interval. We then set the random

seed equal to 3210 so that repeated runs will give exactly the same output. If you'd rather get different results on each run, you can select a different seed or omit the `set.seed` function. We define the data, which for this example is binomial with $n = 6$ and $x = 1$; that is, six trials with one success. The prior distribution is a BETA$(2,2)$ distribution. We run `nsim = 11000` simulations, expecting to discard several at the beginning to assure that we've reached the steady state. Then, inside the loop we propose a new θ called `thetap` where the "p" stands for "proposed." The current value of θ is called `thetac` where the "c" stands for "current.". The heart of the Metropolis-Hastings is the line that computes `alpha`, the acceptance probability. Note that the proposal distribution, the normal in this case, is symmetric so the proposal density $g(|)$ drops out. The next line assigns the variable `decision` to be `"Acc"` or `"Rej"` with probabilities $\min(\alpha,1)$ and $1-\min(\alpha,1)$. Then, if the decision is `"Acc"` we set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{\text{prop}}$; otherwise we set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(\text{i-1})}$. Here is the R code:

```
> pBeta = function(theta,a,b)
> {
>   z = 0
>   if (theta >= 0 && theta <= 1) { z = theta^(a-1) * (1-theta)^(b-1) }
>   return(z)
> }
> thetaLik = function(x,n,theta)  theta^x * (1-theta)^(n-x)
> thetaprop = function(theta)  rnorm(n=1,mean=theta,sd=0.1)
>
> set.seed(3210)
> n = 6
> x = 1
> a = 2
> b = 2
> nsim = 11000
> thetavec = rep(0,nsim)
> thetavec[1] = 0.50
>
> for (i in 2:nsim)
> {
>   thetap = thetaprop( thetavec[i-1] )
>   thetac = thetavec[i-1]
>   alpha =   pBeta(thetap,a,b) * thetaLik(x,n,thetap) /
>           ( pBeta(thetac,a,b) * thetaLik(x,n,thetac) )
>   probAcc = min(1,alpha)
>   decision = sample( c("Acc","Rej") , size=1 , prob=c(probAcc,1-probAcc) )
>   if ( decision == "Acc" )   thetavec[i] = thetap; acc = acc + 1
>   else thetavec[i] = thetac
> }
```

The data (actually datum in this case) is the outome $X \sim \text{BIN}(6,\theta)$ and the prior is BETA$(2,2)$. By the discussion in 16.5, since the beta is a conjugate prior for the binomial distribution, we could conclude immediately that the posterior distribution is the BETA$(2+1,2+6-1)$. We ran the MCMC simulations using the Metropolis algorithm (since the proposal distribution is symmetric,

the Metropolis-Hastings reduces to the slightly simpler Metropolis algorithm) so that we could see how the algorithm works in a simple case.

The Metropolis-Hastings algorithm produces a Markov chain, in the sense that the distribution for the outcome at time t is dependent only on the outcome at time $t-1$, now how we got to the outcome at time $t-1$. The theory says that the steady state distribution of this Markov chain is the posterior distribution that we desire. We must, therefore, assure ourselves that we have reached the steady state. One way to assess whether the Markov chain has reached the steady state is to look at the sequence of values of the Markov chain. If these look like they have settled down to a single probability distribution, then we can infer that we have reached the steady state. There are more precise methods to make this inference, such as the Gelman-Rubin statistic, but we won't go into that here.

The plot of successive values of the Markov chain is called a *trace plot* and for the simple case described here is shown in Figure 16.21. The top plot shows the first 50 simulations, and we can see that there were several cases in the first 50 where the algorithm rejected the proposal (as indicated by consecutive identical points). Points 3, 4, and 5 are identical, indicating that twice in a row we rejected the proposal. The top part of Figure 16.21 indicates rejections with a circle around consecutive identical points. Overall, in the 11,000 simulations we rejected about 21% of the time. From the middle graph, which shows all 11,000 simulations, it appears that the steady state is reached almost immediately when the chain begins. To be safe, we discard the first 1,000 simulations, called the burn-in and keep simulations 1,001 through 11,000, for a total of 10,000 simulations. The bottom part of the graph shows a histogram of these 10,000 simulations in red. This is what we believe the posterior distribution looks like. In this problem we knew that the posterior is a BETA(3,7) distribution so we can plot the (scaled) posterior distribution with the histogram. As the bottom figure shows, the histogram is a very close fit to the true posterior. If we wanted to estimate the probability θ by taking the posterior mean, we would simply average the values of the 10,000 simulations that we kept. This turns out to be 0.299, very close to the mean of the posterior we derived analytically since the mean of the BETA(3,7) distribution is $3/(3+7) = 0.3$. This was from only 10,000 simulations.

Keep in mind that in most realistic cases, the analytic form of the posterior will be unknown and we will have to rely solely on the simulated values.

1.7.2 The Gibbs Sampling Algorithm

The other form of MCMC simulation is called *Gibbs sampling*. We explain Gibbs sampling in the context of a three-parameter problem; the extension from three to p dimensions should be apparent. To implement Gibbs sampling, we must know the conditional distribution of each parameter given all of the other parameters and the data.

Algorithm (Metropolis-Hastings Algorithm for Simulating from a Posterior Distribution)**.** *Suppose we want to simulate from a posterior distribution* $p(\theta_1, \theta_2, \theta_3|\boldsymbol{x})$. *Let* $(\theta_1^{(0)}, \theta_2^{(0)}, \theta_3)^{(0)}$ *denote a starting value for* $(\theta_1, \theta_2, \theta_3)$. *Set* $i = 1$.

Step 1: *Simulate one value from the conditional distribution*

$$p(\theta_1|\theta_2^{(i-1)}, \theta_3^{(i-1)}, \boldsymbol{x})$$

and call this simulated value $\theta_1^{(i)}$.

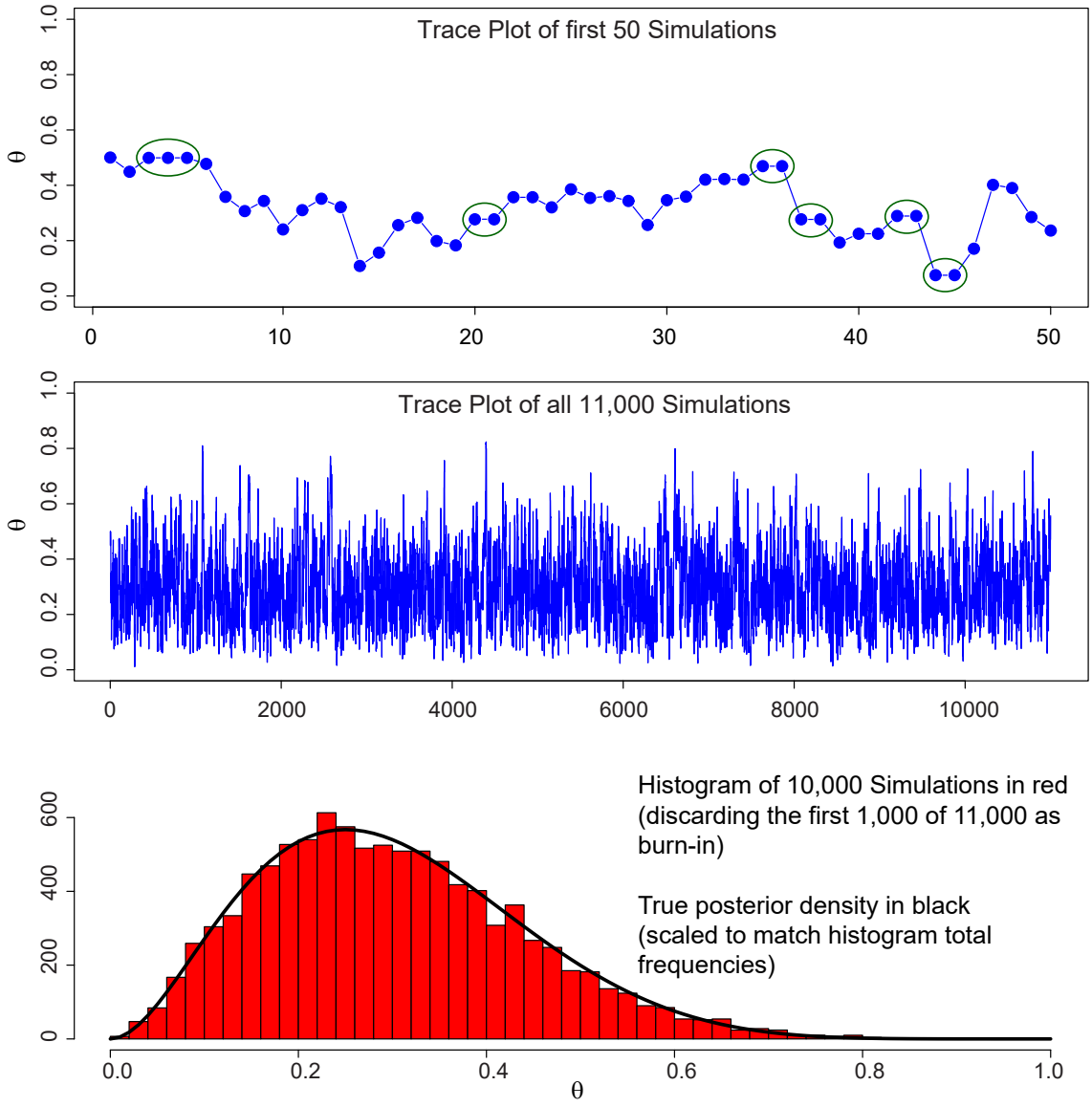


Figure 1.21: Results of the Metropolis algorithm applied to $\text{BIN}(6,\theta)$ data where we observed $x = 1$ success. The prior distribution was $\text{BETA}(2,2)$. The first 50 simulations are shown in the top figure, where we can see several places where the proposal was rejected and the algorithm stayed at the same level. Near the beginning, we rejected the move twice in a row, leading to three consecutive $\boldsymbol{\theta}^{(i)}$ that are identical. Places where the move was rejected are circled.

Step 2: *Simulate one value from the conditional distribution*

$$p(\theta_2|\theta_1^{(i)}, \theta_3^{(i-1)}, \boldsymbol{x})$$

and call this simulated value $\theta_2^{(i)}$.

Step 3: *Simulate one value from the conditional distribution*

$$p(\theta_3|\theta_1^{(i)}, \theta_2^{(i)}, \boldsymbol{x})$$

and call this simulated value $\theta_3^{(i)}$.

Step 4: *If the required number of simulations has been reached, then stop. Otherwise, set* $i \leftarrow i+1$ *and go to Step 1.*

In the Gibbs sampling algorithm we cycle through all p of the parameters in the parameter space $\theta_1, \theta_2, \ldots, \theta_p$, each time conditioning on all of the other components. However, we always condition on the most recent computed value. For example, when we simulate the new $\theta_2^{(i)}$ we condition on $\theta_1^{(i)}$ because it has already been sampled, but we condition on $\theta_3^{(i-1)}$ since it was last updated on the previous step. The same theory applies to Gibbs sampling: if the Markov chain

$$(\theta_1^{(1)}, \theta_2^{(1)}, \theta_3^{(1)}),\ (\theta_1^{(2)}, \theta_2^{(2)}, \theta_3^{(2)}),\ (\theta_1^{(3)}, \theta_2^{(3)}, \theta_3^{(3)}), \ldots$$

converges to a steady state, then the steady state distribution is the desired posterior distribution.

Both Gibbs sampling and the Metropolis-Hastings, in theory, converge so that the steady state distribution is the posterior distribution we are seeking. Thus, the choice of which to use is sometimes arbitrary and may depend on the complexity of the full conditional distributions. As we've mentioned, there are some software packages that will run the MCMC simulations automatically. Most of the older packages such as OpenBUGS, WinBUGS, and JAGS apply Gibbs sampling.[1]

More recent MCMC engines use enhancements of Gibbs sampling and the Metropolis-Hastings algorithm. Stan, named for the father of computer simulation, Stanislaw Ulam, uses what is called Hamiltonian dynamics which uses ideas from mechanics to jump more quickly to the region of high posterior probability. No matter what engine is used to perform the MCMC simulations, the important thing to keep in mind is that we have to sample sufficiently many times so that we can be confident that the steady state has been reached. Once this happens, we simulate additional runs to learn what the posterior distribution. In simple models, the posterior distribution can be derived analytically, but as models get more complex we must rely on simulation in order to gain information about the posterior.

1.8 Hierarchical Bayes Models

Most of what we've done so far involves data from some distribution and parameters of that distribution that must be estimated. This includes the i.i.d. case, such as

$$X_1,\ X_2,\ \ldots,\ X_n \sim \text{i.i.d. } \mathrm{N}(\mu, \sigma^2)$$

and even the non i.i.d. case such as

$$Y_i \overset{\text{ind.}}{\sim} \mathrm{N}(\beta_0 + \beta_1, \sigma^2).$$

[1]The BUGS in OpenBUGS and WinBUGS stands for "Bayesian Using Gibbs Sampling." JAGS stands for "Just Another Gibbs Sampler."

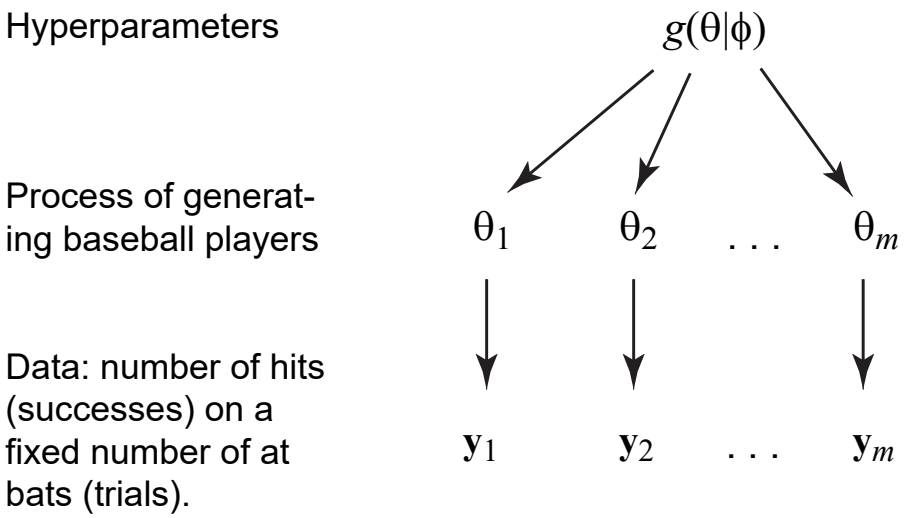


Figure 1.22: The data – process – parameters paradym.

The classical (frequentist) approach and the Bayesian approach diverge in the way we envision parameters. In the classical approach the parameters are fixed but unknown constants. Bayesians treat parameters as uncertain and assess the uncertainty in probabilistic terms.

There are situations where the observed variability is more complex than this simple data – parameters model would suggest. In many cases it is helpful to think of some process that is itself random and then view our data conditionally given this process. Such thinking leads us to hierarchical models, which we will treat from the Bayesian perspective. These ideas are presented in Figure 16.22.

To make these ideas concrete, let's consider a problem from baseball. All you need to know about baseball to understand this problem is that players attempt to get a hit every time they go to bat. Attempts are called "at bats" and successes are called "hits." A player will often get several hundred at bats in a given season. A player's "batting average" is the number of hits divided by the number of at bats. You should recognize that this sounds like the idea of Bernoulli trials we discussed in Chapter 6. The assumption of constant probability of success is questionable, because the chance of getting a hit depends of the strength of the opposing pitcher; good pitchers allow fewer hits. We will assume the usual conditions behind Bernoulli trials for each player.

Let θ_i denote the probability that player i will get a hit on any at bat. The problem is this. Given a very limited amount of data at the beginning of the season, estimate the (unknown) probability θ_i. This of course varies from player to player, with the number being higher for better players. An average player will have a batting average (H/AB) of about 0.250 (i.e., they get a hit about one time if four at bats), but good players will have a batting average (H/AB) of 0.300 or higher.

Let n_i denote the number of at bats (trials) and let X_i denote the number of hits (successes). Early in the year, players will not have many at bats, that is, n_i will be small and as you know the simple estimate $\hat{\theta}_i = X_i/n_i$ has a lot of variability. We consider here the data from the 2018 Major League Baseball season. One player from the Pittsburgh Pirates named Austin Meadows had 18 hits in 44 at bats on May 31, 2018, which is about one-third of the way through the season. This was the highest batting average in Major League Baseball at that time. The at bat total of 44 is low even for this early in the season. Many players will have had 150 to 200 at bats at this point. Should we believe that he is a 0.400 hitter? No player in Major League Baseball has hit 0.400 or higher since Ted Williams in 1941, so it is likely that his true θ_j is quite a bit less than 0.400.

As it turned out, his batting average at the end of the season was 0.287, higher than normal, but much closer to the average of all players. Conversely, a player who is at the low end on May 31, is likely to rebound and end up with a higher batting average. Thus, the end-of-season averages are shrunk toward the overall

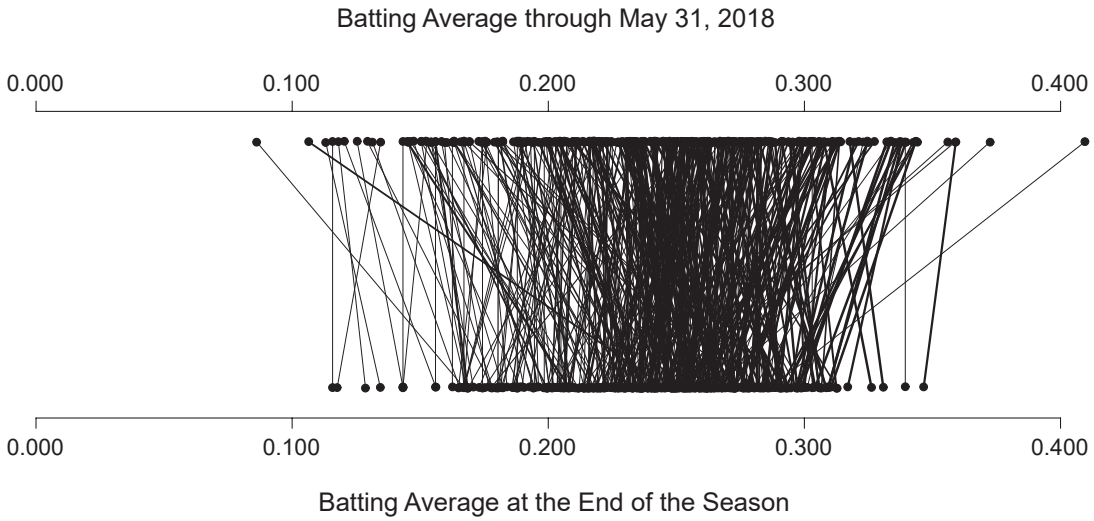


Figure 1.23: Batting averages of all players on May 31, 2018 (top) compared to the end-of-season averages (bottom). Lines connect the same player's two numbers.

mean. Figure 16.23 illustrates the situation. The top shows the batting averages of all 419 players who had at least 25 at bats by May 31, and the bottom shows the end-of-season averages. Lines connect the player's May 31 average with his end-of-season average. We can clearly see this shrinkage toward the mean.

So, how would we model this situation? Let's envision a process that produces baseball players. Specifically, the process produces the probabilities of players getting hits; this is what we've called $\theta_1,\ \theta_2,\ \cdots \theta_n$. These values are unobserved and unobservable. Statisticians call these effects *latent*. We assume that these come from some distribution $p(\theta|\phi)$ where ϕ is itself unknown and possibly a vector of parameters. Even though we can't observe θ_j we can observe the outcomes of random variables that depend on θ_j.

Since $\theta_1,\ \theta_2,\ \cdots\ \theta_n$ are probabilities, we assume that they are selected from a BETA prior distribution. Since we don't know the parameters of this distribution we assume a BETA(a,b) distribution and treat a and b as unknown parameters. These parameters will themselves have prior distributions if we take the Bayesian approach. The θ_is represent the *process* and the parameters, called hyperparameters in a hierarchical model like this, are then a and b. The data consists of the number of hits for player i out of his n_i at bats. If we denote this by X_i, then we can assume $X_i \sim \text{BIN}(n_i,\theta_i)$. Our model can then be described as follows:

$$\begin{aligned}
(a,b) &\sim p(a,b)\\
(\theta_1,\ \theta_2,\ \cdots\ \theta_n)|(a,b) &\sim \\
i.i.d.\text{BETA}(a,b)X_i &\overset{\text{ind.}}{\sim} \text{BIN}(n_i,\theta_i), \qquad i = 1,2,\ldots,419.
\end{aligned}$$

Here $p(a,b)$ is the hyperprior for the parameters a and b which we took to be diffuse exponential distributions.

We sent the code for the model into an MCMC engine (we used JAGS) and turned it loose. The JAGS code is

```
model #   Likelihood    for (i in 1:n)          H[i] ~ dbin( theta[i] , AB[i] )    #   Prior    for
```

We used an interface from R to JAGS run the JAGS code with the right data and starting values for the parameters. Readers interested in the details should consult one of the many books on Bayesian statistics that cover MCMC engines such as JAGS, WinBUGS, and Stan.

In hierarchical models like this, the hyperparameters are often the hardest to get to converge. Figure 16.24 shows the trace plots for a and b on the left side. As you can see, the successive values of the simulations are not independent. When a value is on the high side, the next few tend to also be on the high side. Same for lower values. This phenomenon is called *autocorrelation* or *serial correlation* and is discussed more

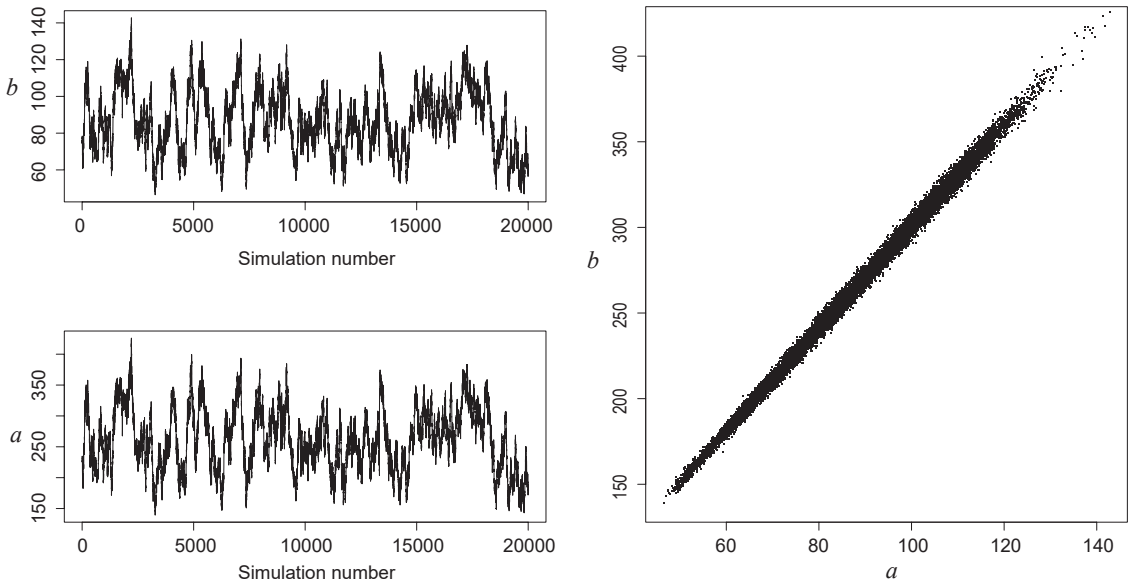


Figure 1.24: Hyperparameters in baseball data.

thoroughly in the next chapter. When we have high autocorrelation like this, it takes many more simulations to provide satisfactory information about the posterior. What is often done in cases like this is that users take every kth simulation and discard all those in between. This is called *thinning*, and in fact, we thinned by taking every tenth simulation; had we not done this we would have even had stronger autocorrelation.

The right part of Figure 16.24 shows the pairs of (a,b) and shows the joint posterior distribution. There is clearly a very high correlation between a and b. Most of the joint posterior probability for a and b lies roughly along the line $b = 3a$, suggesting that we are fairly confident that b is about three times a, but we are rather uncertain about whether this is $a = 60,\ b = 180$ or $a = 120,\ b = 360$, or something in the middle. Recall that the mean of the beta distribution is $a/(a+b)$, and if $b \approx 3a$, then

$$\mu \approxeq \frac{a}{a+b} = \frac{a}{a+3a} = \frac{1}{4} = 0.250.$$

Thus, we're pretty confident that the average across all players is about 0.250, but there is some uncertainty about the spread of θs about this number.

We applied Gibbs sampling using JAGS in this 421-dimensional problem (419 θs and the two hyperparameters, a and b). This means that the algorithm cycled through all 421 parameters, simulating from the conditional distribution of each parameter given the most recently simulated value from the other 420 parameters. It then did this 200,000 times, thinning by a factor of 10, to yield the simulated values of the Markov chain. This was done after a burn-in of 10,000 simulations. The estimates of the parameters were taken to be the posterior means, which was estimated by the average of the simulations after the burn-in. This yielded

$$\hat{a} = 86.7364 \quad \text{and} \quad \hat{b} = 261.0668.$$

The posterior means of the 20,000 saved values of the Markov chain were then used as point estimates for the θs. Austin Meadows, mentioned earlier had a batting average of 0.409 on May 31, and his Bayes estimate of the probability of getting a hit turned out to be 0.270.

If, on the morning of June 1, 2018, we were to make a prediction about what his average would be at the end of the year, we could go with one of two predictions: his current average 0.409, or his hierarchical

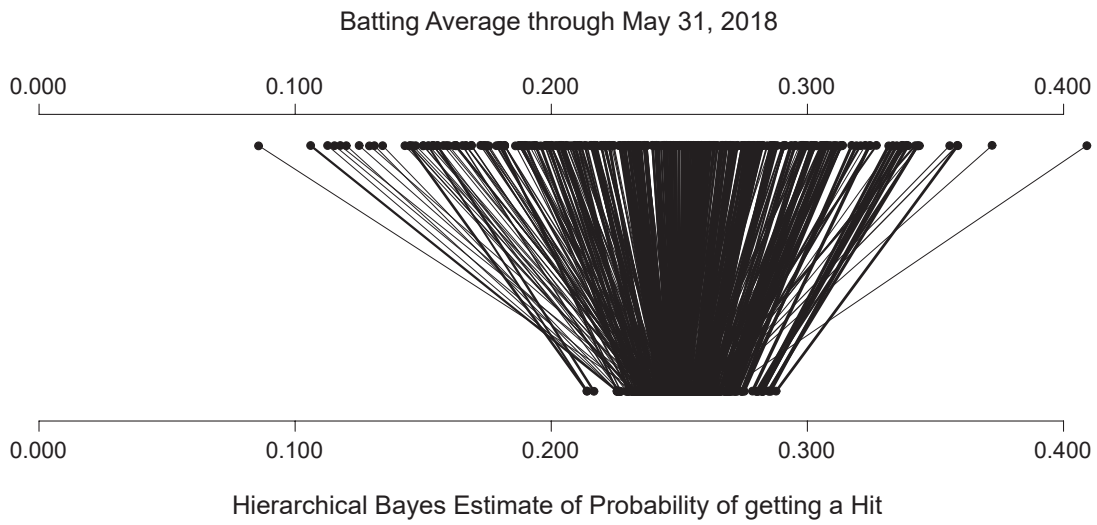


Figure 1.25: Results of hierarchical Bayes model for batting averages. Point estimates for θ_i, $i = 1,2,\ldots,419$ are taken to be the posterior means.

Bayes estimate of 0.270. If we use data from Austin Meadows performance, and his performance only, we would have no other data except his current average of 0.409; this would be our prediction. If we take into account the variability of major league baseball players and the process that generates these, we could apply the hierarchical Bayes method and shrink his, and all other estimates of θ_i toward the overall mean. This would yield a prediction of 0.270, much closer to the overall mean of 0.250. His actual batting average at the end of the year was 0.287, so our hierarchical Bayes estimate of 0.270 was much closer to his end-of-season average than his average of 0.409 on May 31.

Across all 419 players with 25 or more at bats on or before May 31, the mean squared prediction error using the hierarchical Bayes method is

$$\text{MSPE} = \sum_{i=1}^{419} (\hat{\theta}_i^{(\text{HB})} - \theta^{(\text{end})})^2 = 0.00097$$

and the mean squared prediction error using the May 31 average as the predictor is

$$\text{MSPE} = \sum_{i=1}^{419} (\hat{\theta}_i^{(\text{May 31})} - \theta^{(\text{end})})^2 = 0.00113.$$

Thus, using the hierarchical Bayes estimate to predict a player's average at the end of the season, given data through May 31, does a slightly better job of predicting the end-of-season average than the players batting average on May 31.

■ **Example 1.21** A school district is studying the effect of the school and the teacher on student's performance. Suppose that a school district has m schools, and within school i, $i = 1,2,\ldots,m$ there are t_i teachers. Within teacher j's class ($j = 1,2,\ldots,t_j$) there are n_{ij} students. Describe a hierarchical model that might be used to model the effects due to schools and teachers on student performance.

Solution: Let $Y_{i,j,k}$ denote the score of student k in the class of teacher j within school i. We might assume that, given the teacher and the school, the scores might be normally distributed with some mean (which depends on the school and teacher) and a constant variance across all schools and teachers. That is,

$$Y_{i,j,k} \overset{\text{ind.}}{\sim} \text{N}(\delta_{jk}, \sigma^2).$$

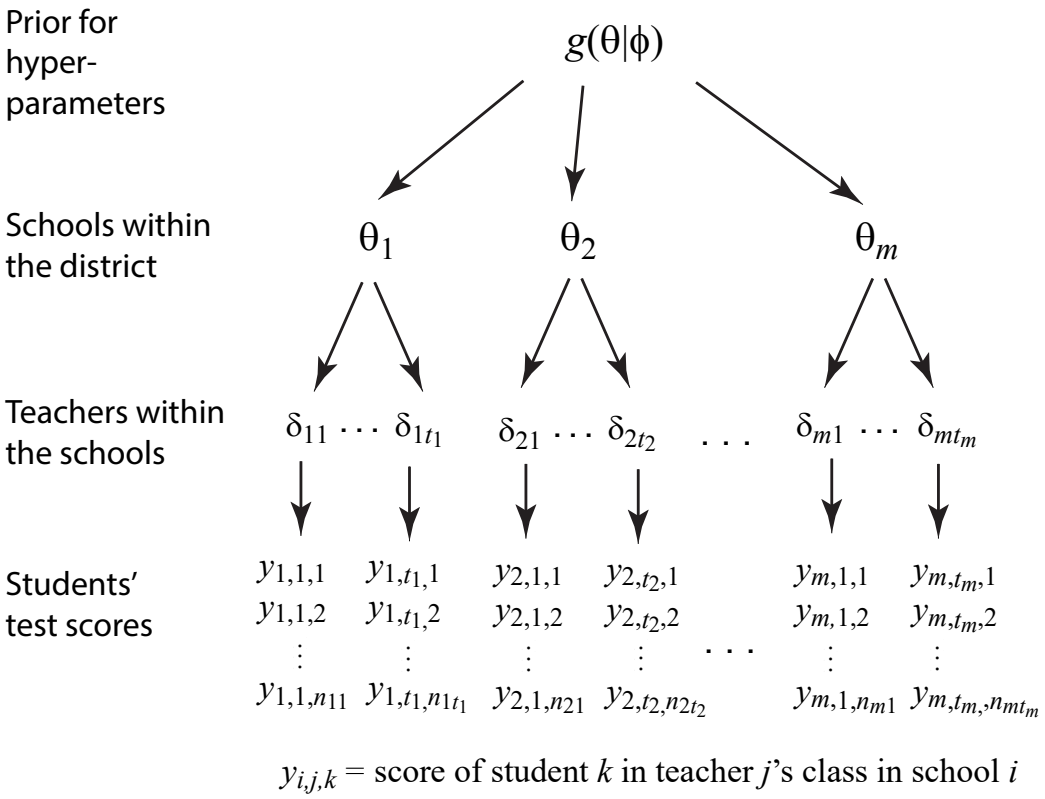


Figure 1.26: Hierarchical structure for study on students' achievement. There are m schools within a district and t_i teachers in school i. There are n_{ij} students in teacher j's class within school i.

Within school j we might assume that the δs within school j make up a random sample:

$$\delta_{j1},\ \delta_{j2},\ \ldots,\ \delta_{jt_j},\sim \text{ i.i.d. } \mathrm{N}(\theta_j, \sigma_{(\mathrm{t})}).$$

The schools then make up a random sample from a normal distribution:

$$\theta_1,\ \theta_2,\ \ldots,\ \theta_m \sim \text{ i.i.d. } \mathrm{N}(\mu, \sigma^2_{(\mathrm{s})}).$$

Finally, if we take a Bayesian approach, we would assign priors to the remaining parameters, σ^2 and $\sigma_{(\mathrm{t})}$, as well as the hyperparameters μ and $\sigma^2_{(\mathrm{s})}$.

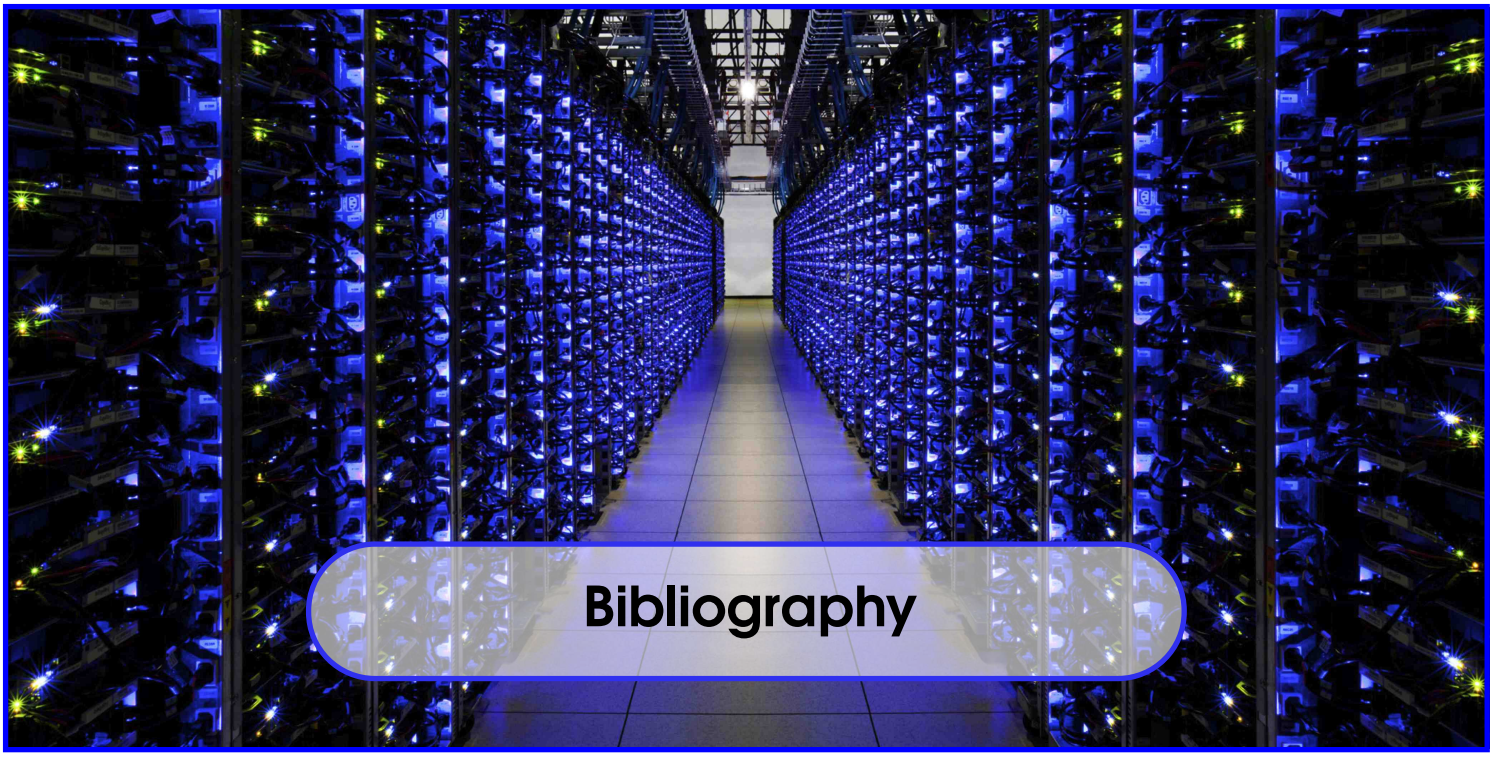
This model could then be run using an MCMC engine such as JAGS. All of the parameters, including all of the δ_{jk}s and θ_is, which are called the *random effects* as well as the parameters μ, σ^2, $\sigma_{(\mathrm{t})}$, and $\sigma^2_{(\mathrm{s})}$, would be estimated in one overall model. The study might focus on the variability of the schools, measured by $\sigma^2_{(\mathrm{s})}$ and by the variability of the teachers within each school.

■

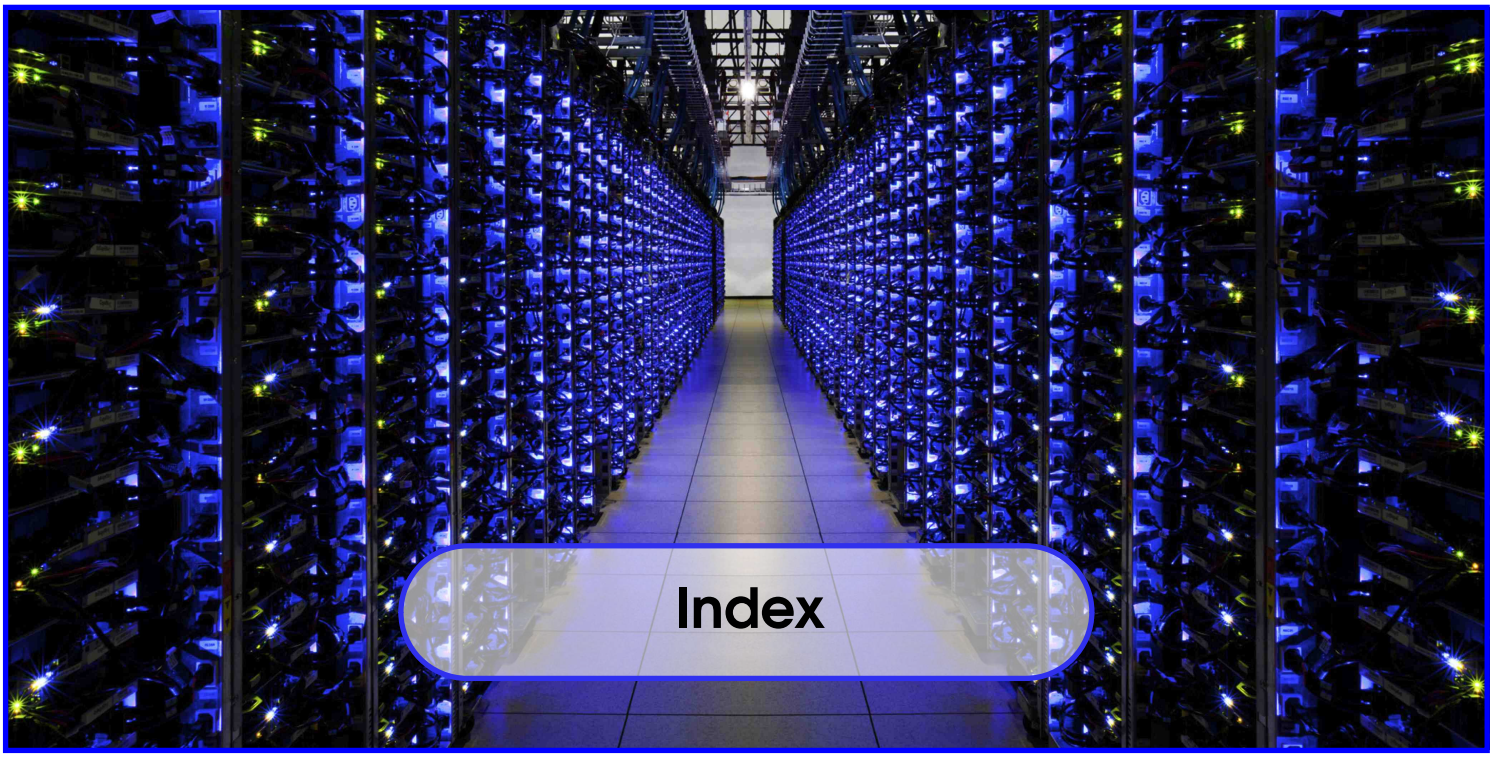
Problem 1.1 What is the average airspeed velocity of an unladen swallow?

Problem 1.2 Refer to Example 16.17. Fill in the steps necessary to obtain the result that

$$E\left(-\frac{\partial^2}{\partial\theta^2}\log f(X|\theta\right) = \frac{n}{\theta(1-\theta)}$$



Bibliography



Index

F distribution, 152
t-distribution, 154

autocorrelation, 47
autocorrelation plot, 47
average, 28
    moving, 45
    population, 29
    sample, 28

bar chart, 54
    side-by-side, 57
    stacked, 54
Bayes' Theorem, 107
bivariate normal distribution, 151
box plot, 65
    repeated, 74

categorical data, 24
Central Limit Theorem (CLT), 160
chi-square distribution, 152
combinations, 102
complementary events, 94
correlation
    sample, 37
covariance
    sample, 36
cumulative density function (cdf), 121, 147

data
    categorical, 24
    continuous, 24
    discrete, 24
    qualitative, 24
    quantitative, 24
descriptive statistics, 6
disjoint events, 90
distribution
    F, 152
    χ^2, 152
    t, 154
    Bernoulli, 120
    binomial, 132
    chi-square, 152
    joint, 151
    negative binomial, 143
    normal, 149
        bivariate, 151
        multivariate, 151
        standard, 149
    Poisson, 144
    sampling, 159
    uniform, 148

error
    measurement, 6
    sampling, 6
events
    complementary, 94

disjoint, 90
independent, 92
expected value, 125

histogram, 58

independence, 93
indicator variable, 24
inferential statistics, 6
interquartile range (IQR), 41

joint distribution, 151

lag, 47
lattice plot, 63
Law of Total Probability, 107

mean, 28
population, 29
sample, 28
trimmed, 31
measurement error, 6
median, 30
moving average, 45
multivariate normal distribution, 151

negative binomial random variable, 143
normal
distribution, 149
random variable, 149

percentile, 39
permutations, 102
pie charts, 58
Poisson random variable, 144
population, 5
probability
axioms, 89
combinations, 102
conditional, 95
density function (pdf), 147
function ($\mathbb{P}$), 89
independence, 92
permutations, 102

qualitative data, 24
quantile, 41
quantitative data, 24
quartile, 41

random sample, 5
random variable, 117
continuous, 147
negative binomial, 143
normal, 149
Poisson, 144
uniform, 148
range, 35

sample, 5
random, 5
sample average, 28
sample mean, 28
sample space (S), 88
sampling distribution, 159
sampling error, 6
scatterplot, 69
standard deviation, 129
sample, 34
standard error, 159
standard normal distribution, 149
standardize, 151
statistic, 6
statistics
descriptive, 6
inferential, 6

time series, 73
trellis plot, 63
trimmed mean, 31

uniform random variable, 148
union, 90

variable
indicator, 24
variance, 129
sample, 33
Venn diagram, 90