WILEY

INTRODUCTION TO

BAYESIAN STATISTICS

WILLIAM M. BOLSTAD

WWW. LINK AVAILABLE

Introduction to
Bayesian Statistics

Introduction to Bayesian Statistics

William M. Bolstad
University of Waikato
Hamilton, New Zealand

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Bolstad, William M., 1943–
Introduction to Bayesian statistics / William M. Bolstad.
p. cm.
Includes bibliographic references and index.
ISBN 0-471-27020-2 (cloth)
1. Bayesian statistical decision theory. I. Title.

QA279.5.B65 2004
519.5'42—dc22 2003057660

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

This book is dedicated to

Sylvie,
Ben, Rachel,
Mary, and Elizabeth

Contents

Preface

How This Text Was Developed

This text grew out of the course notes for an Introduction to Bayesian Statistics course that I have been teaching at the University of Waikato for the past few years. My goal in developing this course was to introduce Bayesian methods at the earliest possible stage, and cover a similar range of topics as a traditional introductory statistics course. There is currently an upsurge in using Bayesian methods in applied statistical analysis, yet the Introduction to Statistics course most students take is almost always taught from a frequentist perspective. In my view, this is not right. Students with a reasonable mathematics background should be exposed to Bayesian methods from the beginning, because that is the direction applied statistics is moving.

Mathematical Background Required

Bayesian statistics uses the rules of probability to make inferences, so students must have good algebraic skills for recognizing and manipulating formulas. A general knowledge of calculus would be an advantage in reading this book. In particular, the student should understand that the area under a curve is found by integration, and that the location of a maximum or a minimum of a continuous differentiable function is found by setting the derivative function equal to zero and solving. The book is self-contained with a calculus appendix students can refer to. However, the actual calculus used is minimal.

Features of the Text

In this text I have introduced Bayesian methods using a step by step development from conditional probability. In Chapter 4, the universe of an experiment is set up with two dimensions, the horizontal dimension is observable, and the vertical dimension is unobservable. Unconditional probabilities are found for each point in the universe using the multiplication rule and the prior probabilities of the unobservable events. Conditional probability is the probability on that part of the universe that occurred, the reduced universe. It is found by dividing the unconditional probability by their sum over all the possible unobservable events. Because of way the universe is organized, this summing is down the column in the reduced universe. The division scales them up so the conditional probabilities sum to one. This result known as *Bayes' theorem* is the key to this course. In Chapter 6 this pattern is repeated with the Bayesian universe. The horizontal dimension is the sample space, the set of all possible values of the observable random variable. The vertical dimension is the parameter space, the set of all possible values of the unobservable parameter. The reduced universe is the vertical slice that we observed. The conditional probabilities given what we observed are the unconditional probabilities found by using the multiplication rule $(prior \times likelihood)$ divided by their sum over all possible parameter values. Again, this sum is taken down the column. The division rescales the probabilities so they sum to one. This gives Bayes' theorem for a discrete parameter and a discrete observation. When the parameter is continuous, the rescaling is done by dividing the joint probability-probability density function at the observed value by its integral over all possible parameter values so it integrates to one. Again, the joint probability-probability density function is found by the multiplication rule and at the observed value is $(prior \times likelihood)$. This is done for binomial observations and a continuous beta prior in Chapter 8. When the observation is also a continuous random variable, the conditional probability density is found by rescaling the joint probability density at the observed value by dividing by its integral over all possible parameter values. Again, the joint probability density is found by the multiplication rule and at the observed value is $prior \times likelihood$. This is done for normal observations and a continuous normal prior in Chapter 10. All these cases follow the same general pattern.

Bayes' theorem allows one to revise his/her belief about the parameter, given the data that occurred. There must be a prior belief to start from. One's prior distribution gives the relative belief weights he/she has for the possible values of the parameters. How to choose ones prior is discussed in detail. Conjugate priors are found by matching first two moments with prior belief on location and spread. When the conjugate shape does not give satisfactory representation of prior belief, setting up a discrete prior and interpolating is suggested.

Details that I consider beyond the scope of this course are included as footnotes. There are many figures that illustrate the main ideas, and there are many fully worked out examples. I have included chapters comparing Bayesian methods with the corresponding frequentist methods. There are exercises at the end of each chapter, some with short answers. In the exercises, I only ask for the Bayesian methods to be

used, because those are the methods I want the students to learn. There are computer exercises to be done in Minitab or R using the included macros. Some of these are small-scale Monte Carlo studies that demonstrate the efficiency of the Bayesian methods evaluated according to frequentist criteria.

Advantages of the Bayesian Perspective

Anyone who has taught an Introduction to Statistics class will know that students have a hard time coming to grips with statistical inference. The concepts of hypothesis testing and confidence intervals are subtle and students struggle with them. Bayesian statistics relies on a single tool, Bayes' theorem to revise our belief given the data. This is more like the kind of plausible reasoning that students use in their everyday life, but structured in a formal way. Conceptually it is a more straightforward method for making inferences. The Bayesian perspective offers a number of advantages over the conventional frequentist perspective.

- The "objectivity" of frequentist statistics has been obtained by disregarding any prior knowledge about the process being measured. Yet in science there usually is some prior knowledge about the process being measured. Throwing this prior information away is wasteful of information (which often translates to money). Bayesian statistics uses both sources of information; the prior information we have about the process and the information about the process contained in the data. They are combined using Bayes' theorem.

- The Bayesian approach allows direct probability statements about the parameters. This is much more useful to a scientist than the confidence statements allowed by frequentist statistics. This is a very compelling reason for using Bayesian statistics. Clients will interpret a frequentist confidence interval as a probability interval. The statistician knows that that interpretation is not correct but also knows that the confidence interpretation relating the probability to all possible data sets that could have occurred, but didn't; is of no particular use to the scientist. Why not use a perspective that allows them to make the interpretation that is useful to them.

- Bayesian statistics has a single tool, Bayes' theorem, which is used in all situations. This contrasts to frequentist procedures, which require many different tools.

- Bayesian methods often outperform frequentist methods, even when judged by frequentist criteria.

- Bayesian statistics has a straightforward way of dealing with nuisance parameters. They are always marginalized out of the joint posterior distribution.

- Bayes' theorem gives the way to find the predictive distribution of future observations. This is not always easily done in a frequentist way.

These advantages have been well known to statisticians for some time. However, there were great difficulties in using Bayesian statistics in actual practice. While it is easy to write down the formula for the posterior distribution,

$$g(\theta|data) = \frac{g(\theta) \times f(data|\theta)}{\int g(\theta) \times f(data|\theta)\, d\theta} \quad ,$$

a closed form existed only in a few simple cases, such as for a normal sample with a normal prior. In other cases the integration required had to be done numerically. This in itself made it more difficult for beginning students. If there were more than a few parameters, it became extremely difficult to perform the numerical integration.

In the past few years, computer algorithms (e.g., the Gibbs Sampler and the Metropolis-Hasting algorithm) have been developed to draw an (approximate) random sample from the posterior distribution, without having to completely evaluate it. We can approximate the posterior distribution to any accuracy we wish by taking a large enough random sample from it. This removes the disadvantage of Bayesian statistics, for now it can be done in practice for problems with many parameters, and for distributions from general samples and having general prior distributions. Of course these methods are beyond the level of an introductory course. Nevertheless, we should be introducing our students the approach to statistics that gives the theoretical advantages from the very start. That is how they will get the maximum benefit.

Outline of a Course Based on This Text

At the University of Waikato we have a one-semester course based on this text. This course consists of 36 one-hour lectures, 12 one-hour tutorial sessions, and several computer assignments. In each tutorial session, the students work through a statistical activity in a hands-on way. Some of the computer assignments involve Monte Carlo studies showing the long run performance of statistical procedures.

- Chapter 1 (one lecture) gives an introduction to the course.
- Chapter 2 (three lectures) covers scientific data gathering including random sampling methods and the need for randomized experiments to make inferences on cause-effect relationships.
- Chapter 3 (two lectures) is on data analysis with methods for displaying and summarizing data. If students have already covered this material in a previous statistics course, this could be covered as a reading assignment only.
- Chapter 4 (three lectures) introduces the rules of probability including joint, marginal, and conditional probability and shows Bayes' theorem is the best method for dealing with uncertainty.
- Chapter 5 (two lectures) introduces discrete and random variables.

- Chapter 6 ((three lectures) shows how Bayesian inference works for an discrete random variable with a discrete prior.
- Chapter 7 (two lectures) introduces continuous random variables.
- Chapter 8 (three lectures) shows how inference is done on the population proportion from a binomial sample using either a uniform or a beta prior. There is discussion on choosing a beta prior that corresponds to your prior belief, and graphing it to confirm that it fits your belief.
- Chapter 9 (three lectures) compares the Bayesian inferences for the proportion with the corresponding frequentist ones. The Bayesian estimator for the proportion is compared with the corresponding frequentist estimator in terms of mean squared error. The difference between the interpretations of Bayesian credible interval and the frequentist confidence interval are discussed.
- Chapter 10 (four lectures) introduces Bayes' theorem for the mean of a normal distribution, using either a "flat" improper prior or a normal prior. There is considerable discussion on choosing a normal prior, and graphing it to confirm it fits with your belief. The predictive distribution of the next observation is developed. *Student's t* distribution is introduced as the adjustment required for the credible intervals when the standard deviation is estimated from the sample. Section 10.5 is at a higher level, and may be omitted.
- Chapter 11 (one lecture) compares the Bayesian inferences for mean with the corresponding frequentist ones.
- Chapter 12 (three lectures) does Bayesian inference for the difference between two normal means, and the difference between two binomial proportions using the normal approximation.
- Chapter 13 (three lectures) does simple linear regression model in a Bayesian manner. Section 13.5 is at a higher level, and may be omitted.
- Chapter 14 (three lectures) introduces robust Bayesian methods using mixture priors. This chapter shows how to protect against misspecified priors, which is one of the main concerns that many people have against using Bayesian statistics. It is at a higher level than the previous chapters and could be omitted and more lecture time given to the other chapters.

Acknowledgements

I would like to acknowledge the help I have had from many people. First, my students over the past three years, whose enthusiasm with the early drafts encouraged me to continue writing. My colleague, James Curran for writing the R macros, and Appendix D on how to implement them, and giving me access to the glass data. Ian Pool, Dharma Dharmalingam and Sandra Baxendine from the University of

Waikato Population Studies Centre for giving me access to the NZFEE data. Fiona Petchey from the University of Waikato Carbon Dating Unit for giving me access to the ^{14}C archeological data. Lance McKay from the University of Waikato Biology Department for giving me access to the slug data. Graham McBride from NIWA for giving me access to the New Zealand water quality data. Harold Henderson and Neil Cox from AgResearch NZ for giving me access to the ^{13}C enriched Octanoic acid breath test data, and the endophyte data. Martin Upsdell from AgResearch NZ made some useful suggestions on an early draft. Renate Meyer from the University of Auckland gave me useful comments on the manuscript. My colleagues Lyn Hunt, Judi McWhirter, Murray Jorgensen, Ray Littler, Dave Whitaker and Nye John for their support and encouragement through this project. Alec Zwart and Stephen Joe for help with LaTeX, and Karen Devoy for her secretarial assistance.

I would like to also thank my editor Rosalyn Farkas at John Wiley & Sons, and Amy Hendrixson, of TeXnology Inc. for their patience and help through the process from rough manuscript to camera-ready copy.

Finally, last but not least, I wish to thank my wife Sylvie for her constant love and support and for her help on producing some of the figures.

WILLIAM M. "BILL" BOLSTAD

Hamilton, New Zealand

1

Introduction to Statistical Science

Statistics is the science that relates data to specific questions of interest. This includes devising methods to gather data relevant to the question, methods to summarize and display the data to shed light on the question, and methods that enable us to draw answers to the question that are supported by the data. Data almost always contain uncertainty. This uncertainty may arise from selection of the items to be measured, or it may arise from variability of the measurement process. Drawing general conclusions from data is the basis for increasing knowledge about the world, and is the basis for all rational scientific inquiry. *Statistical inference* gives us methods and tools for doing this despite the uncertainty in the data. The methods used for analysis depend on the way the data were gathered. It is vitally important that there is a probability model explaining how the uncertainty gets into the data.

Showing a Causal Relationship from Data

Suppose we have observed two variables X and Y. Variable X appears to have an association with variable Y. If high values of X occur with high values of variable Y and low values of X occur with low values of Y, we say the association is positive. On the other hand, the association could be negative in which high values of variable X occur in with low values of variable Y. Figure 1.1 shows a schematic diagram where the association is indicated by the dotted curve connecting X and Y. The unshaded area indicates that X and Y are observed variables. The shaded area indicates that there may be additional variables that have not been observed.

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

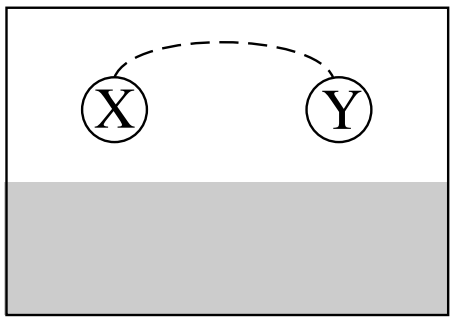


Figure 1.1 Association between two variables.

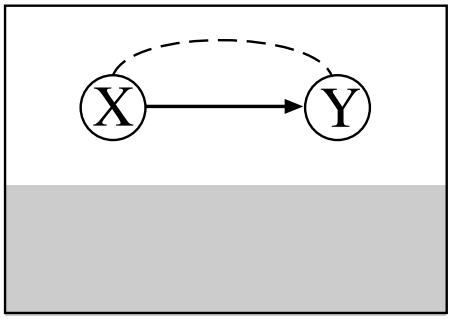


Figure 1.2 Association due to causal relationship.

We would like to determine why the two variables are associated. There are several possible explanations. The association might be a causal one. For example, X might be the cause of Y. This is shown in Figure 1.2, where the causal relationship is indicated by the arrow from X to Y.

On the other hand, there could be an unidentified third variable Z that has a causal effect on both X and Y. They are not related in a direct causal relationship. The association between them is due to the effect of Z. Z is called a *lurking* variable, since it is hiding in the background and it affects the data. This is shown in Figure 1.3.

It is possible that both a causal effect and a lurking variable may both be contributing to the association. This is shown in Figure 1.4. We say that the causal effect and the effect of the lurking variable are *confounded*. This means that both effects are included in the association.

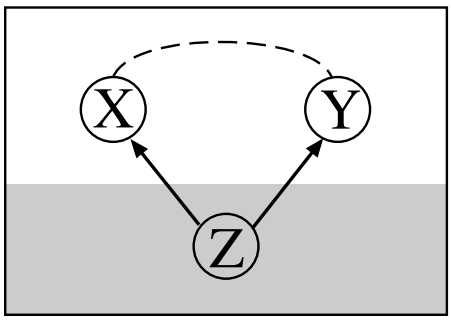


Figure 1.3 Association due to lurking variable.

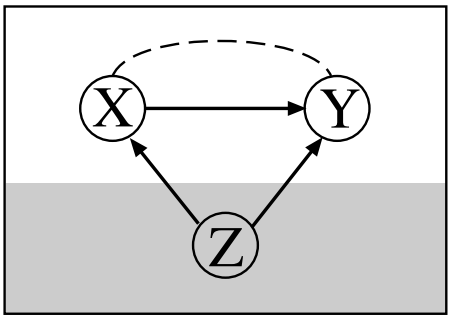


Figure 1.4 Confounded causal and lurking variable effects.

Our first goal is to determine which of the possible reasons for the association holds. If we conclude that it is due to a causal effect, then our next goal is to determine the size of the effect. If we conclude that the association is due to causal effect confounded with the effect of a lurking variable, then our next goal becomes determining the sizes of both the effects.

1.1 THE SCIENTIFIC METHOD: A PROCESS FOR LEARNING

In the Middle Ages, science was deduced from principles set down many centuries earlier by authorities such as Aristotle. The idea that scientific theories should be tested against real world data revolutionized thinking. This way of thinking known as the scientific method sparked the Renaissance.

The scientific method rests on the following premises:

- A scientific hypothesis can never be shown to be absolutely true.

- However, it must potentially be disprovable.
- It is a useful model until it is established that it is not true.
- Always go for the simplest hypothesis, unless it can be shown to be false.

This last principle, elaborated by William of Ockham in the 13^{th} century, is now known as "Ockham's razor" and is firmly embedded in science. It keeps science from developing fanciful overly elaborate theories. Thus the scientific method directs us through an improving sequence of models, as previous ones get falsified. The scientific method generally follows the following procedure:

1. Ask a question or pose a problem in terms of the current scientific hypothesis.
2. Gather all the relevant information that is currently available. This includes the current knowledge about parameters of the model.
3. Design an investigation or experiment that addresses the question from step 1. The predicted outcome of the experiment should be one thing if the current hypothesis is true, and something else if the hypothesis is false.
4. Gather data from the experiment.
5. Draw conclusions given the experimental results. Revise the knowledge about the parameters to take the current results into account.

The scientific method searches for cause and effect relationships between an experimental variable and an outcome variable. In other words, how changing the experimental variable results in a change to the outcome variable. Scientific modelling develops mathematical models of these relationships. Both of them need to isolate the experiment from outside factors that could affect the experimental results. All outside factors that can be identified as possibly affecting the results must be controlled. It is no coincidence that the earliest successes for the method were in physics and chemistry where the few outside factors could be identified and controlled. Thus there were no *lurking* variables. All other relevant variables could be identified, and physically controlled by being held constant. That way they would not affect results of the experiment, and the effect of the experimental variable on the outcome variable could be determined. In biology, medicine, engineering, technology, and the social sciences it isn't that easy to identify the relevant factors that must be controlled. In those fields a different way to control outside factors, because they can't be identified beforehand and physically controlled.

1.2 THE ROLE OF STATISTICS IN THE SCIENTIFIC METHOD

Statistical methods of inference can be used when there is *random* variability in the data. The probability model for the data is justified by the design of the investigation or

experiment. This can extend the scientific method into situations where the relevant outside factors cannot even be identified. Since we cannot identify these outside factors, we cannot control them directly. The lack of direct control means the outside factors will be affecting the data. There is a danger that the wrong conclusions could be drawn from the experiment due to these uncontrolled outside factors.

The important statistical idea of *randomization* has been developed to deal with this possibility. The unidentified outside factors can be "averaged out" by randomly assigning each unit to either treatment or control group. This contributes variability to the data. Statistical conclusions always have some uncertainty or error due to variability in the data. We can develop a probability model of the data variability based on the randomization used. Randomization not only reduces this uncertainty due to outside factors, it also allows us to measure the amount of uncertainty that remains using the probability model. Randomization lets us control the outside factors statistically, by averaging out their effects.

Underlying this is the idea of a statistical *population*, consisting of all possible values of the observations that could be made. The data consists of observations taken from a *sample* of the population. For valid inferences about the population *parameters* from the sample *statistics*, the sample must be "representative" of the population. Amazingly, choosing the sample randomly is the most effective way to get representative samples!

1.3 MAIN APPROACHES TO STATISTICS

There are two main philosophical approaches to statistics. The first is often referred to as the *frequentist* approach. Sometimes it is called the *classical* approach. Procedures are developed by looking at how they perform over all possible random samples. The probabilities don't relate to the particular random sample that was obtained. In many ways this indirect method places the "cart before the horse."

The alternative approach that we take in this book is the *Bayesian* approach. It applies the laws of probability directly to the problem. This offers many fundamental advantages over the more commonly used frequentist approach. We will show these advantages over the course of the book.

Frequentist Approach to Statistics

Most introductory statistics books take the frequentist approach to statistics, which is based on the following ideas:

- Parameters, the numerical characteristics of the population, are fixed but unknown constants.
- Probabilities are always interpreted as long run relative frequency.
- Statistical procedures are judged by how well they perform in the long run over an infinite number of hypothetical repetitions of the experiment.

Probability statements are only allowed for random quantities. The unknown parameters are fixed, not random, so probability statements cannot be made about their value. Instead, a sample is drawn from the population, and a sample statistic is calculated. The probability distribution of the statistic over all possible random samples from the population is determined, and is known as the *sampling distribution* of the statistic. The parameter of the population will also be a parameter of the sampling distribution. The probability statement that can be made about the statistic based on its sampling distribution is converted to a *confidence* statement about the parameter. The confidence is based on the average behavior of the procedure under all possible samples.

Bayesian Approach to Statistics

The Reverend Thomas Bayes first discovered the theorem that now bears his name. It was written up in a paper *An Essay Towards Solving a Problem in the Doctrine of Chances*. This paper was found after his death by his friend Richard Price, who had it published posthumously in the *Philosophical Transactions of the Royal Society* in 1763. Bayes showed how *inverse probability* could be used to calculate probability of antecedent events from the occurrence of the consequent event. His methods were adopted by Laplace and other scientists in the 19^{th} century, but had largely fallen from favor by the early 20^{th} century. By mid 20^{th} century interest in Bayesian methods was renewed by De Finetti, Jeffreys, Savage, and Lindley, among others. They developed a complete method of statistical inference based on Bayes' theorem.

This book introduces the Bayesian approach to statistics. The ideas that form the basis of the this approach are:

- Since we are uncertain about the true value of the parameters we will consider them a random variable.
- The rules of probability are used directly to make inferences about the parameters.
- Probability statements about parameters must be interpreted as "degree of belief." The *prior distribution* must be subjective. Each person can have his/her own prior, which contains the relative weights that person gives to every possible parameter value. It measures how "plausible" the person considers each parameter value to be before observing the data.
- We revise our beliefs about parameters after getting the data by using Bayes' theorem. This gives our *posterior distribution* which gives the relative weights we give to each parameter value after analyzing the data. The posterior distribution comes from two sources: the prior distribution and the observed data.

This has a number of advantages over the conventional frequentist approach. Bayes' theorem is the only consistent way to modify our beliefs about the parameters given the data that actually occurred. This means that the inference is based on the

actual occurring data, not all possible data sets that might have occurred, but didn't! Allowing the parameter to be a random variable lets us make probability statements about it, posterior to the data. This contrasts with the conventional approach where inference probabilities are based on all possible data sets that could have occurred for the fixed parameter value. Given the actual data there is nothing random left with a fixed parameter value, so one can only make *confidence* statements, based on what could have occurred. Bayesian statistics also has a general way of dealing with a *nuisance parameter* . A nuisance parameter is one which we don't want to make inference about, but we don't want them to interfere with the inferences we are making about the main parameters. Frequentist statistics does not have a general procedure for dealing with them. Bayesian statistics is predictive, unlike conventional frequentist statistics. This means that we can easily find the conditional probability distribution of the next observation given the sample data.

Monte Carlo Studies

In frequentist statistics, the parameter is considered a fixed, but unknown constant. A statistical procedure such as a particular estimator for the parameter cannot be judged from the value it takes given the data. The parameter is unknown, so we can't know the value it should be giving. If we knew the parameter value it was supposed to take, we wouldn't be using an estimator.

Instead, statistical procedures are evaluated by looking how they perform in the long run over all possible samples of data, for fixed parameter values over some range. For instance, we fix the parameter at some value. The estimator depends on the random sample, so it is considered a random variable having a probability distribution. This distribution is called the *sampling distribution* of the estimator, since its probability distribution comes from taking all possible random samples. Then we look at how the estimator is distributed around the parameter value. This is called sample space averaging. Essentially it compares the performance of procedures before we take any data.

Bayesian procedures consider the parameter to be a random variable, and its posterior distribution is conditional on the sample data that actually occurred, not all those samples that were possible, but did not occur. However, *before* the experiment, we might want to know how well the Bayesian procedure works at some specific parameter values in the range.

To evaluate the Bayesian procedure using sample space averaging, we have to consider the parameter to be both a random variable and a fixed but unknown value at the same time. We can get past the apparent contradiction in the nature of the parameter because the probability distribution we put on the parameter measures our uncertainty about the true value. It shows the relative belief weights we give to the possible values of the unknown parameter! After looking at the data, our belief distribution over the parameter values has changed. This way we can think of the parameter as fixed, but unknown value at the same time as we think of it being a random variable. This allows us to evaluate the Bayesian procedure using sample

space averaging. This is called *pre-posterior* analysis because it can be done before we obtain the data.

In Chapter 4, we will find out that the laws of probability are the best way to model uncertainty. Because of this, Bayesian procedures will be optimal in the post-data setting, given the data that actually occurred. In Chapters 9 and 11, we will see that Bayesian procedures perform very well in the pre-data setting when evaluated using *pre-posterior* analysis. In fact, it is often the case that Bayesian procedures outperform the usual frequentist procedures even in the pre-data setting.

Monte Carlo studies are a useful way to perform sample space averaging. We draw a large number of samples randomly using the computer and calculate the statistic (frequentist or Bayesian) for each sample. The empirical distribution of the statistic (over the large number of random samples) approximates its sampling distribution (over all possible random samples). We can calculate statistics such as mean and standard deviation on this Monte Carlo sample to approximate the mean and standard deviation of the sampling distribution. Some small-scale Monte Carlo studies are included as exercises.

1.4 PURPOSE AND ORGANIZATION OF THIS TEXT

A very large proportion of undergraduates are required to take a service course in statistics. Almost all of these courses are based on frequentist ideas. Most of them don't even mention Bayesian ideas. As a statistician, I know that Bayesian methods have great theoretical advantages. I think we should be introducing our best students to Bayesian ideas, from the beginning. There aren't many introductory statistics text books based on the Bayesian ideas. Some other texts include Berry (1996), Press (1989), and Lee (1989).

This book aims to introduce students with a good mathematics background to Bayesian statistics. It covers the same topics as a standard introductory statistics text, only from a Bayesian perspective. Students need reasonable algebra skills to follow this book. Bayesian statistics uses the rules of probability, so competence in manipulating mathematical formulas is required. Students will find that general knowledge of calculus is helpful in reading this book. Specifically they need to know that area under a curve is found by integrating, and that a maximum or minimum of a continuous differentiable function is found where the derivative of the function equals zero. However the actual calculus used is minimal. The book is self-contained with a calculus appendix students can refer to.

Chapter 2 introduces some fundamental principles of scientific data gathering to control the effects of unidentified factors. These include the need for drawing samples randomly, and some of random sampling techniques. The reason why there is a difference between the conclusions we can draw from data arising from an observational study and from data arising from a randomized experiment is shown. Completely randomized designs and randomized block designs are discussed.

Chapter 3 covers elementary methods for graphically displaying and summarizing data. Often a good data display is all that is necessary. The principles of designing displays that are true to the data are emphasized.

Chapter 4 shows the difference between deduction and induction. Plausible reasoning is shown to be an extension of logic where there is uncertainty. It turns out that plausible reasoning must follow the same rules as probability. The axioms of probability are introduced and the rules of probability, including conditional probability and Bayes' theorem are developed.

Chapter 5 covers discrete random variables, including joint and marginal discrete random variables. The *binomial* and *hypergeometric* distributions are introduced, and the situations where they arise are characterized.

Chapter 6 covers Bayes' theorem for discrete random variables using a table. We see that two important consequences of the method are that multiplying the prior by a constant, or that multiplying the likelihood by a constant do not affect the resulting posterior distribution. This gives us the "proportional form" of Bayes' theorem. We show that we get the same results when we analyze the observations sequentially using the posterior after the previous observation as the prior for the next observation, as when we analyze the observations all at once using the joint likelihood and the original prior. We show how to use Bayes' theorem for binomial observations with a discrete prior.

Chapter 7 covers continuous random variables, including joint, marginal, and conditional random variables. The *beta* and *normal* distributions are introduced in this chapter.

Chapter 8 covers Bayes' theorem for the population proportion (*binomial*) with a continuous prior. We show how to find the posterior distribution of the population proportion using either a *uniform* prior or a *beta* prior. We explain how to choose a suitable prior. We look at ways of summarizing the posterior distribution.

Chapter 9 compares the Bayesian inferences with the frequentist inferences. We show that the Bayesian estimator (posterior mean using a uniform prior) has better performance than the frequentist estimator (sample proportion) in terms of mean squared error over most of the range of possible values. This kind of frequentist analysis is useful before we perform our Bayesian analysis. We see the Bayesian credible interval has a much more useful interpretation than the frequentist confidence interval for the population proportion. One-sided and two-sided hypothesis tests using Bayesian methods are introduced.

Chapter 10 covers Bayes' theorem for the mean of a normal distribution with known variance. We show how to choose a normal prior. We discuss dealing with nuisance parameters by marginalization. The predictive density of the next observation is found by considering the population mean a nuisance parameter, and marginalizing it out.

Chapter 11 compares Bayesian inferences with the frequentist inferences for the mean of a normal distribution.

Chapter 12 shows how to perform Bayesian inferences for the difference between normal means and how to perform Bayesian inferences for the difference between proportions using the normal approximation.

Chapter 13 introduces the simple linear regression model, and shows how to perform Bayesian inferences on the slope of the model. The predictive distribution of the next observation is found by considering both the slope and intercept to be nuisance parameters, and marginalizing them out.

Chapter 14 shows how we can make Bayesian inference robust against a misspecified prior by using a mixture prior, and marginalizing out the mixture parameter. This chapter is at a somewhat higher level than the others, but it shows how one of the main dangers of Bayesian analysis can be avoided.

Main Points

- An association between two variables does not mean that one causes the other. It may be due to a causal relationship, it may be due to the effect of a third (lurking) variable on both the other variables, or a combination of a causal relationship and the effect of a lurking variable.

- Scientific method is a method for searching for cause-effect relationships, and measuring their strength. It uses controlled experiments, where outside factors that may effect the measurements are controlled. This isolates the relationship between the two variables from the outside factors, so the relationship can be determined.

- Statistical methods extend the scientific method to cases where the outside factors aren't identified, and hence can't be controlled. The principle of *randomization* is used to statistically control these unidentified outside factors by averaging out their effects. This contributes to *variability* in the data.

- We can use the probability model (based on the randomization method) to measure the uncertainty.

- The frequentist approach to statistics considers the parameter to be a fixed but unknown constant. The only kind of probability allowed is long run relative frequency. These probabilities are only for observations and sample statistics, given the unknown parameters. Statistical procedures are judged by how they perform in an infinite number of hypothetical repetitions of the experiment.

- The Bayesian approach to statistics allows the parameter to be considered a random variable. Probabilities can be calculated for parameters as well as observations and sample statistics. Probabilities calculated for parameters are interpreted as "degree of belief," and must be subjective. The rules of probability are used to revise our beliefs about the parameters, given the data.

- Frequentist estimators are evaluated by looking at their sampling distribution for a fixed parameter value, and how it is distributed over all possible repetitions of the experiment.

- If we look at the sampling distribution of a Bayesian estimator for a fixed parameter value it is called pre-posterior analysis since it can be done prior to taking the data.
- A Monte Carlo study is where we perform the experiment a large number of times, and calculate the statistic for each experiment. We use the empirical distribution of the statistic over all the samples we took in our study instead of its sampling distribution over all possible repetitions.

2

Scientific Data Gathering

Scientists gather data purposefully, in order to find answers to particular questions. Statistical science has shown that data should be relevant to the particular questions, yet be gathered using randomization. The development of methods to gather data purposefully, yet using randomization is one of the greatest contributions the field of statistics has made to the practice of science.

Variability in data solely due to chance can be averaged out by increasing the sample size. Variability due to other causes cannot be. Statistical methods have been developed for gathering data randomly, yet relevant to a specific question. These methods can be divided into two fields. Sample survey theory is the study of methods for sampling from a finite real population. Experimental design is the study of methods for designing experiments that focus on the desired factors, and are not affected by other possibly unidentified ones.

Inferences always depend on the probability model which we assume generated the observed data being the correct one. When data are not gathered randomly, there is a risk that the observed pattern is due to lurking variables that were not observed, instead of being a true reflection of the underlying pattern. In a properly designed experiment, treatments are assigned to subjects in such a way as to reduce the effects of any lurking variables that are present, but unknown to us.

When we make inferences from data gathered according to a properly designed random survey or experiment, the probability model for the observations follows from the design of the survey or experiment, and we can be confident that it is correct. This puts our inferences on a solid foundation. On the other hand, when we

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

make inferences from data gathered from a nonrandom design, we don't have any underlying justification for the probability model, we just assume it is true! There is the possibility the assumed probability model for the observations is not correct, and our inferences will be on shaky ground.

2.1 SAMPLING FROM A REAL POPULATION

First, we will define some fundamental terms.

- *Population.* The entire group of objects or people the investigator wants information about. For instance, the population might consist of New Zealand residents over the age of eighteen. Usually we want to know some specific attribute about the population. Each member of the population has a number associated with it, for example, his/her annual income. Then we can consider the model population to be the set of numbers for each individual in the real population. Our model population would be the set of incomes of all New Zealand residents over the age of eighteen. We want to learn about the distribution of the population. Specifically, we want information about the population *parameters*, which are numbers associated with the distribution of the population, such as the population mean, median, and standard deviation. Often it is not feasible to get information about all the units in the population. The population may be too big, or spread over too large an area, or it may cost too much to obtain data for the complete population. So we don't know the parameters because it is infeasible to calculate them.

- *Sample.* A subset of the population. The investigator draws one sample from the population, and gets information from the individuals in that sample. Sample *statistics* are calculated from sample data. They are numerical characteristics that summarize the distribution of the sample, such as the sample mean, median, and standard deviation. A statistic has a similar relationship to a sample that a parameter has to a population. However, the sample is known, so the statistic can be calculated.

- *Statistical inference.* Making a statement about population parameters on basis of sample statistics. Good inferences can be made if the sample is representative of the population as a whole! The distribution of the sample must be similar to the distribution of the population from which it came! *Sampling bias*, a systematic tendency to collect a sample which is not representative of the population, must be avoided. It would cause the distribution of the sample to be dissimilar to that of the population, and thus lead to very poor inferences.

Even if we are aware of something about the population and try to represent it in the sample, there is probably some other factors in the population that we are unaware of, and the sample would end up being nonrepresentative in those factors.

Example 1 *Suppose we are interested in estimating the proportion of Hamilton voters who approve the Hamilton City Council's financing a new rugby stadium. We*

decide to go downtown one lunch break, and draw our sample from people passing by. We might decide that our sample should be balanced between males and females the same as the voting age population. We might get a sample evenly balanced between males and females, but not be aware that the people we interview during the day are mainly those on the street during working hours. Office workers would be over represented, while factory workers would be underrepresented. There might be other biases inherent in choosing our sample this way, and we might not have a clue as to what these biases are. Some groups would be systematically underrepresented, and others systematically overrepresented. We can't make our sample representative for classifications we don't know.

Surprisingly, *random samples* give more representative samples than any nonrandom method such as quota samples or judgment samples. They not only minimize the amount of error in the inference, they also allow a (probabilistic) measurement of the error that remains.

Simple Random Sampling (without Replacement)

Simple random sampling requires a *sampling frame* , which is a list of the population numbered from 1 to N. A sequence of n random numbers are drawn from the numbers 1 to N. Each time a number is drawn, it is removed from consideration, so it cannot be drawn again. The items on the list corresponding to the chosen numbers are included in the sample. Thus, at each draw, each item not yet selected has an equal chance of being selected. Every item has equal chance of being in the final sample. Furthermore, every possible sample of the required size is equally likely.

Suppose we are sampling from the population of registered voters in a large city. It is likely that the proportion of males in the sample is close to the proportion of males in the population. Most samples are near the correct proportions, however, we are not certain to get the exact proportion. All possible samples of size n are equally likely, including those that are not representative with respect to sex.

Stratified Random Sampling

Since we know what the proportions of males and females are from the voters list, we should take that information into account in our sampling method. In stratified random sampling, the population is divided into subpopulations called *strata*. In our case this would be males and females. The sampling frame would be divided into separate sampling frames for the two strata. A simple random sample is taken from each *stratum* where each stratum sample size is proportional to stratum size. Every item has equal chance of being selected. And every possible sample that has each stratum represented in the correct proportions is equally likely. This method will give us samples that are exactly representative with respect to sex. Hence inferences from these type samples will be more accurate than those from simple random sampling when the variable of interest has different distributions over the strata. If the variable of interest is the same for all the strata, stratified random sampling will be no more

(and no less) accurate than simple random sampling. Stratification has no potential downside as far as accuracy of the inference. However, it is more costly, as the sampling frame has to be divided into separate sampling frames for each stratum.

Cluster Random Sampling

Sometimes we don't have a good sampling frame of individuals. In other cases the individuals are scattered across a wide area. In cluster random sampling, we divide that area into neighborhoods called clusters. Then we make a sampling frame for clusters. A random sample of clusters is selected. All items in the chosen clusters are included in the sample. This is very cost effective because the interviewer won't have as much travel time between interviews. The drawback is that items in a cluster tend to be more similar than items in different clusters. For instance, people living in the same neighborhood usually come from the same economic level because the houses were built at the same time and in the same price range. This means that each observation gives less information about the population parameters. It is less efficient in terms of sample size. However, often it is very cost effective, since getting a larger sample is usually cheaper by this method.

Nonsampling Errors in Sample Surveys

Errors can arise in sample surveys or in a complete population census for reasons other than the sampling method used. These nonsampling errors include response bias; the people who respond may be somewhat different than those who do not respond. They may have different views on the matters surveyed. Since we only get observations from those who respond, this difference would bias the results. A well planned survey will have callbacks, where those in the sample who haven't responded will be contacted again, in order to get responses from as many people in the original sample as possible. This will entail additional costs, but is important as we have no reason to believe that nonrespondents have the same views as the respondents. Errors can also arise from poorly worded questions. Survey questions should be trialed in a pilot study to determine if there is any ambiguity.

Randomized Response Methods

Social science researchers and medical researchers often wish to obtain information about the population as a whole, but the information that they wish to obtain is sensitive to the individuals who are surveyed. For instance, the distribution of the number of sex partners over the whole population would be indicative of the overall population risk for sexually transmitted diseases. Individuals surveyed may not wish to divulge this sensitive personal information. They might refuse to respond, or even worse, they could give an untruthful answer. Either way, this would threaten the validity of the survey results. *Randomized response* methods have been developed to get around this problem. There are two questions, the sensitive question and the dummy question. Both questions have the same set of answers. The respondent uses

a randomization that selects which question he or she answers, and also the answer if the dummy question is selected. Some of the answers in the survey data will be to the sensitive question and some will be to the dummy question. The interviewer will not know which is which. However, the incorrect answers are entering the data from known randomization probabilities. This way information about the population can be obtained without actually knowing the personal information of the individuals surveyed, since only that individual knows which question he or she answered. Bolstad, Hunt, and McWhirter (2001) describe a *Sex, Drugs, and Rock & Roll Survey* that gets sensitive information about a population (Introduction to Statistics class) using randomized response methods.

2.2 OBSERVATIONAL STUDIES AND DESIGNED EXPERIMENTS

The goal of scientific inquiry is to gain new knowledge about the cause and effect relationship between a factor and a response variable. We gather data to help us determine these relationships, and to develop mathematical models to explain them. The world is complicated. There are many other factors that may affect the response. We may not even know what these other factors are. If we don't know what they are, we cannot control them directly. Unless we can control them, we can't make inferences about cause and effect relationships! Suppose, for example, we want to study a herbal medicine for its effect on weight loss. Each person in the study is an *experimental unit*. There is great variability between experimental units, because people are all unique individuals with their own hereditary body chemistry and dietary and exercise habits. The variation among experimental units makes it more difficult to detect the effect of a treatment. Figure 2.1 shows a collection of experimental units. The degree of shading shows they are not the same with respect to some unidentified variable. The response variable in the experiment may depend on that unidentified variable, which could be a lurking variable in the experiment.

Observational Study

If we record the data on a group of subjects that decided to take the herbal medicine and compared that with data from a control group who did not, that would be an *observational study*. The treatments have *not* been randomly assigned to treatment and control group. Instead they self select. Even if we observe a substantial difference between the two groups, we cannot conclude there is a causal relationship from an observational study. We can't rule out that the association was due to an unidentified lurking variable. In our study, those who took the treatment may have been more highly motivated to lose weight than those who did not. Or there may be other factors that differed between the two groups. Any inferences we make on an observational study are dependent on the assumption that there are no differences between the distribution of the units assigned to the treatment groups and the control group. We can't know whether this assumption is actually correct in an observational study.

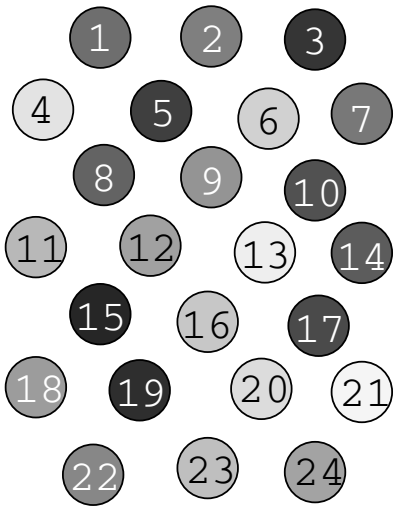


Figure 2.1 Variation among experimental units.

Designed Experiment

We need to get our data from a designed experiment if we want to be able to make sound inferences about cause-effect relationships. The experimenter uses randomization to decide which subjects get into the treatment group(s) and control group respectively. For instance, he/she uses a table of random numbers, or flips a coin.

We are going to divide the experimental units into four treatment groups (one of which may be a control group). We must ensure that each group gets a similar range of units. If we don't, we might end up attributing a difference between treatment groups to the different treatments, when in fact it was due to the lurking variable and a biased assignment of experimental units to treatment groups.

Completely randomized design. We will randomly assign experimental units to groups so that each experimental unit is equally likely to go to any of the groups. Each experimental unit will be assigned (nearly)independently of other experimental units. The only dependence between assignments is that having assigned one unit to treatment group 1 (for example), the probability of the other unit being assigned to group 1 is slightly reduced because there is one less place in group 1. This is known as a completely randomized design. Having a large number of (nearly) independent randomizations ensures that the comparisons between treatment groups and control group are fair since all groups will contain a similar range of experimental units. Units having high values and units having low values of the lurking variable will be

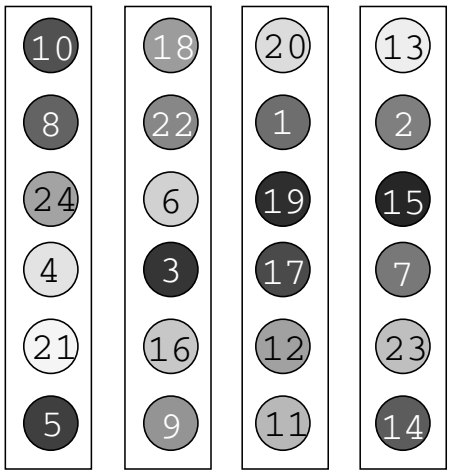


Figure 2.2 Completely randomized design. Units have been randomly assigned to four treatment groups.

in all treatment groups in similar proportions. In Figure 2.2 we see the four treatment groups have similar range of experimental units with respect to the unidentified lurking variable.

The randomization averages out the differences between experimental units assigned to the groups. The expected value of the lurking variable is the same for all groups, because of the randomization. The average value of the lurking variable for each group will be close to its mean value in the population because there are a large number of independent randomizations. The larger the number of units in the experiment, the closer the average values of the lurking variable in each group will be to its mean value in the population. If we find an association between the treatment and the response, it will be unlikely that the association was due to any lurking variable. For a large-scale experiment, we can effectively rule out any lurking variable, and conclude that the association was due to the effect of different treatments.

Randomized block design. If we identify a variable, we can control for it directly. It ceases to be a lurking variable. One might think that using judgment about assigning experimental units to the treatment and control groups would lead to similar range of units being assigned to them. The experimenter could get similar groups according to the criterion (identified variable) he/she was using. However, there would be no protection against any other lurking variable that hadn't been considered. We can't expect it to be averaged out if we haven't done the assignments randomly!

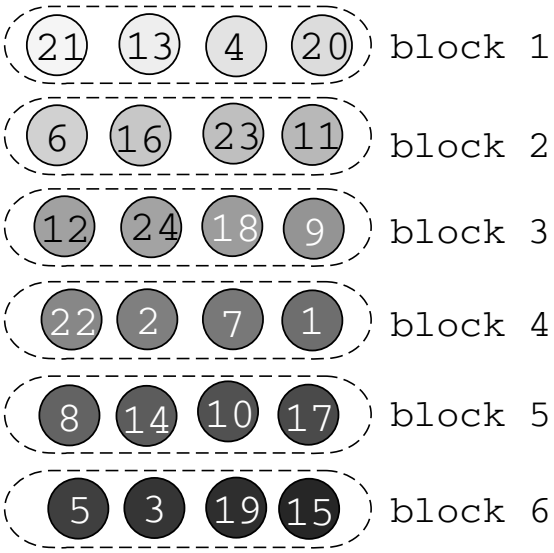


Figure 2.3 Similar units have been put into blocks.

Any prior knowledge we have about the experimental units should be used before the randomization. Units that have similar values of the identified variable should be formed into *blocks*. This is shown in Figure 2.3. The experimental units in each block are similar with respect to that variable. Then the randomization is be done within blocks. One experimental unit in each block is randomly assigned to each treatment group. The blocking controls that particular variable, as we are sure all units in the block are similar, and one goes to each treatment group. By selecting which one goes to each group randomly, we are protecting against any other lurking variable by randomization. It is unlikely that any of the treatment groups was unduly favored or disadvantaged by the lurking variable. On the average, all groups are treated the same. Figure 2.4 shows the treatment groups found by a randomized block design. We see the four treatment groups are even more similar than those from the randomized block design.

For example, if we wanted to determine which of four varieties of wheat gave better yield, we would divide the field into blocks of four adjacent plots because plots that are adjacent are more similar in their fertility than plots that are distant from each other. Then within each block, one plot would be randomly assigned to each variety. This randomized block design ensures that the four varieties each have been assigned to similar groups of plots. It protects against any other lurking variable, by the within block randomization.

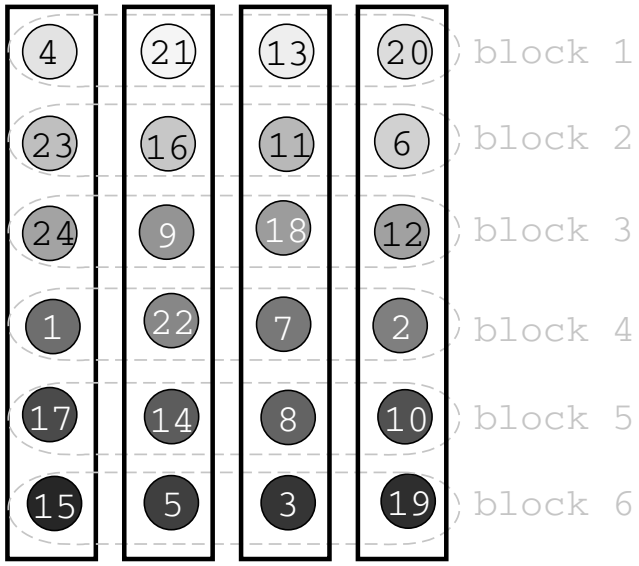


Figure 2.4 Randomized block design. One unit in each block randomly assigned to each treatment group. Randomizations in different blocks are independent of each other.

When the response variable is related to the trait we are blocking on, the blocking will be effective, and the randomized block design will lead to more precise inferences about the yields than a completely randomized design with the same number of plots. This can be seen by comparing the treatment groups from the completely randomized design shown in Figure 2.2 with the treatment groups from the randomized block design shown in Figure 2.4. The treatment groups from the randomized block design are more similar than those from the completely randomized design.

Main Points

- *Population*. The entire set of objects or people that the study is about. Each member of the population has a number associated with it, so we often consider the population as a set of numbers. We want to know about the distribution of these numbers.
- *Sample*. The subset of the population from which we obtain the numbers.
- *Parameter*. A number that is a characteristic of the population distribution, such as the mean, median, standard deviation, and interquartile range of the whole population.
- *Statistic*. A number that is a characteristic of the sample distribution, such as the mean, median, standard deviation, and interquartile range of the sample.

- *Statistical inference*. Making a statement about population parameters on the basis of sample statistics.
- *Simple random sampling*. At each draw every item that has not already been drawn has an equal chance of being chosen to be included in the sample.
- *Stratified random sampling*. The population is partitioned into subpopulations called strata, and simple random samples are drawn from each stratum where the stratum sample sizes are proportional to the stratum proportions in the population. The stratum samples are combined to form the sample from the population.
- *Cluster random sampling*. The area the population lies in is partitioned into areas called clusters. A random sample of clusters is drawn, and all members of the population in the chosen clusters are included in the sample.
- *Randomized response methods*. These allow the respondent to randomly determine whether to answer a sensitive question or the dummy question, which both have the same range of answers. Thus the respondents personal information is not divulged by the answer, since the interviewer does not know which question it applies to.
- *Observational study*. The researcher collects data from a set of experimental units not chosen randomly, or not allocated to experimental or control group by randomization. There may be lurking variables due to the lack of randomization.
- *Designed experiment*. The researcher allocates experimental units to the treatment group(s) and control group by some form of randomization.
- *Completely randomized design*. The researcher randomly assigns the units into the treatment groups (nearly) independently. The only dependence is the constraint that the treatment groups are the correct size.
- *Randomized block design*. The researcher first groups the units into blocks which contain similar units. Then the units in each block are randomly assigned, one to each group. The randomizations in separate blocks are performed independent of each other.

Monte Carlo Exercises

2.1 **Monte Carlo study comparing methods for random sampling.** We will use a Monte Carlo computer simulation to evaluate the methods of random sampling. Now, if we want to evaluate a method, we need to know how it does in the long run. In a real life situation, we can't judge a method by the sample estimate it gives, because if we knew the population parameter, we would not be taking a sample and estimating it with a sample statistic.

One way to evaluate a statistical procedure is to evaluate the *sampling distribution* which summarizes how the estimate based on that procedure varies in the long run (over all possible random samples) for a case when we know the population parameters. Then we can see how closely the sampling distribution is centered around the true parameter. The closer it is, the better the statistical procedure, and the more confidence we will have in it for realistic cases when we don't know the parameter.

If we use computer simulations to run a large number of hypothetical repetitions of the procedure with known parameters, this is known as a Monte Carlo study named after the famous casino. Instead of having the theoretical sampling distribution, we have the empirical distribution of the sample statistic over those simulated repetitions. We judge the statistical procedure by seeing how closely the empirical distribution of the estimator is centered around the known parameter.

The population. Suppose there is a population made up of 100 individuals, and we want to estimate the mean income of the population from a random sample of size 20. The individuals come from three ethnic groups with population proportions of 40%, 40%, and 20% respectively. There are twenty neighborhoods and five individuals live in each one. Now, the income distribution may be different for the three ethnic groups. Also, individuals in the same neighborhood tend to be more similar than individuals in different neighborhoods.

Details about the population are contained in the Minitab worksheet *sscsample.mtw*. Each row contains the information for an individual. Column 1 contains the income, column 2 contains the ethnic group, and column 3 contains the neighborhood. Compute the mean income for the population. That will be the true parameter value that we are trying to estimate.

In the Monte Carlo study we will approximate the *sampling distribution* of the sample means for three types of random sampling, simple random sampling, stratified random sampling, and cluster random sampling. We do this by drawing a large number (in this case 200) random samples from the population using each method of sampling, calculating the sample mean as our estimate. The empirical distribution of these 200 sample means approximates the sampling distribution of the estimate.

(a) Display the incomes for the three ethnic groups (strata) using boxplots on the same scale. Compute the mean income for the three ethnic groups. Do you see any difference between the income distributions?

(b) Draw 200 random samples of size 20 from the population using simple random sampling using *sscsample.mac* and put the output in columns c6-c9. Details of how to use this macro are in Appendix 3. Answer the following questions from the output:

   i. Does simple random sampling always have the strata represented in the correct proportions?

ii. On the average, does simple random sampling give the strata in their correct proportions?

iii. Does the mean of the *sampling distribution* of the sample mean for simple random sampling appear to be close enough to the population mean that we can consider the difference to be due to chance alone? (We only took 200 samples, not all possible samples.)

(c) Draw 200 stratified random samples using the macro and store the output in c11-c14. Answer the following questions from the output:

i. Does stratified random sampling always have the strata represented in the correct proportions?

ii. On the average, does stratified random sampling give the strata in their correct proportions?

iii. Does the mean of the *sampling distribution* of the sample mean for stratified random sampling appear to be close enough to the population mean that we can consider the difference to be due to chance alone? (We only took 200 samples, not all possible samples.)

(d) Draw 200 cluster random samples using the macro and put the output in columns c16-c19. Answer the following questions from the output:

i. Does cluster random sampling always have the strata represented in the correct proportions?

ii. On the average, does cluster random sampling give the strata in their correct proportions?

iii. Does the mean of the *sampling distribution* of the sample mean for cluster random sampling appear to be close enough to the population mean that we can consider the difference to be due to chance alone? (We only took 200 samples, not all possible samples.)

(e) Compare the spreads of the sampling distributions (standard deviation and interquartile range). Which method of random sampling seems to be more effective in giving sample means more concentrated about the true mean?

(f) Give reasons for this.

2.2 **Monte Carlo study comparing completely randomized design and randomized block design.** Often we want to set up an experiment to determine the magnitude of several treatment effects. We have a set of experimental units that we are going to divide into treatment groups. There is variation among the experimental units in the underlying response variable that we are going to measure. We will assume that we have an additive model where each of the treatments has a constant effect. That means the measurement we get for an experimental unit i given treatment j will be the underlying value for unit i

plus the effect of the treatment for the treatment it receives

$$y_{i,j} = u_i + T_j \, ,$$

where u_i is the underlying value for experimental unit i and T_j is the treatment effect for treatment j. The assignment of experimental units to treatment groups is crucial.

There are two things that the assignment of experimental units into treatment groups should deal with. First, there may be a "lurking variable" that is related to the measurement variable, either positively or negatively. If we assign experimental units that have high values of that lurking variable into one treatment group, that group will be either advantaged or disadvantaged, depending if there is a positive or negative relationship. We would be quite likely to conclude that treatment is good or bad relative to the other treatments, when in fact the apparent difference would be due to the effect of the lurking variable. That is clearly a bad thing to occur. We know that to prevent this, the experimental units should be assigned to treatment groups according to some randomization method. On the average, we want all treatment groups to get a similar range of experimental units with respect to the lurking variable. Otherwise, the experimental results may be biased.

Second, the variation in the underlying values of the experimental units may mask the differing effects of the treatments. It certainly makes it harder to detect a small difference in treatment effects. The assignment of experimental units into treatment groups should make the groups as similar as possible. Certainly, we want the group means of the underlying values to be nearly equal.

The *completely randomized design* randomly divides the set of experimental units into treatment groups. Each unit is randomized (almost) independently. We want to insure that each treatment group contains equal numbers of units. Every assignment that satisfies this criterion is equally likely. This design does not take the values of the other variable into account. It remains a possible lurking variable.

The *randomized block design* takes the other variable value into account. First blocks of experimental units having similar values of the other variable are formed. Then one unit in each block is randomly assigned to each of the treatment groups. In other words, randomization occurs within blocks. The randomizations in different blocks are done independently of each other. This design makes use of the other variable. It ceases to be a lurking variable and becomes the blocking variable.

In this assignment we compare the two methods of randomly assigning experimental units into treatment groups. Each experimental unit has an underlying value of the response variable and a value of another variable associated with it. (If we don't take the other variable in account, it will be a lurking variable.)

We will run a small-scale Monte Carlo study to compare the performance of these two designs in two situations.

(a) First we will do a small-scale Monte Carlo study of 500 random assignments using each of the two designs when the response variable is strongly related to the other variable. We let the correlation between them be $k1 = .8$. The details of how to use the Minitab macro *Xdesign.mac* or the R function *Xdesign* are in Appendix 3 and Appendix 4, respectively. Look at the boxplots and summary statistics.

   i. Does it appear that, on average, all groups have the same underlying mean value for the other (lurking) variable when we use a completely randomized design?

   ii. Does it appear that, on average, all groups have the same underlying mean value for the other (blocking) variable when we use a randomized block design?

   iii. Does the distribution of the other variable over the treatment groups appear to be the same for the two designs? Explain any difference.

   iv. Which design is controlling for the other variable more effectively? Explain.

   v. Does it appear that, on average, all groups have the same underlying mean value for the response variable when we use a completely randomized design?

   vi. Does it appear that, on average, all groups have the same underlying mean value for the response variable when we use a randomized block design?

   vii. Does the distribution of the response variable over the treatment groups appear to be the same for the two designs? Explain any difference.

   viii. Which design will give us a better chance for detecting a small difference in treatment effects? Explain.

   ix. Is blocking on the other variable effective when the response variable is strongly related to the other variable?

(b) Next we will do a small-scale Monte Carlo study of 500 random assignments using each of the two designs when the response variable is weakly related to the other variable. We let the correlation between them be $k1 = .4$. Look at the boxplots and summary statistics.

   i. Does it appear that, on average, all groups have the same underlying mean value for the other (lurking) variable when we use a completely randomized design?

   ii. Does it appear that, on average, all groups have the same underlying mean value for the other (blocking) variable when we use a randomized block design?

iii. Does the distribution of the other variable over the treatment groups appear to be the same for the two designs? Explain any difference.

iv. Which design is controlling for the other variable more effectively? Explain.

v. Does it appear that, on average, all groups have the same underlying mean value for the response variable when we use a completely randomized design?

vi. Does it appear that, on average, all groups have the same underlying mean value for the response variable when we use a randomized block design?

vii. Does the distribution of the response variable over the treatment groups appear to be the same for the two designs? Explain any difference.

viii. Which design will give us a better chance for detecting a small difference in treatment effects? Explain.

ix. Is blocking on the other variable effective when the response variable is strongly related to the other variable?

(c) Next we will do a small-scale Monte Carlo study of 500 random assignments using each of the two designs when the response variable is not related to the other variable. We let the correlation between them be $k1 = 0$. This will make the response variable independent of the other variable. Look at the boxplots for the treatment group means for the other variable.

i. Does it appear that, on average, all groups have the same underlying mean value for the other (lurking) variable when we use a completely randomized design?

ii. Does it appear that, on average, all groups have the same underlying mean value for the other (blocking) variable when we use a randomized block design?

iii. Does the distribution of the other variable over the treatment groups appear to be the same for the two designs? Explain any difference.

iv. Which design is controlling for the other variable more effectively? Explain.

v. Does it appear that, on average, all groups have the same underlying mean value for the response variable when we use a completely randomized design?

vi. Does it appear that, on average, all groups have the same underlying mean value for the response variable when we use a randomized block design?

vii. Does the distribution of the response variable over the treatment groups appear to be the same for the two designs? Explain any difference.

viii. Which design will give us a better chance for detecting a small difference in treatment effects? Explain.

ix. Is blocking on the other variable effective when the response variable is independent from the other variable?

x. Can we lose any effectiveness by blocking on a variable that is not related to the response?

3

Displaying and Summarizing Data

We use statistical methods to extract information from data and gain insight into the underlying process that generated the data. Frequently our data set consists of measurements on one or more variables over the experimental units in one or more samples. The distribution of the numbers in the sample will give us insight into the distribution of the numbers for the whole population.

It is very difficult to gain much understanding by looking at a set of numbers. Our brains were not designed for that. We need to find ways to present the data that allow us to note the important features of the data. The visual processing system in our brain enables us to quickly perceive the overview we want, when the data are represented pictorially in a sensible way. They say a picture is worth a thousand words. That is true, provided the we have the correct picture. If the picture is incorrect, we can mislead ourselves and others very badly!

3.1 GRAPHICALLY DISPLAYING A SINGLE VARIABLE

Often our data set consists of a set of measurements on a single variable for a single sample of subjects or experimental units. We want to get some insight into the distribution of the measurements of the whole population. A visual display of the measurements of the sample helps with this.

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

Table 3.1 Earth density measurements by Cavendish

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

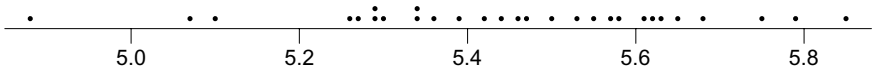


Figure 3.1 Dotplot of Earth density measurements by Cavendish.

Example 2 *In 1798 the English scientist Cavendish performed a series of 29 measurements on the density of the Earth using a torsion balance. This experiment and the data set are described by Stigler (1977). Table 3.1 contains the 29 measurements.*

Dotplot

A dotplot is the simplest data display for a single variable. Each observation is represented by a dot at its value along horizontal axis. This shows the relative positions of all the observation values. It is easy to get a general idea of the distribution of the values. Figure 3.1 shows the dotplot of Cavendish's Earth density measurements.

Boxplot (Box-and-Whisker Plot)

Another simple graphical method to summarize the distribution of the data is to form a boxplot. First we have to sort and summarize the data.

Originally, the sample values are $y_1, \cdots, y_n$. The subscript denotes the order (in time) the observation was taken, y_1 is the first, y_2 is the second, and so on up to y_n which is last. When we order the sample values by size from smallest to largest we get the *order statistics*. They are denoted $y_{[1]}, \cdots, y_{[n]}$, where $y_{[1]}$ is the smallest, $y_{[2]}$ is the second smallest, on up to the largest $y_{[n]}$. We divide the ordered observations into quarters with the quartiles. Q_1, the lower quartile, is the value that 25% of the observations are less than or equal to it, and 75% or more of the observations are greater than or equal to it. Q_2, the middle quartile, is the value that 50% or more of the observations are less than or equal to it, and 50% or more of the observations are greater than or equal to it. Q_2 is also known as the sample median. Similarly Q_3, the upper quartile is the value that 75% of the observations are less than or equal to it, and 25% of the observations are greater than or equal to it. We can find these from

the order statistics:

$$Q_1 = y_{[\frac{n+1}{4}]} ,$$

$$Q_2 = y_{[\frac{n+1}{2}]} ,$$

$$Q_3 = y_{[\frac{3(n+1)}{4}]} .$$

If the subscripts are not integers, we take the weighted average of the two closest order statistics. For example, Cavendish's Earth density data $n = 29$,

$$Q_1 = y_{[\frac{30}{4}]} .$$

This is halfway between the 7'th and 8'th order statistics, so

$$Q_1 = \tfrac{1}{2} \times y_{[7]} + \tfrac{1}{2} \times y_{[8]} .$$

The five number summary of a data set is $y_{[1]}, Q_1, Q_2, Q_3, y_{[n]}$. This gives the minimum, the three quartiles, and the maximum of the observations. The *boxplot* or *box-and-whisker plot* is a pictorial way of representing the five number summary. The steps are:

- Draw and label an axis.
- Draw a box with ends at the first and third quartiles.
- Draw a line through the box at the second quartile (median).
- Draw a line (whisker) from the lower quartile to the lowest observation, and draw a line (whisker) from the upper quartile to the highest observation.
- **Warning**: Minitab extends the whiskers only to a maximum length of 1.5 $\times$ the interquartile range. Any observation further out than that is identified with an asterisk (*) to indicate the observation may be an outlier. This can seriously distort the picture of the sample, because the criterion does not depend on the sample size. A large sample can look very heavy-tailed because the asterisks show that there are many possibly outlying values, when the proportion of outliers is well within the normal range. In Exercise 6, we show how this distortion works, and how we can control it by editing the attribute in the Minitab dialog box.

The boxplot divides the observations into quarters. It shows you a lot about the shape of the data distribution. Examining the length of the whiskers compared to the box length shows whether the data set has light, normal, or heavy tails. Comparing the lengths of the whiskers show whether the distribution of the data appears to be skewed or symmetric. Figure 3.2 shows the boxplot for Cavendish's Earth density measurements. It shows the data distribution is fairly symmetric but with a slightly longer lower tail.

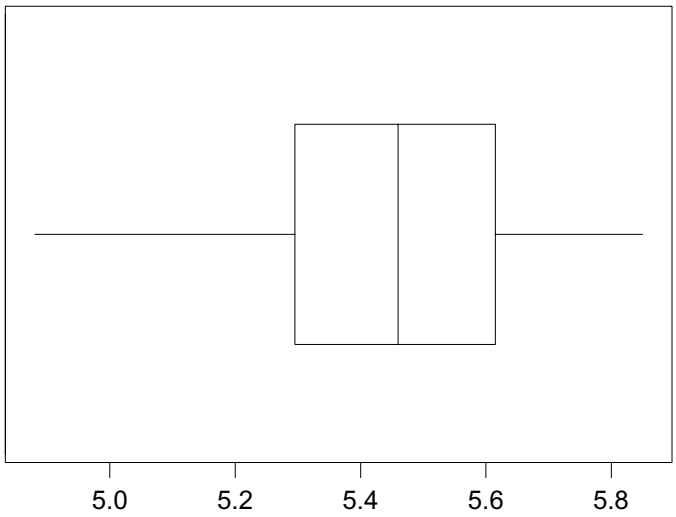


Figure 3.2 Boxplot of Earth density measurements by Cavendish.

Stem-and-Leaf Diagram

The stem-and-leaf diagram is a quick and easy way of extracting information about the distribution of a sample of numbers. The *stem* represents the leading digit(s) to a certain depth (power of 10) of each data item, and the leaf represents the next digit of the data item. A stem-and-leaf diagram can be constructed by hand for a small data set. It is often the first technique used on a set of numbers. The steps are

- Draw a vertical axis (stem) and scale it for the stem units. Always use a *linear* scale!
- Plot leaf for the next digit. We could round off the leaf digit, but usually we don't bother if we are doing it by hand. In any case, we may have lost some information by rounding off or by truncating.
- Order the leaves with the smallest near stem to the largest farthest away.
- State the leaf unit on your diagram.

The stem-and-leaf plot gives a picture of the distribution of the numbers when we turn it on its side. It retains the actual numbers to within the accuracy of the leaf unit. We can find the order statistics counting up from the lower end. This helps to find the quartiles and the median. The Figure 3.3 shows a stem-and-leaf diagram for Cavendish's Earth density measurements. We use a two digit stem, units and tenths, and a one digit leaf, hundredths.

	leaf unit .01
48	8
49	
50	7
51	0
52	6799
53	04469
54	2467
55	03578
56	12358
57	59
58	5

Figure 3.3 Stem-and-leaf plot for Cavendish's Earth density measurements.

There are 29 measurements. We can count down to the $X_{\frac{29+1}{2}} = X_{15}$ to find that the median is 5.46. We can count down to $X_{\frac{29+1}{4}} = X_{7\frac{1}{2}}$. Thus the first quartile $Q_1 = \frac{1}{2} \times X_7 + \frac{1}{2} \times X_8$ which is 5.295

Frequency Table

Another main approach to simplify a set of numbers is to put it in a frequency table. This is sometimes referred to as *binning* the data. The steps are:

- Partition possible values into nonoverlapping groups (bins). Usually we use equal width groups. However this is not required.
- Put each item into the group it belongs in.
- Count the number of items in each group.

Frequency tables are a useful tool for summarizing data into an understandable form. There is a trade-off between the loss of information in our summary, and the ease of understanding the information that remains. We have lost information when we put a number into a group. We know it lies between the group boundaries, but its exact value is no longer known. The fewer groups we use, the more concise the summary, but the greater loss of information. If we use more groups we lose less information, but our summary is less concise and harder to grasp. Since we no longer have the information about exactly where each value lies in a group, it seems logical that the best assumption we can then make is that each value in the group is equally possible. The Earth density measurements made by Cavendish are shown as a frequency table in Table 3.2.

If there are two many groups, some of them may not contain any observations. In that case, it is better to lump two or more adjacent groups into a bigger one to

Table 3.2 Frequency table of Earth density measurements by Cavendish

Boundaries	Frequency
$4.80 < x \leq 5.00$	1
$5.00 < x \leq 5.20$	2
$5.20 < x \leq 5.40$	9
$5.40 < x \leq 5.60$	9
$5.60 < x \leq 5.80$	7
$5.80 < x \leq 6.00$	1

get some observations in every group. There are two ways to show the data in a frequency table pictorially. They are *histograms* and *cumulative frequency polygons*.

Histogram

This is the most common way to show the distribution of data in the frequency table. The steps for constructing a histogram are:

- Put group boundaries on horizontal axis drawn on a *linear* scale.
- Draw a rectangular bar for each group where the *area* of bar is proportional to the frequency of that group. For example, this means that if a group is twice as wide as the others, its height is half that group's frequency. The bar is flat across the top to show our assumption that each value in the group is equally possible.
- Do not put any gaps between the bars if the data are continuous.
- The scale on the vertical axis is density, which is group frequency divided by group width. When the groups have equal width, the scale is proportional to frequency, or relative frequency, and they could be used instead of density. This is not true if unequal width groups are used. It is not necessary to label the vertical axis on the graph. The shape of the graph is the important thing, not its vertical scale.
- **Warning**: If you use unequal group widths in Minitab, you must click on *density* in the *options* dialog box; otherwise, the wrong shape histogram will result.

The histogram gives us a picture of how the sample data are distributed. We can see the shape of the distribution and relative tail weights. We look at it as a representing a picture of the underlying population the sample came from. This underlying

population distribution[1] would generally be reasonably smooth. There is always a trade-off between too many and too few groups. If we use too many groups, the histogram has a "saw tooth" appearance and the histogram is not representing the population distribution very well. If we use too few groups, we lose details about the shape. Figure 3.4 shows histogram of the Earth density measurements by Cavendish using 12, 6, and 4 groups, respectively. This illustrates the trade-off between too many and too few groups. We see the histogram with 12 groups has gaps, and a saw tooth appearance. The histogram with 6 groups gives a better representation of the underlying distribution of Earth density measurements. The histogram with 4 groups has lost too much detail. The last histogram has unequal width groups. The height of the wider bars is shortened to keep the area proportional to frequency.

Cumulative Frequency Polygon

The other way for displaying the data from a frequency table is to construct a *cumulative frequency polygon*, sometimes called an *ogive*. It is particularly useful because you can estimate the median and quartiles from the graph. The steps are:

- Group boundaries on horizontal axis drawn on a *linear* scale.
- Frequency or percentage shown on vertical axis.
- Plot *(lower boundary of lowest class, 0)*.
- For each group, plot *(upper class boundary, cumulative frequency)*. We don't know the exact value of each observation in the group. However, we do know that all the values in a group must be less than or equal to the upper boundary.
- Join the plotted points with a straight line. Joining them with a straight line shows that we consider each value in the group to be equally possible.

We can estimate the median and quartiles easily from the graph. To find median go up to 50 % on vertical scale, draw line over to the cumulative frequency polygon, and down to horizontal axis. The value where it hits the axis is the estimate of the median. Similarly to find the quartiles, go up to 25% or 75%, across to cumulative frequency polygon, and down to horizontal axis to find lower and upper quartile respectively. The underlying assumption behind these estimates is that all values in a group are evenly spread across the group. Figure 3.5 shows the cumulative frequency polygon for the Earth density measurements by Cavendish.

[1] In this case, the *population* is the set of all possible Earth density measurements that Cavendish could have obtained from his experiment. This population is theoretical, as each of its elements was only brought into existence by Cavendish performing the experiment.

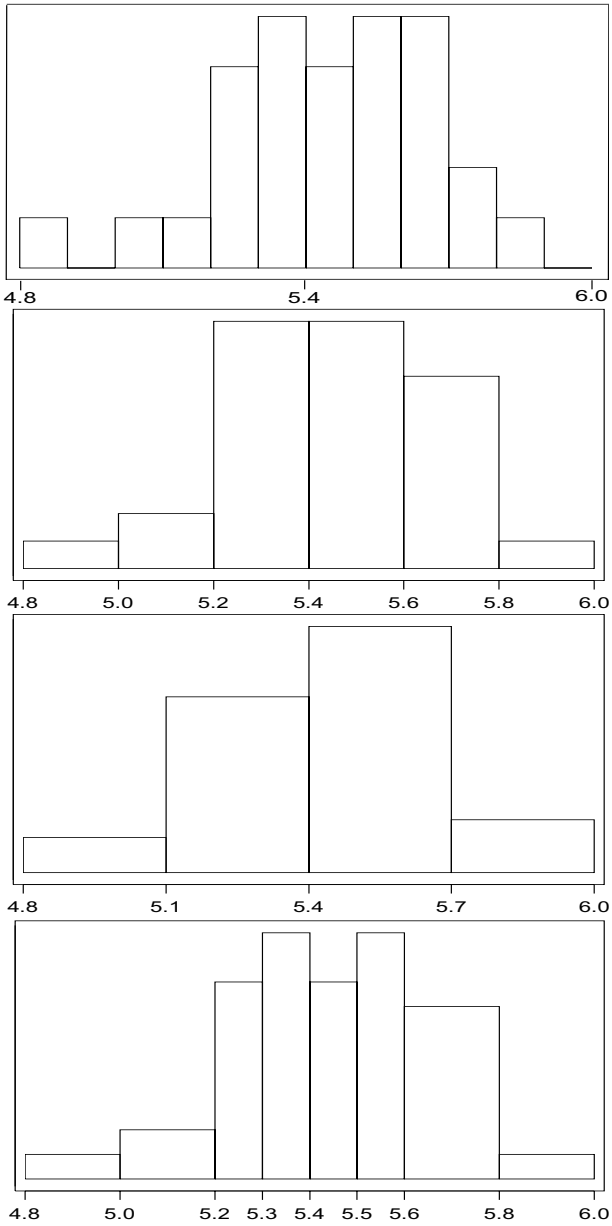


Figure 3.4 Histograms of Earth density measurements by Cavendish with different boundaries. Note the area is always proportional to frequency.

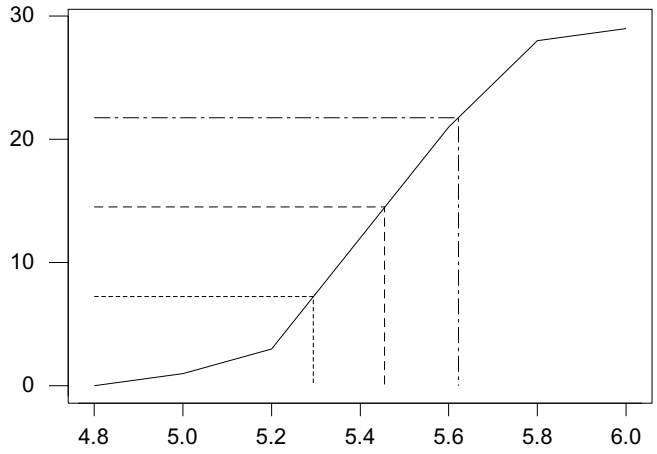


Figure 3.5 Cumulative frequency polygon of Earth density measurements by Cavendish.

3.2 GRAPHICALLY COMPARING TWO SAMPLES

Sometimes we have the same variable recorded for two samples. For instance, we may have responses for the treatment group and control group from a randomized experiment. We want to determine whether or not the treatment has been effective.

Often a picture can clearly show us this, and there is no need for any sophisticated statistical inference. The key to making visual comparisons between two data samples is "Don't compare apples to oranges." By that, we mean that the pictures for the two samples must be lined up, and with the same scale. Stacked dotplots and stacked boxplots where they are lined up on the same axis give a good comparison of the samples. Back-to-back stem-and-leaf diagrams are another good way of comparing two small data sets. The two samples use common stem, and the leaves from one sample are on one side of the stem, and the leaves from the other sample are on the other side of the stem. The leaves of the two sample are ordered, from smallest closest to stem to largest farthest away. We can put histograms back-to-back or stack them. We can plot the cumulative frequency polygons for the two samples on the same axis. If one is always to the left of the other, we can deduce its distribution is shifted relative to the other.

All of these pictures can show us whether there are any differences between the two distributions. For example, do the distributions seem to have the same location on the number line, or does one appear to be shifted relative to the other? Do the distributions seem to have the same spread, or is one more spread out than the other? Are the shapes similar? If we have more than two samples, we can do any of these pictures that is stacked. Of course, back-to-back ones only work for two samples.

Example 3 *Between 1879 and 1882 scientists were devising experiments for determining the speed of light. Table 3.3 contains measurements collected by Michelson in a series of experiments on the speed of light. The first 20 measurements were*

Table 3.3 Michelson's speed of light measurements. Value in table plus 2999000km/s.

Michelson (1879)		Michelson (1882)	
850	740	883	816
900	1070	778	796
930	850	682	711
950	980	611	599
980	880	1051	781
1000	980	578	796
930	650	774	820
760	810	772	696
1000	1000	573	748
960	960	748	797
		851	809
		723	

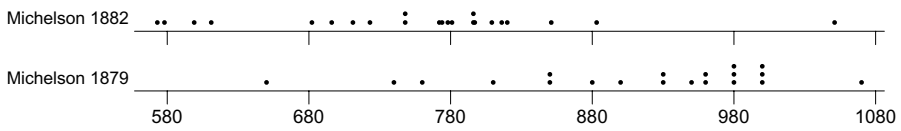


Figure 3.6 Dotplots of Michelsons speed of light measurements.

made in 1879, and the next 23 supplementary measurements were made in 1882. The experiment and the data are described in Stigler (1977).

Figure 3.6 shows stacked dotplots for the two data sets. Figure 3.7 shows stacked boxplots for the two data sets. The true value of the speed of light in the air is 2999710. We see from these plots that there was a systematic error (bias) in the first series of measurements that was greatly reduced in the second.

Back-to-back stem-and-leaf diagrams are another good way to show the relationship between two data sets. The stem goes in the middle. We put the leaves for one data set on the right side, and leaves for the other on the left. The leaves are ascending order moving away from the stem. Back-to-back stem-and-leaf diagrams are shown for Michelson's data in Figure 3.8. The stem is hundreds, and the leaf unit is 10.

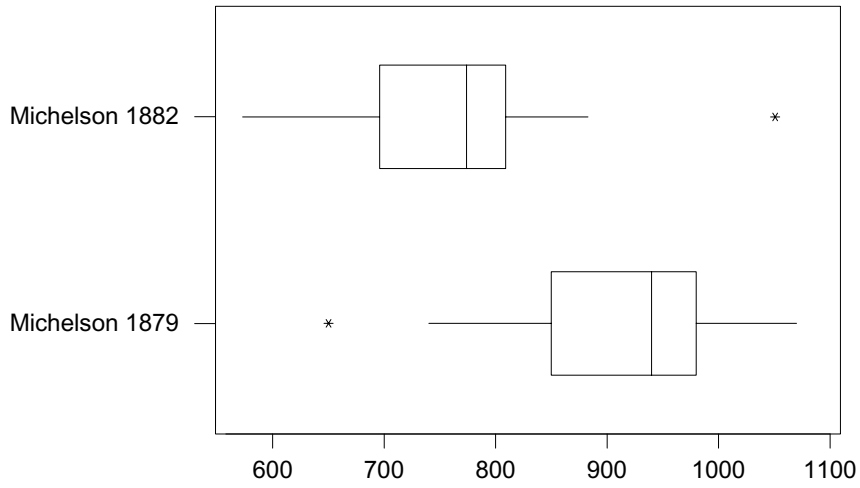


Figure 3.7 Boxplot of Michelsons speed of light measurements.

977	5	leaf unit 10
1	6	
98	6	5
4412	7	4
9998777	7	6
210	8	1
85	8	558
	9	033
	9	566888
	10	000
5	10	7

Figure 3.8 Back-to-back stem-and-leaf plots for Michelson's data.

3.3 MEASURES OF LOCATION

Sometimes we want to summarize our data set with numbers. The most important aspect of the data set distribution is determining a value that summarizes its location on the number line. The most commonly used measures of location are the mean and the median. We will look at each ones's advantages and disadvantages.

Both the mean and the median are members of the trimmed mean family which also includes compromise values between them, depending on the amount of trimming. We do not consider the mode (most common value) to be a suitable measure of location for the following reasons. For continuous data values, each value is unique

if we measure it accurately enough. In many cases, the mode is near one end of the distribution, not the central region. The mode may not be unique.

Mean: Advantages and Disadvantages

The mean is the most commonly used measure of location, because of its simplicity, and its good mathematical properties. The mean of a data set $y_1, \cdots, y_n$ is simply the arithmetic average of the numbers.

$$\bar{y} \;=\; \frac{1}{n} \times \sum_{i=1}^{n} y_i \;=\; \frac{1}{n} \times (y_1 + \cdots + y_n)\,.$$

The mean is simple and very easy to calculate. You just make one pass through the numbers and add them up. Then divide the sum by the size of the sample.

The mean has good mathematical properties. The mean of a sum is the sum of the means. For example, if y is total income, u is "earned income" (wages and salaries), v is "unearned income" (interest, dividends, rents), and w is "other income" (social security benefits and pensions, etc.). Clearly, a persons total income is the sum of the incomes he or she receives from each source $y_i = u_i + v_i + w_i$. Then

$$\bar{y} = \bar{u} + \bar{v} + \bar{w}\,.$$

So it doesn't matter if we take the means from each income source and then add them together to find the mean total income, or add the each individuals incomes from all sources to get his/her total income and then take the mean of that. We get the same value either way.

The mean combines well. The mean of a combined set is the weighted average of the means of the constituent sets, where weights are proportions each constituent set is to the combined set. For example, the data may come from two sources, males and females who had been interviewed separately. The overall mean would be the weighted average of the male mean and the female mean where the weights are the proportions of males and females in the sample.

The mean is the first moment or center of gravity of the numbers. We can think of the mean as the balance point if an equal weight was placed at each of the data points on the (weightless) number line. The mean would be the balance point of the line. This leads to the main disadvantage of the mean. It is strongly influenced by *outliers*. A single observation much bigger than the rest of the observations has a large effect on the mean. That makes using the mean problematic with highly skewed data such as personal income. Figure 3.9 shows how the mean is influenced by an outlier.

Calculating mean for grouped data. When the data have been put in a frequency table, we only know between which boundaries each observation lies. We

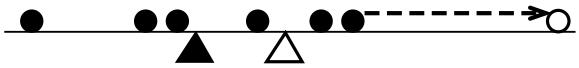


Figure 3.9 The mean as the balance point of the data is affected by moving the outlier.

no longer have the actual data values. In that case there are two assumptions we can make about the actual values.

1. All values in a group lie at the group midpoint.
2. All the values in a group are evenly spread across the group.

Fortunately, both these assumptions lead us to the same calculation of the mean value. The total contribution for all the observations in a group is the midpoint times the frequency under both assumptions.

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{J} n_j \times m_j$$

$$= \sum_{j=1}^{J} \frac{n_j}{n} \times m_j \,,$$

where n_j is the number of observations in the j^{th} interval, n is the total number of observations, and m_j is the midpoint of the j^{th} interval.

Median: Advantages and Disadvantages

The median of a set of numbers is the number such that 50% of the numbers are less than or equal to it, and 50% of the numbers are greater than or equal to it. Finding the median requires us to sort the numbers. It is the middle number when the sample size is odd, or it is the average of the two numbers closest to middle when the sample size is even.

$$m = y_{[\frac{n+1}{2}]} \,.$$

The median is not influenced by outliers at all. This makes it very suitable for highly skewed data like personal income. This is shown in Figure 3.10. However it does not have same good mathematical properties as mean. The median of a sum is not necessarily the sum of the medians. Neither does it have good combining properties similar to those of the mean. The median of the combined sample is not necessarily the weighted average of the medians. For these reasons, the median is not used as often as the mean. It is mainly used for very skewed data such as incomes where there are outliers which would unduly influence the mean, but don't affect the median.

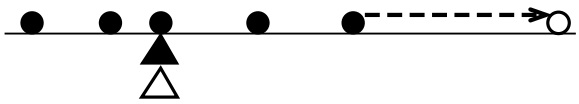


Figure 3.10 The median as the middle point of the data is not affected by moving the outlier.

Trimmed mean. We find the trimmed mean with degree of trimming equal to *k* by first ordering the observations, then trimming the lower *k* and upper *k* order statistics, and taking the average of those remaining.

$$\bar{x}_k = \frac{\sum_{i=k+1}^{n-k} x_{[i]}}{n-2k} .$$

We see that $\bar{x}_0$ (where there is no trimming) is the mean. If n is odd and we let $k = \frac{n}{2}$ then $\bar{x}_k$ is the median. Similarly if n is even and we let $k = \frac{n-2}{2}$ then $\bar{x}_k$ is the median. If k is small, the trimmed mean will have properties similar to the mean. If k is large, the trimmed mean has properties similar to the median.

3.4 MEASURES OF SPREAD

After we have determined where the data set is located on the number line, the next important aspect of the data set distribution is determining how spread out the data distribution is. If the data are very variable, the data set will be very spread out. So measuring spread gives a measure of the variability. We will look at some of the common measures of variability.

Range: Advantage and Disadvantage

The range is the largest observation minus smallest:

$$R = y_{[n]} - y_{[1]} .$$

The range is very easy to find. However, the largest and smallest observation are the observations that are most likely to be outliers. Clearly, the range is extremely influenced by outliers.

Interquartile Range: Advantages and Disadvantages

The interquartile range measures the spread of the middle 50% of the observations. It is the third quartile minus first quartile

$$IQR = Q_3 - Q_1 .$$

The quartiles are not outliers, so the interquartile range is not influenced by outliers. Nevertheless it is not used very much in inference because like the median it doesn't have good math or combining properties.

Variance: Advantages and Disadvantages

The variance of a data set is the average squared deviation from the mean.[2]

$$Var(y) = \frac{1}{n} \times \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

In physical terms, it is the second moment of inertia about the mean. Engineers refer to the variance as the *MSD*, mean squared deviation. It has good mathematical properties, although more complicated than those for the mean. The variance of a sum (of independent variables) is the sum of the individual variances.

It has good combining properties, although more complicated than those for the mean. The variance of a combined set is the weighted average of the variances of the constituent sets, plus the weighted average of the squares of the constituent means away from the combined mean, where weights are proportions each constituent set is to the combined set.

Squaring the deviations from the mean emphasizes the observations far from the mean. Those observations have large magnitude in a positive or negative direction already, and squaring them makes them much larger still, and all positive. Thus the variance is very influenced by outliers. The variance is in squared units. Thus its size is not comparable to mean.

Calculating variance for grouped data. The variance is the average squared deviation from the mean. When the data have been put in a frequency table, we no longer have the actual data values. In that case there are two assumptions we can make about the actual values.

1. All values in a group lie at the group midpoint.

2. All the values in a group are evenly spread across the group.

Unfortunately, these two assumptions lead us to different calculation of the variance. Under the first assumption we get the approximate formula

$$Var(y) = \frac{1}{n} \sum_{j=1}^{J} n_j \times (m_j - \bar{y})^2 ,$$

where n_j is the number of observations in the j^{th} interval, n is the total number of observations, m_j is the midpoint of the j^{th} interval. This formula only contains between group variation, and ignores the variation for the observations in the same

[2]Note that we are defining the variance of a data set using the divisor n. We aren't making any distinction over whether our data set is the whole population or only a sample from the population. Some books define the variance of a sample data set using divisor $n - 1$. One *degree of freedom* has been lost because for a sample, we are using the sample mean instead of the unknown population mean. When we use the divisor $n - 1$ we are calculating the *sample estimate of the variance*, not the variance itself.

group. Under the second assumption we add in the variation within each group to get the formula

$$Var(y) = \frac{1}{n}\sum_{j=1}^{J}\left(n_j \times (m_j - \bar{y})^2 + n_j \times \frac{R_j^2}{12}\right),$$

where R_j is the upper boundary minus the lower boundary for the j^{th} group.

Standard Deviation: Advantages and Disadvantages

The standard deviation is the square root of the variance.

$$sd(y) = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n}(y_i - \bar{y})^2}\,.$$

Engineers refer to it as the *RMS*, root mean square. It is not as affected by outliers as variance, but still quite affected. It inherits good mathematical properties and good combining properties from the variance. The standard deviation is the most widely used measure of spread. It is in the same units as mean, so its size is directly comparable to the mean.

3.5 DISPLAYING RELATIONSHIPS BETWEEN TWO OR MORE VARIABLES

Sometimes our data are measurements for two variables for each experimental unit. This is called *bivariate* data. We want to investigate the relationship between the two variables.

Scatterplot

The scatterplot is just a two-dimensional dotplot. Mark off the horizontal axis for the first variable, the vertical axis for the second. Each point is plotted on the graph. The shape of the "point cloud" gives us an idea as to whether the two variables are related, and if so, what type relationship.

When we have two samples of bivariate data, and want to see if the relationship between the variables is similar in the two samples, we can plot the points for both samples on the same scatterplot using different symbols so we can tell them apart.

Example 4 *The Bears.mtw file stored in Minitab contains 143 measurements on wild bears that were anesthetized, measured, tagged, and released. Figure 3.11 shows a scatterplot of head length versus head width for these bears. From this we can observe that head length and head width are related. Bears with large width heads tend to have heads that are long. We can also see that male bears tend to have larger heads than female bears.*

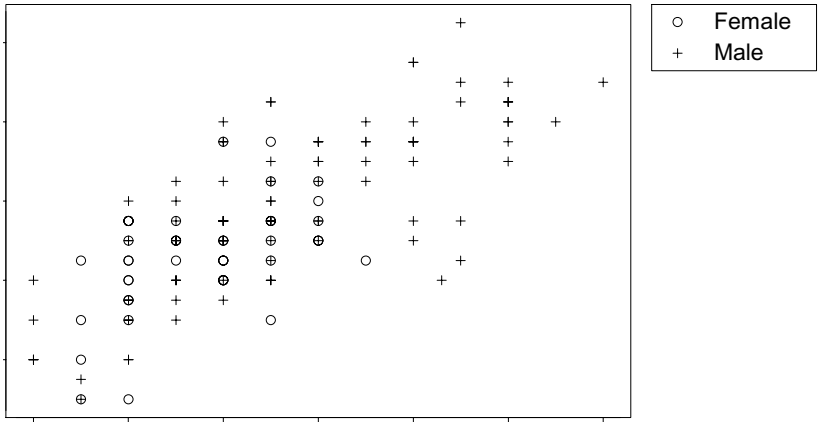


Figure 3.11 Head length versus head width in black bears.

Scatterplot Matrix

Sometimes our data consists of measurements of several variables on each experimental unit. This is called *multivariate* data. To investigate the relationships between the variables, form a *scatterplot matrix*. This means that we construct the scatterplot for each pair of variables, then display them in an array like a matrix. We look at each scatterplot in turn to investigate the relationship between that pair of the variables. More complicated relationships between three or more of the variables may be hard to see on this plot.

Example 4 (continued) *Figure 3.12 shows a scatterplot matrix showing scatterplots of head length, head width, neck girth, length, chest girth, and weight for the bear measurement data. We see there are strong positive relationships among the variables, and some of them appear to be nonlinear.*

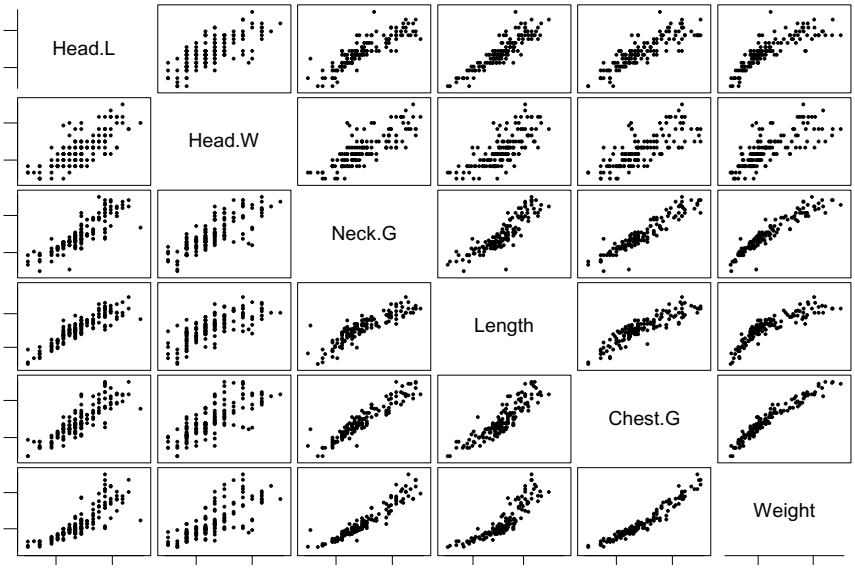


Figure 3.12 Scatterplot matrix of bear data.

3.6 MEASURES OF ASSOCIATION FOR TWO OR MORE VARIABLES

Covariance and Correlation between Two Variables

The covariance of two variables is the average of *first variable minus its mean* times *second variable minus its mean*:

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\,.$$

This measures how the variables vary together. Correlation between two variables is the covariance of the two variables divided by product of standard deviations of the two variables. This standardizes the correlation to lie between -1 and $+1$.

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x) \times Var(y)}}\,.$$

Correlation measures the strength of the *linear* relationship between two variables. A correlation of +1 indicates the points lie on a straight line with positive slope. A correlation of -1 indicates the points lie on a straight line with negative slope. A positive correlation that is less than one indicates that the points are scattered, but generally low values of the first variable are associated with low values of the second,

Table 3.4 Correlation matrix for bear data

	Head.L	Head.W	Neck.G	Length	Chest.G	Weight
Head.L	1.000	.744	.862	.895	.854	.833
Head.W	.744	1.000	.805	.736	.756	.756
Neck.G	.862	.805	1.000	.873	.940	.943
Length	.895	.736	.873	1.000	.889	.875
Chest.G	.854	.756	.940	.889	1.000	.966
Weight	.833	.756	.943	.875	.966	1.000

and high values of the first are associated with high values of the second. The higher the correlation, the more closely the points are bunched around a line. A negative correlation has low values of the first associated with high values of the second, and high values of the first associated with low values of the second. A correlation of 0 indicates that there is no association of low values or high values of the first with either high or low values of the second. It does not mean the variables are not related, only that they are not linearly related.

When we have more than two variables, we put the correlations in a matrix. The correlation between x and y equals the correlation between y and x, so the correlation matrix is symmetric about the main diagonal. The correlation of any variable with itself equals one.

Example 4 (continued) *The correlation matrix for the bear data is given in Table 3.4. We see that all the variables are correlated with each other. Looking at the matrix plot we see that Head.L and Head.W have a correlation of .744, and the scatterplot of those two variables is spread out. We see that the Head.L and Length have a higher correlation of .895, and on the scatterplot of those variables, we see the points lie much closer to a line. We see that Chest.G and Weight are highly correlated at .966. On the scatterplot we see those points lie much closer to a line, although we can also see that actually they seem to lie on a curve that is quite close to a line.*

Main Points

- Data should always be looked at in several ways as the first stage in any statistical analysis. Often a good graphical display is enough to show what is going on, and no further analysis is needed. Some elementary data analysis tools are:
  - *Order Statistics*. The data when ordered smallest to largest. $y_{[1]}, \cdots, y_{[n]}$.
  - *Median*. The value that has 50% of the observations above it and 50% of the observations below it. This is

$$y_{[\frac{n+1}{2}]}.$$

It is the middle value of the order statistics when n is odd. When n is even, the median is the weighted average of the two closest order statistics:

$$y_{[\frac{n+1}{2}]} = \frac{1}{2} \times y_{[\frac{n}{2}]} + \frac{1}{2} \times y_{[\frac{n}{2}+1]} .$$

The median is also known as the second quartile.

- *Lower quartile*. The value that 25 % of the observations are below it and 75 % of the observations are above it. It is also known as the first quartile. It is

$$Q_1 = y_{[\frac{n+1}{4}]} .$$

If $\frac{n+1}{4}$ is not an integer, we find it by taking the weighted average of the two closest order statistics.

- *Upper quartile*. The value that 75 % of the observations are below it and 25 % of the observations are above it. It is also known as the upper quartile. It is

$$Q_3 = x_{[\frac{3(n+1)}{4}]} .$$

If $\frac{3(n+1)}{4}$is not an integer, the quartile is found by taking the weighted average of the two closest order statistics .

- When we are comparing samples graphically, it is important that they be on the same scale. We have to be able to get the correct visual comparison without reading the numbers on the axis. Some elementary graphical data displays are:

  - *Stem-and-leaf diagram*. An quick and easy graphic which allows us to extract information from a sample. A vertical stem is drawn with a numbers up to stem digit along linear scale. Each number is represented using its next digit as a leaf unit at the appropriate place along the stem. The leaves should be ordered away from the stem. It is easy to find (approximately) the quartiles by counting along the graphic. Comparisons are done with back-to-back stem-and-leaf diagrams.
  - *Boxplot*. A graphic along a linear axis where the central box contains the middle 50% of the observation, and a whisker goes out from each end of the box to the lowest and highest observation. There is a line through the box at the median. So it is a visual representation of the five numbers $y_{[1]}, Q_1, Q_2, Q_3, y_{[n]}$ that give a quick summary of the data distribution. Comparisons are done with stacked boxplots.
  - *Histogram*. A graphic where the group boundaries are put on a linear scaled horizontal axis. Each group is represented by a vertical bar where the *area* of the bar is proportional to the frequency in the group.
  - *Cumulative frequency polygon* (ogive). A graphic where the group boundaries are put on a linearly scaled horizontal axis. The point *(lower boundary of lowest group, 0)* and the points *(upper group boundary, cumulative frequency)* are plotted and joined by straight lines. The median and quartiles can be found easily using the graph.

- It is also useful to summarize the data set using a few numerical summary statistics. The most important summary statistic of a variable is a measure of location which indicates where the values lie along the number axis. Some possible measures of location are:
  - *Mean*. The average of the numbers. It is easy to use, has good mathematical properties, and combines well. It is the most widely used measure of location. It is sensitive to outliers, so it is not particularly good for heavy tailed distributions.
  - *Median*. The middle order statistic, or the average of the two closest to the middle. This is harder to find as it requires sorting the data. It is not affected by outliers. The median doesn't have the good mathematical properties or good combining properties of the mean. Because of this, it is not used as often as the mean. Mainly it is used with distributions that have heavy tails or outliers, where it is preferred to the mean.
  - *Trimmed mean*. This is a compromise between the mean and the median. Discard the k largest and the k smallest order statistics and take the average of the rest.
- The second important summary statistic is a measure of spread, which shows how spread out are the numbers. Some commonly used measures of spread are:
  - *Range*. This is the largest order statistic minus the smallest order statistic. Obviously very sensitive to outliers.
  - *Interquartile range (IQR)*. This is the upper quartile minus the lower quartile. It measures the spread of the middle 50% of the observations. It is not sensitive to outliers.
  - *Variance*. The average of the squared deviations from the mean. Strongly influenced by outliers. The variance has good mathematical properties, and combines well, but it is in squared units and is not directly comparable to the mean.
  - *Standard deviation*. The square root of the variance. This is less sensitive to outliers than the variance and is directly comparable to the mean since it is in the same units. It inherits good mathematical properties and combining properties from the variance.
- Graphical display for relationship between two or more variables.
  - *Scatterplot*. Look for pattern.
  - *Scatterplot matrix*. An array of scatterplots for all pairs of variables.
- *Correlation* is a numerical measure of the strength of the *linear relationship* between the two variables. It is standardized to always lie between -1 and $+1$. If the points lie on a line with negative slope, the correlation is -1, and if

they lie on a line with positive slope, the correlation is $+1$. A correlation of 0 doesn't mean there is no relationship, only that there is no *linear* relationship.

Exercises

3.1 A study on air pollution in a major city measured the concentration of sulphur dioxide on 25 summer days. The measurements were:

3	9	16	23	29
3	11	17	25	35
5	13	18	26	43
7	13	19	27	44
9	14	23	28	46

(a) Form a stem-and-leaf diagram of the sulphur dioxide measurements.

(b) Find the median, lower quartile, and upper quartile of the measurements.

(c) Sketch a boxplot of the measurements.

3.2 Dutch elm disease is spread by bark beetles that breed in the diseased wood. A sample of 100 infected elms was obtained, and the number of bark beetles on each tree was counted. The data are summarized in the following table:

Boundaries	Frequency
$0 < x \leq 50$	8
$50 < x \leq 100$	24
$100 < x \leq 150$	33
$150 < x \leq 200$	21
$200 < x \leq 400$	14

(a) Graph a histogram for the bark beetle data.

(b) Graph a cumulative frequency polygon of the bark beetle data. Show the median and quartiles on your cumulative frequency polygon.

3.3 A manufacturer wants to determine whether the distance between two holes stamped into a metal part is meeting specifications. A sample of 50 parts was taken, and the distance was measured to nearest tenth of a millimeter. The results were:

300.6	299.7	300.2	300.0	300.1
300.0	300 .1	299.9	300.2	300.1
300.5	299.6	300.7	299.9	300.2
299.9	300.4	299.8	300.4	300.4
300.4	300.2	299.4	300.6	299.8
299.7	300.1	299.9	300.0	300.0
300.5	300.1	299.9	299.8	300.2
300.7	300.4	300.0	300.1	300.0
300.2	300.3	300.5	300.0	300.1
300.3	299.9	300.1	300.2	299.5

(a) Form a stem-and-leaf diagram of the measurements.

(b) Find the median, lower quartile, and upper quartile of the measurements.

(c) Sketch a boxplot of the measurements.

(d) Put the measurements in a frequency table with the following classes:

Boundaries	Frequency
$299.2 < x \leq 299.6$	
$299.6 < x \leq 299.8$	
$299.8 < x \leq 300.0$	
$300.0 < x \leq 300.2$	
$300.2 < x \leq 300.4$	
$300.4 < x \leq 300.8$	

(e) Construct a histogram of the measurements.

(f) Construct a cumulative frequency polygon of the measurements. Show the median and quartiles.

3.4 The manager of a government department is concerned about the efficiency in which his department serves the public. Specifically he is concerned about the delay experienced by members of the public waiting to be served. He takes a sample of 50 arriving customers, and measures the time each waits until service begins. The times (rounded off to the nearest second) are:

98	5	6	39	31
46	129	17	1	64
40	121	88	102	50
123	50	20	37	65
75	191	110	28	44
47	6	43	60	12
150	16	182	32	5
106	32	26	87	137
44	13	18	69	107
5	53	54	173	118

(a) Form a stem-and-leaf diagram of the measurements.

(b) Find the median, lower quartile, and upper quartile of the measurements.

(c) Sketch a boxplot of the measurements.

(d) Put the measurements in a frequency table with the following classes:

Boundaries	Frequency
$0 < x \leq 20$	
$20 < x \leq 40$	
$40 < x \leq 60$	
$60 < x \leq 80$	
$80 < x \leq 100$	
$100 < x \leq 200$	

(e) Construct a histogram of the measurements.

(f) Construct a cumulative frequency polygon of the measurements. Show the median and quartiles.

3.5 A random sample of 50 families reported the dollar amount they had available as a liquid cash reserve. The data have been put in the following frequency table:

Boundaries	Frequency
$0 < x \leq 500$	17
$500 < x \leq 1000$	15
$1000 < x \leq 2000$	7
$2000 < x \leq 4000$	5
$4000 < x \leq 6000$	3
$6000 < x \leq 10000$	3

(a) Construct a histogram of the measurements.

(b) Construct a cumulative frequency polygon of the measurements. Show the median and quartiles.

(c) Calculate the grouped mean for the data.

3.6 In this exercise we see how the default setting in the Minitab boxplot command can be misleading, since it doesn't take the sample size into account. We will generate three samples of different sizes from the same distribution, and compare their Minitab boxplots. Generate 250 *normal* $(0, 1)$ observations and put them in column c1 by pulling down the *calc* menu to the *random data* command over to *normal* and filling in the dialog box. Generate 1000 *normal* $(0, 1)$ observations the same way and put them in column c2, and generate 4000 *normal* $(0, 1)$ observations the same way and put them in column c3. Stack these three columns by pulling down the *manip* menu down to *stack/unstack* and over to *stack columns* and filling in the dialog box to put the stacked column into c4, with subscripts into c5. Form stacked boxplots by pulling down *graph* menu to boxplot command and filling in dialog box. Y is c4 and x is c5.

(a) What do you notice from the resulting boxplot?

(b) Which sample seems to have a heavier tail?

(c) Why is this misleading?

(d) Redo the boxplot highlighting the *outlier symbol* in the dialog box, and clicking on *edit attributes* and select *dot*.

(e) Is the graph still as misleading as the original?

3.7 McGhie and Barker (1984) collected 100 slugs from the species *Limax maximus* around Hamilton, New Zealand. They were preserved in a relaxed state, and their length in mm and weight in gm were recorded. Thirty of the observations are shown below. The full data are in the Minitab worksheet *slug.mtw*.

length (mm)	weight (gm)	length (mm)	weight (gm)	length (mm)	weight (gm)
73	3.68	21	0.14	75	4.94
78	5.48	26	0.35	78	5.48
75	4.94	26	0.29	22	0.36
69	3.47	36	0.88	61	3.16
60	3.26	16	0.12	59	1.91
74	4.36	35	0.66	78	8.44
85	6.44	36	0.62	90	13.62
86	8.37	22	0.17	93	8.70
82	6.40	24	0.25	71	4.39
85	8.23	42	2.28	94	8.23

(a) Plot weight on length using Minitab. What do you notice about the shape of the relationship?

(b) Often when we have a nonlinear relationship, we can transform the variables by taking logarithms and achieve linearity. In this case, weight is related to volume which is related to length times width times height. Taking logarithms of weight and length should give a more linear relationship. Plot log(weight) on log(length) using Minitab. Does this relationship appear to be linear?

(c) From the scatterplot of log(weight) on log(length) can you identify any points that do not appear to fit the pattern?

4

Logic, Probability, and Uncertainty

Most situations we deal with in everyday life are not completely predictable. If I think about the weather tomorrow at noon, I cannot be certain whether it will or will not be raining. I could contact the Meteorological Service and get the most up to date weather forecast possible, which is based on the latest available data from ground stations and satellite images. The forecast could be that it will be a fine day. I decide to take that forecast into account, and not take my umbrella. Despite the forecast it could rain and I could get soaked going to lunch. There is always uncertainty.

In this chapter we will see that deductive logic can only deal with certainty. This is of very limited use in most real situations. We need to develop inductive logic that allows us to deal with uncertainty.

Since we can't completely eliminate uncertainty, we need to model it. In real life when we are faced with uncertainty, we use plausible reasoning. We adjust our belief about something, based on the occurrence or nonoccurrence of something else. We will see how plausible reasoning should be based on the rules of probability which were originally derived to analyze the outcome of games based on random chance. Thus the rules of probability extend logic to include plausible reasoning where there is uncertainty.

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

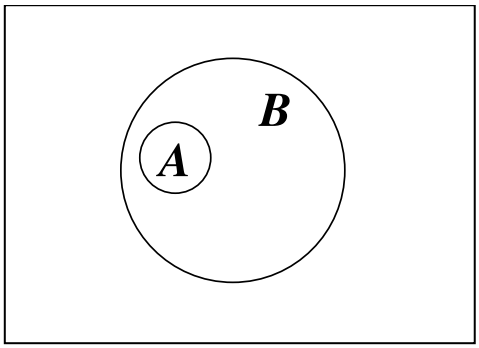


Figure 4.1 "If *A* is true then *B* is true." Deduction is possible.

4.1 DEDUCTIVE LOGIC AND PLAUSIBLE REASONING

Suppose we know "If proposition *A* is true, then proposition B is true." We are then told "proposition A is true." Therefore we know that "B is true." It is the only conclusion consistent with the condition. This is a deduction.

Again suppose we know "If proposition *A* is true, then proposition B is true." Then we are told "*B* is not true." Therefore we know that "*A* is not true." This is also a deduction. When we determine a proposition is true by deduction using the rules of logic, it is certain. Deduction works from the general to the particular.

We can represent propositions using diagrams. Propositions "*A* is true" and "*B* is true" are each represented by the interior of a circle. The proposition "*if A is true then B is true*" is represented by having circle representing *A* lie completely inside *B*. This is shown in Figure 4.1. The essence of the first deduction is that if we are in a circle *A* that lies completely inside circle *B*, then we must be inside circle *B*. Similarly, the essence of the second induction is that if we are outside of a circle *B* that completely contains circle *A*, then we must be outside circle *A*.

Other propositions can be seen in the diagram. Proposition "*A and B are both true*" is represented by the *intersection*, the region in both the circles simultaneously. In this instance, the intersection equals *A* by itself. The proposition "*A* or *B* is true" is represented by the *union*, region in either one or the other, or both of the circles. In this instance, the union equals *B* by itself.

On the other hand, suppose we are told "*A* is not true." What can we now say about *B*? Traditional logic has nothing to say about this. Both "*B* is true" and "*B* is not true" are consistent with the conditions given. Some points outside circle *A* are inside circle *B*, and some are outside circle *B*. No deduction is possible. Intuitively though, we would now believe that it was less plausible that *B* is true than we previously did before we were told "*A* is not true." This is because one of the ways B could be true, namely that *A* and *B* are both true is now no longer a possibility. And the ways that *B* could be false have not been affected.

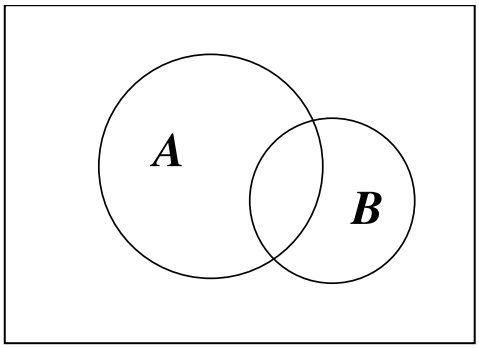


Figure 4.2 Both "*A* is true" and " *A* is false" are consistent with both "*B* is true" and "*B* is false." No deduction is possible here.

Similarly, when we are told "*B* is true," traditional logic has nothing to contribute. Both "*A* is true" and "*A* is not true" are consistent with the conditions given. Nevertheless, we see that "*B* is true" increases the plausibility of "*A* is true" because one of the ways *A* could be false, namely both *A* and *B* are false is no longer possible, and the ways that *A* are true have not been affected.

Often propositions are related in such a way that no deduction is possible. Both "*A* is true" and "*A* is false" are consistent with both "*B* is true" and "*B* is false." Figure 4.2 shows this by having the two circles intersect, and neither is completely inside the other.

Suppose we try to use numbers to measure plausibility of propositions. When we change our plausibility for some proposition on the basis of the occurrence of some other proposition, we are making an *induction*. Induction works from the particular to the general.

Desired Properties of Plausibility Measures

1. Degrees of plausibility are represented by nonnegative real numbers.
2. They qualitatively agree with common sense. Larger numbers mean greater plausibility.
3. If a proposition can be represented more than one way, then all representations must give the same plausibility.
4. We must always take all the relevant evidence into account.
5. Equivalent states of knowledge are always given the same plausibility.

R. T. Cox showed that any set of plausibilities that satisfies the desired properties given above, must operate according to the same rules as probability. Thus the sensible way to revise plausibilities is by using the rules of probability. Bayesian

statistics uses the rules of probability to revise our belief given the data. Probability is used as an extension of logic to cases where deductions cannot be made. Jaynes (1995) gives an excellent discussion on using probability as logic.

4.2 PROBABILITY

We start this section with the idea of a random experiment. In a random experiment, though we make the observation under known repeatable conditions, the outcome is uncertain. When we repeat the experiment under identical conditions, we may get a different outcome. We start with the following definitions:

- *Random experiment*. An experiment that has an outcome that is not completely predictable. We can repeat the experiment under the same conditions and not get the same result. Tossing a coin is an example of a random experiment.
- *Outcome*. The result of one single trial of the random experiment.
- *Sample space*. The set of all possible outcomes of one single trial of the random experiment. We denote it Ω. The sample space contains everything we are considering in this analysis of the experiment, so we also can call it the *universe*. In our diagrams we will call it U.
- *Event*. Any set of possible outcomes of a random experiment.

Possible events include the universe, U, and the set containing no outcomes, the empty set ϕ. From any two events E and F we can create other events by the following operations.

- *Union of two events*. The union of two events E and F is the set of outcomes in *either* E or F (inclusive or). Denoted $E \cup F$
- *Intersection of two events*. The intersection of two events E and F is the set of outcomes in both E and F simultaneously. Denoted $E \cap F$.
- *Complement of an event*. The complement of an event E is the set of outcomes not in E. Denoted $\tilde{E}$

We will use Venn diagrams to illustrate the relationship between events. Events are denoted as regions in the universe. The relationship between two events depends on the outcomes they have in common. If all the outcomes in one event are also in the other event, the first event is a subset of the other. This is shown in Figure 4.3.

If the events have some outcomes in common, but each has some outcomes that are not in the other, they are intersecting events. This is shown in Figure 4.4. Neither event is contained in the other.

If the two events have no outcomes in common, they are *mutually exclusive* events. In that case the occurrence of one of the events excludes the occurrence of the other, and vice versa. They are also referred to as *disjoint* events. This is shown in Figure 4.5

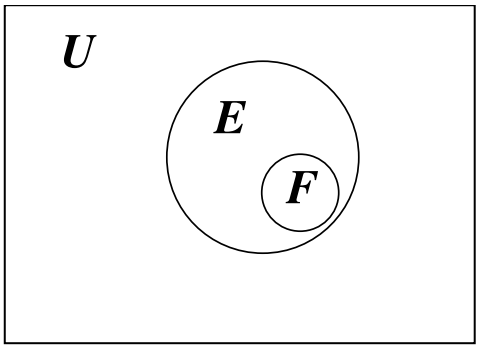


Figure 4.3 Event F is a subset of event E.

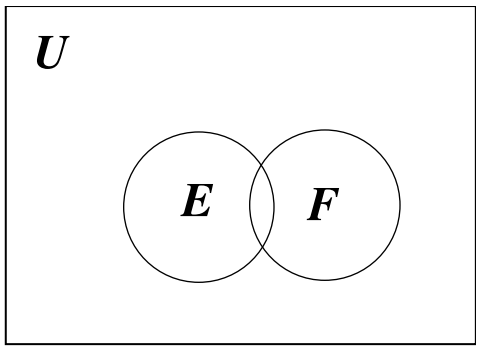


Figure 4.4 E and F are *intersecting* events.

4.3 AXIOMS OF PROBABILITY

The probability assignment for a random experiment is an assignment of probabilities to all possible events the experiment generates. These probabilities are real numbers between 0 and 1. The higher the probability of an event is, the more likely it is to occur. A probability that equals 1 means that event is certain to occur, and a probability of 0 means the event cannot possibly occur. To be consistent, the assignment of probabilities to events must satisfy the following axioms.

1. $P(A) \geq 0$ for any event A. (Probabilities are nonnegative.)
2. $P(U) = 1$. (Probability of universe = 1. Some outcome occurs every time you conduct the experiment.)
3. If A and B are mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$. (Probability is additive over *disjoint* events.)

The other rules of probability can be proved from the axioms.

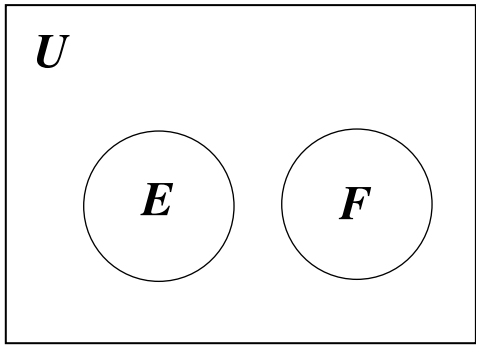


Figure 4.5 Event E and event F are *mutually exclusive* or *disjoint* events.

1. $P(\phi) = 0$. (The empty set has zero probability.)
   - $U = U \cup \phi$ and $U \cap \phi = \phi$. Therefore by axiom 3
   - $1 = 1 + P(\phi)$.
     qed
2. $P(\tilde{A}) = 1 - P(A)$. (The probability of a complement of an event.)
   - $U = A \cup \tilde{A}$ and $A \cap \tilde{A} = \phi$. Therefore by axiom 3
   - $1 = P(A) + P(\tilde{A})$.
     qed
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. (The addition rule of probability.)
   - $A \cup B = A \cup (\tilde{A} \cap B)$ and they are disjoint. Therefore by axiom 3
   - $P(A \cup B) = P(A) + P(\tilde{A} \cap B)$.
   - $B = (A \cap B) \cup (\tilde{A} \cap B)$, and they are disjoint. Therefore by axiom 3
   - $P(B) = P(A \cap B) + P(\tilde{A} \cap B)$. Substituting this in previous equation gives
   - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
     qed

   An easy way to remember this rule is to look at the Venn diagram of the events. The probability of the part $A \cap B$ has been included twice, once in $P(A)$ and once in $P(B)$, so it has to be subtracted out once.

4.4 JOINT PROBABILITY AND INDEPENDENT EVENTS

Figure 4.6 shows the Venn diagram for two events A and B in the universe U. The joint probability of events A and B is the probability that both events occur simultaneously,

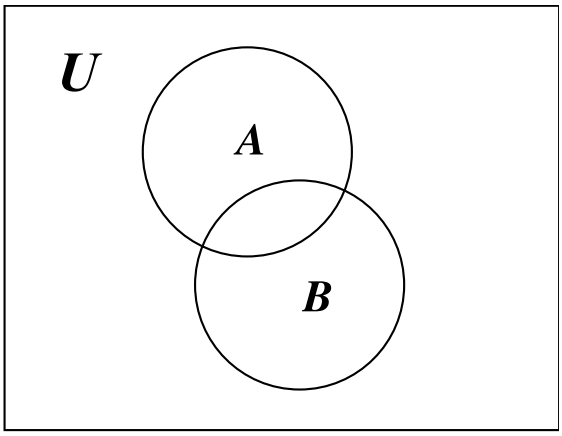


Figure 4.6 Two events A and B in the universe U.

on the same repetition of the random experiment. This would be the probability of the set of outcomes that are in both event A and event B, the intersection $A \cap B$. In other words the joint probability of events A and B is $P(A \cap B)$, the probability of their intersection.

If event A and event B are independent, then $P(A \cap B) = P(A) \times P(B)$. The joint probability is the product of the individual probabilities. If that does not hold the events are called *dependent* events. Note that whether or not two events A and B are independent or dependent depends on the probabilities assigned.

Distinction between independent events and mutually exclusive events. People often get confused between independent events and mutually exclusive events. This semantic confusion arises because the word *independent* has several meanings. The primary meaning of something being independent of something else is that the second thing has no affect on the first. This is the meaning of the word independent we are using in the definition of independent events. The occurrence of one event does not affect the occurrence or nonoccurrence of the other events.

There is another meaning of the word independent. That is the political meaning of independence. When a colony becomes independent of the mother country, it becomes a distinct separate country. That meaning is covered by the definition of *mutually exclusive* or *disjoint* events.

Independence of two events is not a property of the events themselves, rather it is a property that comes from the probabilities of the events and their intersection. This is in contrast to *mutually exclusive* events, which have the property that they contain no elements in common. Mutually exclusive events with nonnegative probability cannot be independent. Their intersection is the empty set, so it must have probability zero, which cannot equal the product of the probabilities of the two events!

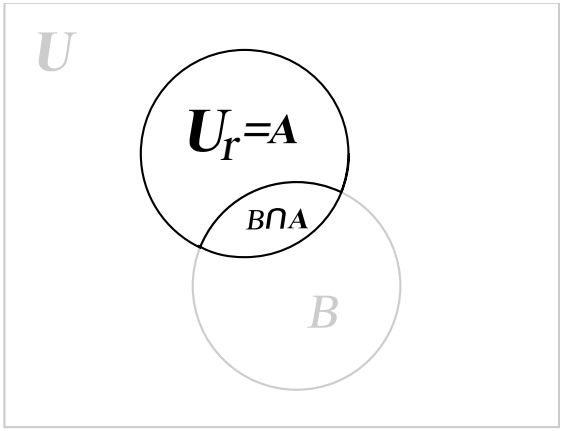


Figure 4.7 The reduced universe, given event A has occurred.

Marginal probability. The probability of one of the events A, in the joint event setting is called its marginal probability. It is found by summing $P(A \cap B)$ and $P(A \cap \tilde{B})$ using the axioms of probability.

- $A = (A \cap B) \cup (A \cap \tilde{B})$, and they are disjoint. Therefore by axiom 3
- $P(A) = P(A \cap B) + P(A \cap \tilde{B})$. The marginal probability of event A is found by summing its *disjoint* parts.
  qed

4.5 CONDITIONAL PROBABILITY

If we know that one event has occurred, does that affect the probability that another event has occurred? To answer this, we need to look at conditional probability.

Suppose we are told that the event A has occurred. Everything outside of A is no longer possible. We only have to consider outcomes inside event A. The *reduced universe* $U_r = A$. The only part of event B that is now relevant is that part which is also in A. This is $B \cap A$. Figure 4.7 shows that, given event A has occurred, the reduced universe is now the event A, and the only relevant part of event B is $B \cap A$.

Given that event A has occurred, the total probability in the reduced universe must equal 1. The probability of B given A is the unconditional probability of that part of B that is also in A, multiplied by the scale factor $\frac{1}{P(A)}$. That gives the conditional probability of event B given event A:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \,. \tag{4.1}$$

We see the conditional probability $P(B|A)$ is proportional to the joint probability $P(A \cap B)$ but has been rescaled so the probability of the reduced universe equals 1.

Conditional probability for independent events. Notice that when A and B are independent events

$$P(B|A) = P(B)\,,$$

since $P(B \cap A) = P(B) \times P(A)$ for independent events, and the factor $P(A)$ will cancel out. Knowledge about A does not affect the probability of B occurring when A and B are independent events! This shows that the definition we used for independent events is a reasonable one.

Multiplication rule. Formally, we could reverse the roles of the two events A and B. The conditional probability of A given B would be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}\,.$$

However, we will not consider the two events the same way. B is an unobservable event. That is, the occurrence or nonoccurrence of event B is not observed. A is an observable event that can occur either with event B or with its complement $\tilde{B}$. However, the chances of A occurring may depend on which one of B or $\tilde{B}$ has occurred. In other words, the probability of event A is conditional on the occurrence or nonoccurrence of event B. When we clear the fractions in the conditional probability formula we get

$$P(A \cap B) = P(B) \times P(A|B)\,. \tag{4.2}$$

This is known as the multiplication rule for probability. It restates the conditional probability relationship of an observable event given an unobservable event in a way that is useful for finding the joint probability $P(A \cap B)$. Similarly

$$P(A \cap \tilde{B}) = P(\tilde{B}) \times P(A|\tilde{B})\,.$$

4.6 BAYES' THEOREM

From the definition of conditional probability

$$P(B|A) = \frac{P(A \cap B)}{P(A)}\,.$$

We know that the *marginal* probability of event A is found by summing the probabilities of its *disjoint* parts. Since $A = (A \cap B) \cup (A \cap \tilde{B})$, and clearly $(A \cap B)$ and $(A \cap \tilde{B})$ are disjoint,

$$P(A) = P(A \cap B) + P(A \cap \tilde{B})\,.$$

We substitute this into the definition of conditional probability to get

$$P(B|A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap \tilde{B})}\,.$$

Now we use the multiplication rule to find each of these joint probabilities. This gives Bayes' theorem for a single event:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A|B) \times P(B) + P(A|\tilde{B}) \times P(\tilde{B})} \,. \tag{4.3}$$

Summarizing, we see Bayes' theorem is a restatement of the conditional probability $P(B|A)$ where:

1. The probability of A is found as the sum of the probabilities of its disjoint parts, $(A \cap B)$ and $(A \cap \tilde{B})$, and
2. Each of the joint probabilities are found using the multiplication rule.

The two important things to note are that the *union* of B and $\tilde{B}$ is the whole universe U, and that they are *disjoint*. We say that events B and $\tilde{B}$ partition the universe.

A set of events partitioning the universe. Often we have a set of more than two events that partition the universe. For example, suppose we have n events $B_1, \cdots, B_n$ such that:

- The union $B_1 \cup B_2 \cup \cdots \cup B_n = U$, the universe, and
- Every distinct pair of the events are disjoint, $B_i \cap B_j = \phi$ for $i = 1, \ldots, n$, $j = 1, \ldots, n$, and $i \neq j$.

Then we say the set of events $B_1, \cdots, B_n$ *partitions* the universe. An observable event A will be partitioned into parts by the partition. $A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots (A \cap B_n)$. $(A \cap B_i)$ and $(A \cap B_j)$ are disjoint since B_i and B_j are disjoint. Hence

$$P(A) = \sum_{j=1}^{n} P(A \cap B_j) \,.$$

This is known as the law of total probability. It just says the probability of an event A is the sum of the probabilities of its disjoint parts. Using the multiplication rule on each joint probability gives

$$P(A) = \sum_{j=1}^{n} P(A|B_j) \times P(B_j) \,.$$

The conditional probability $P(B_i|A)$ for $i = 1, \ldots, n$ is found by dividing each joint probability by the probability of the event A.

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} \,.$$

Using the multiplication rule to find the joint probability in the numerator, and the law of total probability in the denominator gives

$$P(B_i|A) = \frac{P(A|B_i) \times P(B_i)}{\sum_{j=1}^{n} P(A|B_j) \times P(B_j)} \,. \tag{4.4}$$

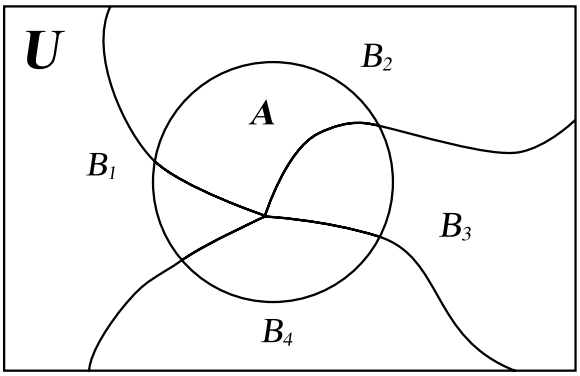


Figure 4.8 Four events B_i for $i = 1, \ldots, 4$ that partition the universe U, and event A.

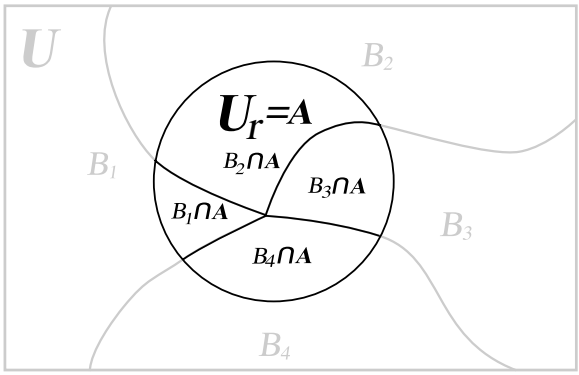


Figure 4.9 The reduced universe given event A has occurred, together with the four events partitioning the universe.

This is a result known as Bayes' theorem published posthumously in 1763 after the death of its discoverer, Reverend Thomas Bayes.

Example 5 *Suppose $n = 4$. Figure 4.8 shows the four unobservable events $B_1, \ldots, B_4$ that partition the universe U, and an observable event A. Now let us look at the conditional probability of B_i given A has occurred. Figure 4.9 shows the reduced universe, given event A has occurred. The conditional probabilities are the probabilities on the reduced universe, scaled up so they sum to 1. They are given by Equation 4.4.*

Bayes' theorem is really just a restatement of the conditional probability formula, where the joint probability in the numerator is found by the multiplication rule, and the marginal probability found in the denominator is found using the law of total probability followed by the multiplication rule. Note how the events A and B_i for

$i = 1, \ldots, n$ are not treated symmetrically. The events B_i for $i = 1, \ldots, n$ are considered unobservable. We never know which one of them occurred. The event A is an observable event. The marginal probabilities $P(B_i)$ for $i = 1, \ldots, n$ are assumed known before we start, and called our *prior* probabilities.

Bayes' Theorem: The Key to Bayesian Statistics

To see how we can use Bayes' theorem to revise our beliefs on the basis of evidence we need to look at each part. Let $B_1, \ldots, B_n$ be a set of unobservable events which partition the universe. We start with $P(B_i)$ for $i = 1, \ldots, n$, the *prior* probability for the events B_i, for $i = 1, \ldots, n$. This distribution gives the weight we attach to each of the B_i from our prior belief. Then we find that A has occurred.

The *likelihood* of the unobservable events $B_1, \ldots, B_n$ is the conditional probability that A has occurred given B_i for $i = 1, \ldots, n$. Thus the *likelihood* of event B_i is given by $P(A|B_i)$. We see the *likelihood* is a function defined on the events $B_1, \ldots, B_n$. The *likelihood* is the weight given to each of the B_i events given by the occurrence of A.

$P(B_i|A)$ for $i = 1, \ldots, n$ is the *posterior* probability of event B_i given A has occurred. This distribution contains the weight we attach to each of the events B_i for $i = 1, \ldots n$ after we know event A has occurred. It combines our prior beliefs with the evidence given by the occurrence of event A.

The Bayesian universe. We can get better insight into Bayes' theorem if we think of the universe as having two dimensions, one observable, and one unobservable. We let the observable dimension be horizontal, and let the unobservable dimension be vertical. The unobservable events no longer partition the universe haphazardly. Instead, they partition the universe as rectangles that cut completely across the universe in a horizontal direction. The whole universe consists of these horizontal rectangles in a vertical stack. Since we don't ever observe which of these events occurred, we never know what vertical position we are in the Bayesian universe.

Observable events are vertical rectangles, that cut the universe from top to bottom. We observe that vertical rectangle A has occurred, so we observe the horizontal position in the universe.

Each event $B_i \cap A$ is a rectangle at the intersection of B_i and A. The probability of the event $B_i \cap A$ is found by multiplying the prior probability of B_i times the conditional probability of A given B_i. This is the multiplication rule.

The event A is the union of the disjoint parts $A \cap B_i$ for $i = 1, \ldots, n$. The probability of A is clearly the sum of the probabilities of each of the disjoint parts. The probability of A is found by summing the probabilities of each disjoint part down the vertical column represented by A. This is the *marginal* probability of A.

The posterior probability of any particular B_i given A is the proportion of A that is also in B_i. In other words, the probability of that $B_i \cap A$ divided by the sum of $B_j \cap A$ summed over all $j = 1, \ldots, n$.

In Bayes' theorem, each of the joint probabilities are found by multiplying the *prior* probability $P(B_i)$ times the *likelihood* $P(A|B_i)$. In Chapter 5, we will see that

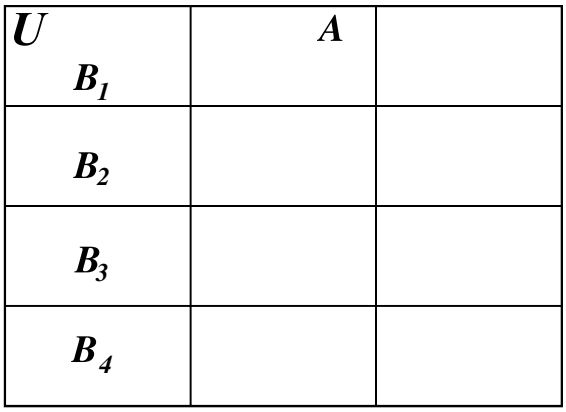


Figure 4.10 The Bayesian universe U with four unobservable events B_i for $i = 1, \ldots, 4$ which partition it shown in the vertical dimension, and the observable event A shown in the horizontal dimension.

the universe set out with two dimensions for two jointly distributed discrete random variables is very similar to that shown in Figures 4.10 and 4.11. One random variable will be observed, and we will determine the conditional probability distribution of the other random variable, given our observed value of the first. In Chapter 6, we will develop Bayes' theorem for two discrete random variables in an analogous manner to our development of Bayes' theorem for events in this chapter.

Example 5 (continued) *Figure 4.10 shows the four unobservable events B_i for $i = 1, \ldots, 4$ that partition the Bayesian universe, together with event A which is observable. Figure 4.11 shows the reduced universe, given event A has occurred. These figures will give us better insight than Figures 4.8 and 4.9. We know where in the Bayesian universe we are in the horizontal direction since we know event A occurred. However we don't know where we are in the vertical direction since we don't know which one of the B_i occurred.*

Multiplying by constant. The numerator of Bayes' theorem is the prior probability times the likelihood. The denominator is the sum of the prior probabilities times likelihoods over the whole partition. This division of the prior probability times likelihood by the sum of prior probabilities times likelihoods makes the posterior probability sum to 1.

Note, if we multiplied each of the likelihoods by a constant, the denominator would also be multiplied by the same constant. The constant would cancel out in the division, and we would be left with the same posterior probabilities. Because of this, we only need to know the likelihood to within a constant of proportionality. The *relative* weights given to each of the possibilities by the likelihood is all we need. Similarly, we could multiply each prior probability by a constant. The denominator would again be multiplied by the same constant, so we would be left with the same posterior probabilities. The only thing we need in the prior is the *relative* weights we

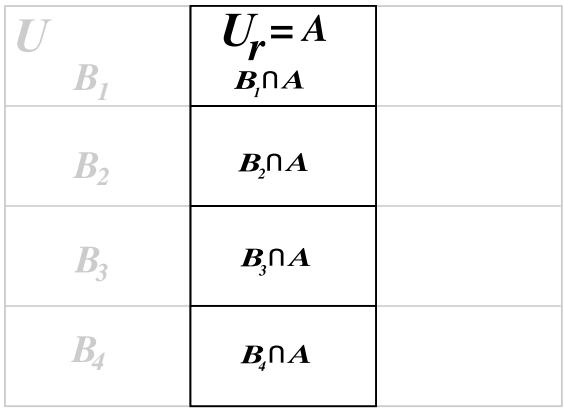


Figure 4.11 The reduced Bayesian universe, given A has occurred, together with the four unobservable events B_i for $i = 1, \ldots, 4$ that partition it.

give to each of the possibilities. We often write Bayes theorem in its proportional form as

$$posterior \propto prior \times likelihood$$

This gives the relative weights for each of the events B_i for $i = 1, \ldots, n$ after we know A has occurred. Dividing by the sum of the relative weights rescales the relative weights so they sum to 1. This makes it a probability distribution.

We can summarize the use of Bayes' theorem for events by the following three steps:

1. Multiply *prior* times *likelihood* for each of the B_i. This finds the probability of $B_i \cap A$ by the multiplication rule.
2. Sum them for $i = 1, \ldots, n$. This finds the probability of A by the law of total probability.
3. Divide each of the prior times likelihood values by their sum. This finds the conditional probability of that particular B_i given A.

4.7 ASSIGNING PROBABILITIES

Any assignment of probabilities to all possible events must satisfy the probability axioms. Of course, to be useful the probabilities assigned to events must correspond to the real world. There are two methods of probability assignment that we will use:

1. *Long run relative frequency probability assignment* : the probability of an event is considered to be the proportion of times it would occur if the experiment was repeated an infinite number of repetitions. This is the method of assigning probabilities used in frequentist statistics. For example, if I was trying to

assign the probability of getting a head on a toss of a coin, I would toss it a large number of times, and use the proportion of heads that occurred as an approximation to the probability.

2. *Degree of belief probability assignment*: the probability of an event is what I believe it is from previous experience. This is subjective. Someone else can have a different belief. For example, I could say that I believe the coin is a fair one, so for me, the probability of getting a head equals .5. Someone else might look at the coin and observing a slight asymmetry he/she might decide the probability of getting a head equals .49.

In Bayesian statistics, we will use long run relative frequency assignments of probabilities for events that are outcomes of the random experiment, given the value of the unobservable variable. We call the unobservable variable the *parameter*. Think about repeating the experiment over and over again an infinite number of times while holding the parameter (unobservable) at a fixed value. The set of all possible observable values of the experiment is called the *sample space* of the experiment. The probability of an event is long run relative frequency of the event over all these hypothetical repetitions. We see the *sample space* is the observable (horizontal) dimension of the *Bayesian universe*.

The set of all possible values of the parameter (unobservable) is called the *parameter space*. It is the unobservable (vertical) dimension of the *Bayesian universe*. In Bayesian statistics we also consider the parameter value to be random. The probability I assign to an event "the parameter has a certain value" can't be assigned by long run relative frequency. To be consistent with the idea of a fixed but unknown parameter value, I must assign probabilities by degree of belief. This shows the relative plausibility I give to all the possible parameter values, before the experiment. Someone else would have different probabilities assigned according to his/her belief.

I am modelling my uncertainty about the parameter value by a single random draw from my prior distribution. I do not consider hypothetical repetitions of this draw. I want to make my inference about the parameter value drawn this particular time, given this particular data. Earlier in the chapter we saw that using the rules of probability is the only consistent way to update our beliefs given the data. So probability statements about the parameter value are always subjective, since they start with subjective prior belief.

4.8 ODDS RATIOS AND BAYES FACTOR

Another way of dealing with uncertain events that we are modelling as random, is to form the *odds ratio* of the event. The odds ratio for an event C equals the probability of the event occurring divided by the probability of the event not occurring:

$$odds(C) = \frac{P(C)}{P(\tilde{C})}\,.$$

Since the probability of the event not occurring equals one minus the probability of the event, there is a one to one relationship between the odds of an event and its probability.

$$odds(C) = \frac{P(C)}{(1 - P(C))} .$$

If we are using prior probabilities, we get the *prior* odds ratio. In other words, the ratio before we have analyzed the data. If we are using posterior probabilities we get the *posterior* odds ratio.

Solving the equation for the probability of event C we get

$$P(C) = \frac{odds(C)}{(1 + odds(C))} .$$

We see that there is a one-to-one correspondence between odds ratios and probabilities.

Bayes Factor (B)

The Bayes factor B contains the evidence in the data D that occurred relevant to the question about C occurring. It is the factor by which the prior odds is changed to the posterior odds:

$$prior\ odds(C) \times B = posterior\ odds(C) .$$

We can solve this relationship for the Bayes factor to get

$$B = \frac{posterior\ odds}{prior\ odds} .$$

We can substitute in the ratio of probabilities for both the posterior and prior odds ratios to find

$$B = \frac{P(D|C)}{P(D|\tilde{C})} .$$

Thus the Bayes factor is the ratio of the probability of getting the data which occurred given the event, to the probability of getting the data which occurred given the complement of the event. If the Bayes factor is greater than 1, then the data has made us believe that event is more probable than we thought before. If the Bayes factor is less than 1, then the data has made us believe the event is less probable than we originally thought.

Main Points

- *Deductive logic*. A logical process for determining the truth of a statement from knowing the truth or falsehood of other statements that the first statement is a consequence of. Deduction works from the general to the particular. We

can make a deduction from a known population distribution to determine the sampling distribution of a statistic.

- Deductions do not have the possibility of error.
- *Inductive logic*. A process, based on plausible reasoning, for inferring the truth of the statement from knowing the truth or falsehood of other statements which are consequences of the first statement. It works from the particular to the general. Statistical inference is an inductive process for making inferences about the parameter, on the basis of the observed statistic from the sampling distribution given the parameter.
- There is always the possibility of error when making an inference.
- Plausible reasoning should be based on the rules of probability to be consistent. They are:
  - Probability of an event is a nonnegative number.
  - Probability of the sample space (universe) equals 1.
  - The probability is additive over disjoint events.
- A *random experiment* is an experiment where the outcome is not exactly predictable, even when the experiment is repeated under the identical conditions.
- The set of all possible outcomes of a random experiment is called the *sample space* Ω. In frequentist statistics, the sample space is the universe for analyzing events based on the experiment.
- The *union* of two events A and B is the set of outcomes in A or B. This is an inclusive or. The union is denoted $A \cup B$.
- The *intersection* of two events A and B is the set of outcomes in both A and B simultaneously. The intersection is denoted $A \cap B$.
- The *complement* of event A is the set of outcomes not in A. The complement of event A is denoted $\tilde{A}$.
- *Mutually exclusive* events have no elements in common. There intersection $P(A \cap B)$ equals the empty set, ϕ.
- The conditional probability of event B given event A is given by

$$P(B|A) = \frac{P(A \cap B)}{P(A)} .$$

- The event B is unobservable. The event A is observable. We could nominally write the conditional probability formula for $P(A|B)$, but the relationship is not

used in that form. We do not treat the events symmetrically. The multiplication rule is the definition of conditional probability cleared of the fraction.

$$P(A \cap B) = P(B) \times P(A|B) \,.$$

It is used to assign probabilities to compound events.

- The *law of total probability* says that given events $B_1, \ldots, B_n$ that partition the sample space (universe), and another event A, then

$$P(A) = \sum_{j=1}^{n} P(B_j \cap A)$$

because probability is additive over the disjoint events, $(A \cap B_1) \ldots (A \cap B_n)$. When we find the probability of each of the intersections $A \cap B_j$ by the multiplication rule, we get

$$P(A) = \sum_{j} P(B_j) \times P(A|B_j) \,.$$

- Bayes' theorem is the key to Bayesian statistics:

$$P(B_i|A) = \frac{P(B_i) \times P(A|B_i)}{\sum_j P(B_j) \times P(A|B_j)} \,.$$

This comes from the definition of conditional probability. The marginal probability of the event A is found by the law of total probability, and each of the joint probabilities is found from the multiplication rule. $P(B_i)$ is called the prior probability of event B_i, and $P(B_i|A)$ is called the posterior probability of event B_i.

- In the Bayesian universe, the unobservable events $B_1, \ldots, B_n$ which partition the universe are horizontal slices, and the observable event A is a vertical slice. The probability $P(A)$ is found by summing the $P(A \cap B_i)$ down the column. Each of the $P(A \cap B_i)$ is found by multiplying the prior $P(B_i)$ times the likelihood $P(A|B_i)$. So Bayes' theorem can be summarized by saying the posterior probability is the prior times likelihood divided by the sum of the prior times likelihood.

- The Bayesian universe has two dimensions. The sample space forms the observable (horizontal) dimension of the Bayesian universe. The parameter space is the unobservable (vertical) dimension. In Bayesian statistics, the probabilities are defined on both dimensions of the Bayesian universe.

- The odds ratio of an event A is the ratio of the probability of the event to the probability of its complement:

$$odds(A) = \frac{P(A)}{P(\tilde{A})} \,.$$

If it is found before analyzing the data, it is the prior odds ratio. If it is found after analyzing the data, it is the posterior odds ratio.

- The Bayes factor is the amount of evidence in the data that changes the prior odds to the posterior odds:

$$prior\ odds = B \times posterior\ odds\,.$$

Exercises

4.1 There are two events A and B. $P(A) = .4$ and $P(B) = .5$. The events A and B are independent.

(a) Find $P(\tilde{A})$.

(b) Find $P(A \cap B)$.

(c) Find $P(A \cup B)$.

4.2 There are two events A and B. $P(A) = .5$ and $P(B) = .3$. The events A and B are independent.

(a) Find $P(\tilde{A})$.

(b) Find $P(A \cap B)$.

(c) Find $P(A \cup B)$.

4.3 There are two events A and B. $P(A) = .4$ and $P(B) = .4$. $P(\tilde{A} \cap B) = .24$.

(a) Are A and B independent events? Explain why or why not.

(b) Find $P(A \cup B)$.

4.4 There are two events A and B. $P(A) = .7$ and $P(B) = .8$. $P(\tilde{A} \cap \tilde{B}) = .1$.

(a) Are A and B independent events? Explain why or why not.

(b) Find $P(A \cup B)$.

4.5 A single fair die is rolled. Let the event A be "the face showing is even." Let the event B be "the face showing is divisible by 3."

(a) List out the sample space of the experiment.

(b) List the outcomes in A, and find $P(A)$.

(c) List the outcomes in B, and find $P(B)$.

(d) List the outcomes in $A \cap B$, and find $P(A \cap B)$.

(e) Are the events A and B independent? Explain why or why not.

4.6 Two fair dice, one red and one green, are rolled. Let the event A be "the sum of the faces showing is equal to seven." Let the event B be "the faces showing on the two dice are equal."

(a) List out the sample space of the experiment.
(b) List the outcomes in A, and find $P(A)$.
(c) List the outcomes in B, and find $P(B)$.
(d) List the outcomes in $A \cap B$, and find $P(A \cap B)$.
(e) Are the events A and B independent? Explain why or why not.
(f) How would you describe the relationship between event A and event B?

4.7 Two fair dice, one red and one green, are rolled. Let the event A be "the sum of the faces showing is an even number." Let the event B be "the sum of the faces showing is divisible by 3."

(a) List the outcomes in A, and find $P(A)$.
(b) List the outcomes in B, and find $P(B)$.
(c) List the outcomes in $A \cap B$, and find $P(A \cap B)$.
(d) Are the events A and B independent? Explain why or why not.

4.8 Two dice are rolled. The red die has been loaded. Its probabilities are $P(1) = P(2) = P(3) = P(4) = \frac{1}{5}$ and $P(5) = P(6) = \frac{1}{10}$. The green die is fair. Let the event A be "the sum of the faces showing is an even number." Let the event B be "the sum of the faces showing is divisible by 3."

(a) List the outcomes in A, and find $P(A)$.
(b) List the outcomes in B, and find $P(B)$.
(c) List the outcomes in $A \cap B$, and find $P(A \cap B)$.
(d) Are the events A and B independent? Explain why or why not.

4.9 Suppose there is a medical diagnostic test for a disease. The *sensitivity* of the test is .95. This means that if a person has the disease, the probability that the test gives a positive response is .95. The *specificity* of the test is .90. This means that if a person does not have the disease, the probability that the test gives a negative response is .90, or that the *false positive* rate of the test is .10. In the population, 1% of the people have the disease. What is the probability that a person tested has the disease, given the results of the test is positive? Let D be the event "the person has the disease" and let T be the event "the test gives a positive result."

4.10 Suppose there is a medical screening procedure for a specific cancer that has *sensitivity* = .90, and *specificity* = .95. Suppose the underlying rate of the cancer in the population is .001. Let B be the event "the person has that specific cancer" and A be the event "the screening procedure gives a positive result."

(a) What is the probability that a person has the disease given the results of the screening is positive?
(b) Does this show that screening is effective in detecting this cancer?

5

Discrete Random Variables

In the previous chapter, we looked at random experiments in terms of events. We also introduced probability defined on events as a tool for understanding random experiments. We showed how conditional probability is the logical way to change our belief about an unobserved event, given we observed another related event. In this chapter, we introduce discrete random variables and probability distributions.

A random variable describes the outcome of the experiment in terms of a number. If the only possible outcomes of the experiment are distinct numbers separated from each other (e.g., counts), we say that the random variable is discrete. There are good reasons why we introduce random variables and their notation:

- It is quicker to describe an outcome as a random variable having a particular value than to describe that outcome in words. Any event can be formed from outcomes described by the random variable using union, intersection, and complements.

- The probability distribution of the discrete random variable is a numerical function. It is easier to deal with a numerical function than with probabilities being a function defined on sets (events). The probability of any possible event can be found from the probability distribution of the random variable using the rules of probability. So instead of having to know the probability of every possible event, we only have to know the probability distribution of the random variable.

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

Table 5.1 Typical results of rolling a fair die

	Proportion After					
Value	10 Rolls	100 Rolls	1,000 Rolls	10,000 Rolls	...	Probability
1	.1	.15	.198	.1704	...	.1666
2	.3	.08	.163	.1661	...	.1666
3	.1	.20	.156	.1670	...	.1666
4	.2	.16	.153	.1698	...	.1666
5	.0	.22	.165	.1583	...	.1666
6	.3	.19	.165	.1684	...	.1666

- It becomes much easier to deal with compound events made up from repetitions of the experiment.

5.1 DISCRETE RANDOM VARIABLES

A number that is determined by the outcome of a random experiment is called a random variable. Random variables are denoted with uppercase letters, e.g., Y. The value the random variable takes is denoted by lowercase letter, e.g., y. A discrete random variable, Y, can only take on separated values y_k. There can be a finite possible number of values, for example, the random variable defined as "number of heads in n tosses of a coin" has possible values $0, 1, \ldots, n$. Or there can be a countably infinite number of possible values, for example the random variable defined as "number of tosses until the first head" has possible values $1, 2, \cdots, \infty$. The key thing for discrete random variables is that the possible values are separated by gaps.

Thought Experiment 1: *Roll of a die*
Suppose we have a fair six-faced die. Our random experiment is to roll it, and we let the random variable Y be the number on the top face. There are six possible values $1, 2, \ldots, 6$. Since the die is fair, those six values are equally likely. Now, suppose we take independent repetitions of the random variable, and record each occurrence of Y. Table 5.1 shows the proportion of times each face has occurred in a typical sequence of rolls of the die, after 10, 100, 1,000, and 10,000 rolls. The last column shows the true probabilities for a fair die.

We note that the proportions taking any value are getting closer and closer to the true probability of that value as n increases to ∞. We could draw graphs of the proportions having each value. These are shown in Figure 5.1. The graphs are at zero for any other y value, and have a spike at each possible value where the spike height equals the proportion of times that value occurred. The sum of spike heights equals one.

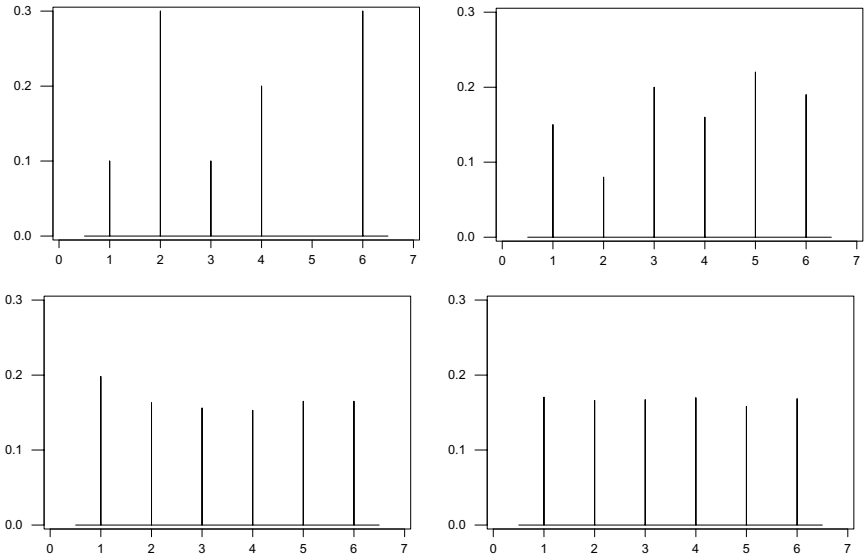


Figure 5.1 Proportions resulting from 10, 100, 1,000, and 10,000 rolls of a fair die.

Thought Experiment 2: *Random sampling from a finite population*

Suppose we have a finite population of size N. There can be at most a finite number of possible values, and they must be discrete, since there must be a gap between every pair of two real numbers. Some members of the population have the same value, so there are only K possible values $y_1, \ldots, y_K$. The probability of observing the value y_k is the proportion of population having that value.

We start by randomly drawing from the population with replacement. Each draw is done under identical conditions. If we continue doing the sampling, eventually we have seen all possible values. After each draw we update the proportions in the accumulated sample that have each value. We sketch a graph with a spike at each value in the sample equal to the proportion in the sample having that value. The updating of the graph at step n is made by scaling all the existing spikes down by the ratio $\frac{n-1}{n}$ and adding $\frac{1}{n}$ to the spike at the value observed. The scaling changes the proportions after the first $n-1$ observations to the proportions after the first n observations. As the sample size increases, the sample proportions get less variable. In the limit as the sample size n approaches infinity, the spike at each value approaches its probability.

Thought Experiment 3: *Number of tails before first head from independent coin tosses*

Each toss of a coin results in either a head or a tail. The probability of getting a head remains the same on each toss. The outcomes of each toss are independent of each other. This is an example of what we call Bernoulli trials. The outcome of a trial is either a success (head) or failure (tail), the probability of success remains

constant over all trials, and we are taking independent trials. We are counting the number of failures before the first success. Every nonnegative integer is a possible value, and there are an infinite number of them. They must be discrete, since there is a gap between every pair of nonnegative integers.

We start by tossing the coin and counting the number of tails until the first head occurs. Then we repeat the whole process. Eventually we reach a state where most of the time we get a value we have gotten before. After each sequence of trials until the first head, we update the proportions that have each value. We sketch a graph with a spike at each value equal to the proportion having that value. As in the previous example, the updating of the graph at step n is made by scaling all the existing spikes down by the ratio $\frac{n-1}{n}$ and adding $\frac{1}{n}$ to the spike at the value observed. The sample proportions get less variable as the sample size increases, and in the limit as n approaches infinity, the spike at each value approaches its probability.

5.2 PROBABILITY DISTRIBUTION OF A DISCRETE RANDOM VARIABLE

The proportion functions that we have seen in the three thought experiments are spike functions. They have a spike at each possible value, zero at all other values, and the sum of the spike heights equals one. In the limit as the sample size approaches infinity, the proportion of times a value occurs approaches the probability of that value, and the proportion graphs approach the probability function

$$f(y_k) = P(Y = y_k)$$

for all possible values $y_1, \ldots, y_k$ of the discrete random variable. For any other value y it equals zero.

Expected Value of a Discrete Random Variable

The expected value of a discrete random variable Y is defined to be the sum over all possible values of each possible value times its probability:

$$E(Y) = \sum_{k=1} y_k \times f(y_k)\,. \tag{5.1}$$

The expected value of a random variable is often called the mean of the random variable, and denoted μ. It is like the sample mean of an infinite sample of independent repetitions of the random variable. The sample mean of a random sample of size n repetitions of the random variable is

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i\,.$$

Here y_i is the value that occurs on the i^{th} repetition. We are summing over all

repetitions. Grouping together all repetitions that have the same possible value we get

$$\bar{y} = \sum_k \frac{n_k}{n} \times y_k \,,$$

where n_k is the number of observations that have value y_k, and we are now summing over all possible values. Note that each of the y_i (observed values) equals one of the y_k (possible values). But in the limit as n approaches ∞, the relative frequency $\frac{n_k}{n}$ approaches the probability $f(y_k)$, so the sample mean, $\bar{y}$, approaches the expected value, $E(Y)$. This shows that the expected value of a random variable is like the sample mean of an infinite size random sample of that variable.

The Variance of a Discrete Random Variable

The variance of a random variable is the expected value of square of the variable minus its mean.

$$Var(Y) = E(Y - E(Y))^2 = \sum_k (y_k - \mu)^2 \times f(y_k) \,. \tag{5.2}$$

This is like the sample variance of an infinite size random sample of that variable. We note that if we square the term in brackets, break the sum into three sums, and factor the constant terms out of each sum, we get

$$\begin{aligned} Var(Y) &= \sum_k y_k^2 \times f(y_k) - 2\mu \times \sum_k y_k f(y_k) + \mu^2 \times \sum_k f(y_k) \\ &= E(Y^2) - \mu^2 \,. \end{aligned}$$

Since $\mu = E(Y)$ this gives another useful formula for computing the variance.

$$Var(Y) = E(Y^2) - [E(Y)]^2 \,. \tag{5.3}$$

Example 6 *Let Y be a discrete random variable with probability function given in the following table.*

y_i	$f(y_i)$
0	*.20*
1	*.15*
2	*.25*
3	*.35*
4	*.05*

To find $E(Y)$ we use Equation 5.1 which gives

$$\begin{aligned} E(Y) &= 0 \times .20 + 1 \times .15 + 2 \times .25 + 3 \times .35 + 4 \times .05 \\ &= 1.90 \,. \end{aligned}$$

Note that the expected value does not have to be a possible value of the random variable Y. It represents an average. We will find $Var(Y)$ in two ways and see that they give equivalent results. First, we use the definition of variance given in Equation 5.2.

$$\begin{aligned} Var(Y) &= (0-1.90)^2 \times .20 + (1-1.90)^2 \times .15 + (2-1.90)^2 \times .25 \\ &\quad +(3-1.90)^2 \times .35 + (4-1.90)^2 \times .05 \\ &= 1.49\,. \end{aligned}$$

Second, we will use Equation 5.3. We calculate

$$\begin{aligned} E(Y^2) &= 0^2 \times .20 + 1^2 \times .15 + 2^2 \times .25 + 3^2 \times .35 + 4^2 \times .05 \\ &= 5.10\,. \end{aligned}$$

Putting that result in Equation 5.3, we get

$$\begin{aligned} Var(Y) &= 5.10 - 1.90^2 \\ &= 1.49\,. \end{aligned}$$

The Mean and Variance of a Linear Function of a Random Variable

Suppose $W = a \times Y + b$, where Y is a discrete random variable. Clearly, W is another number that is the outcome of the same random experiment that Y came from. Thus W, a linear function of a random variable Y, is another random variable. We wish to find its mean.

$$\begin{aligned} E(aY+b) &= \sum_k (ay_k + b) \times f(y_k) \\ &= \sum_k ay_k \times f(y_k) + \sum b \times f(y_k) \\ &= a \sum y_k f(y_k) + b \sum f(y_k)\,. \end{aligned}$$

Since $\sum y_k f(y_k) = \mu$ and $\sum f(y_k) = 1$, the mean of the linear function is the linear function of the mean:

$$E(aY+b) = aE(Y) + b\,. \tag{5.4}$$

Similarly we may wish to know its variance.

$$\begin{aligned} Var(aY+b) &= \sum_k (ay_k + b - E(aY+b))^2 f(y_k) \\ &= \sum_k [a(y_k - E(Y)) + b - b)]^2 f(y_k) \\ &= a^2 \sum_k (y_k - E(Y))^2 f(y_k)\,. \end{aligned}$$

Thus the variance of a linear function is the square of the multiplicative constant a times the variance :

$$Var(aY + b) = a^2 Var(Y)\,. \tag{5.5}$$

The additive constant b doesn't enter into it.

Example 6 (continued) *Suppose $W = -2Y + 3$. Then from Equation 5.4, we have*

$$\begin{aligned} E(W) &= -2E(Y) + 3 \\ &= -2 \times 1.90 + 3 \\ &= -.80 \end{aligned}$$

and from Equation 5.5, we have

$$\begin{aligned} Var(W) &= (-2)^2 \times Var(Y) \\ &= 4 \times 1.49 \\ &= 5.96\,. \end{aligned}$$

5.3 BINOMIAL DISTRIBUTION

Let us look at three situations and see what characteristics they have in common.

Coin tossing. Suppose we toss the same coin n times, and count the number of heads that occur. We consider that any one toss is not influenced by the outcomes of previous tosses, in other words, the outcome of one toss is independent of the outcomes of previous tosses. Since we are always tossing the same coin, the probability of getting a head on any particular toss remains constant for all tosses. The possible values of the total number of heads observed in the n tosses are $0, \ldots, n$.

Drawing from an urn with replacement. An urn contains balls of two colors, red and green. The proportion of red balls is π. We draw a ball at random from the urn, record its color, then return it to the urn, and remix the balls before the next random draw. We make a total of n draws, and count the number of times we drew a red ball. Since we replace and remix the balls between draws, each draw takes place under identical conditions. The outcome of any particular draw is not influenced by the previous draw outcomes. The probability of getting a red ball on any particular draw remains equal to π, the proportion of red balls in the urn. The possible values of the total number of red balls drawn are $0, \ldots, n$.

Random sampling from a very large population. Suppose we draw a random sample of size n from a very large population. The proportion of items in the population having some attribute is π. We count the number of items in the sample that have the attribute. Since the population is very large compared to the sample size, removing a few items from the population does not perceptibly change the proportion of remaining items having the attribute. For all intents and purposes it remains π. The random draws are taken under almost identical conditions. The outcome of any draw is not influenced by the previous outcomes. The possible values of the number of items drawn that have the attribute is $0, \ldots, n$.

Characteristics of the Binomial Distribution

These three cases all have the following things in common.

- There are n *independent* trials. Each trial can result either in a "success" or a "failure."
- The probability of "success" is constant over all the trials. Let π be the probability of "success."
- Y is the number of "successes" that occurred in the n trials. Y can take on integer values $0, 1, \ldots, n$.

These are the characteristics of the *binomial* (n, π) distribution. The probability function of the binomial random variable Y given the parameter value π is written as

$$f(y|\pi) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y} \tag{5.6}$$

for $y = 0, 1, \ldots, n$ where the binomial coefficient

$$\left(\begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{y! \times (n-y)!} \, .$$

Mean of binomial. The mean of the binomial(n, π) distribution is the sample size times the probability of success since

$$\begin{aligned} E(Y|\pi) &= \sum_{y=0}^{n} y \times f(y|\pi) \\ &= \sum_{y=0}^{n} y \times \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y} \, . \end{aligned}$$

We write this as a conditional mean because it is the mean of Y given the value of the parameter π. The first term in the sum is 0, so we can start the sum at $y = 1$. We cancel y in the remaining terms, and factor out $n\pi$. This gives

$$E(Y|\pi) = \sum_{y=1}^{n} n\pi \left(\begin{array}{c} n-1 \\ y-1 \end{array} \right) \pi^{y-1} (1-\pi)^{n-y} \, .$$

Factoring $n\pi$ out of the sum and substituting $n' = n - 1$ and $y' = y - 1$ we get

$$E(Y|\pi) = n\pi \sum_{y'=0}^{n'} \left(\begin{array}{c} n' \\ y' \end{array} \right) \pi^{y'} (1-\pi)^{n'-y'} \, .$$

We see the sum is a binomial probability function summed over all possible values. Hence it equals one, and the mean of the binomial is

$$E(Y|\pi) = n\pi \, . \tag{5.7}$$

Variance of binomial. The variance is the sample size times the probability of success times the probability of failure. We write this as a condtional variance since it is the variance of Y given the value of the parameter π. Note that

$$\begin{aligned} E(Y(Y-1)|\pi) &= \sum_{y=0}^{n} y(y-1) \times f(y|\pi) \\ &= \sum_{y=0}^{n} y(y-1) \times \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y} . \end{aligned}$$

The first two terms in the sum equal 0, so we can start summing at $y = 2$. We cancel $y(y-1)$ out of the remaining terms and factor out $n(n-1)\pi^2$ to get

$$E(Y(Y-1)|\pi) = \sum_{y=2}^{n} n(n-1)\pi^2 \left(\begin{array}{c} n-2 \\ y-2 \end{array} \right) \pi^{y-2} (1-\pi)^{n-y} .$$

Substituting $y' = y - 2$ and $n' = n - 2$ we get

$$\begin{aligned} E(Y(Y-1)|\pi) &= n(n-1)\pi^2 \sum_{y'=0}^{n-2} \left(\begin{array}{c} n' \\ y' \end{array} \right) \pi^{y'} (1-\pi)^{n'} \\ &= n(n-1)\pi^2 \end{aligned}$$

since we are summing a binomial distribution over all possible values. The variance can be found by

$$\begin{aligned} Var(Y|\pi) &= E(Y^2|\pi) - [E(Y|\pi)]^2 \\ &= E(Y(Y-1)|\pi) + E(Y|\pi) - [E(Y|\pi)]^2 \\ &= n(n-1)\pi^2 + n\pi - [n\pi]^2 . \end{aligned}$$

Hence the variance of the binomial is the sample size times the probability of success times the probability of failure.

$$Var(Y|\pi) = n\pi(1-\pi) . \tag{5.8}$$

5.4 HYPERGEOMETRIC DISTRIBUTION

The hypergeometric distribution models sampling from an urn without replacement. There is an urn containing N balls, R of which are red. A sequence of n balls is drawn randomly from the urn *without replacement*. Drawing a red ball is called a "success." The probability of success π does not stay constant over all the draws. At each draw the probability of "success" is the proportion of red balls remaining in the urn, which does depend on the outcomes of previous draws. Y is the number of "successes" in the n trials. Y can take on integer values $0, 1, \ldots, n$.

Probability Function of Hypergeometric

The probability function of the hypergeometric random variable Y given the parameters N, n, R is written as

$$f(y|N,R,n) = \frac{\left(\begin{array}{c} R \\ y \end{array}\right) \times \left(\begin{array}{c} N-R \\ n-y \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}$$

for possible values $y = 0, 1, \ldots, n$.

Mean and variance of hypergeometric. The conditional mean of the hypergeometric distribution is given by

$$E(Y|N,R,n) = n \times \frac{R}{N}\,.$$

The conditional variance of the hypergeometric distribution is given by

$$Var(Y|N,R,n) = n \times \frac{R}{N} \times \left(1 - \frac{R}{N})\right) \times \left(\frac{N-n}{N-1}\right)\,.$$

We note that $\frac{R}{N}$ is the proportion of red balls in the urn. The mean and variance of the hypergeometric are similar to that of the binomial, except that the variance is smaller due to the finite population correction factor $\frac{N-n}{N-1}$.

5.5 JOINT RANDOM VARIABLES

When two (or more) numbers are determined from the outcome of a random experiment, we call it a joint experiment. The two numbers are called joint random variables and denoted X, Y. If both the random variables are discrete, they each have separated possible values x_i for $i = 1, \ldots, I$, and y_j for $j = 1, \ldots, J$. The *universe* for the experiment is the set of all possible outcomes of the experiment which are all possible ordered pairs of possible values. The *universe* of the joint experiment is shown in Table 5.2.

The joint probability function of two discrete joint random variables is defined at each point in the universe:

$$f(x_i, y_j) = P(X = x_i, Y = y_j)$$

for $i = 1, \ldots, I$, and $j = 1, \ldots, J$. This is the probability that $X = x_i$ and $Y = y_j$ simultaneously, in other words, the probability of the intersection of the events $X = x_i$ and $Y = y_j$. These joint probabilities can be put in a table.

We might want to consider the probability distribution of just one of the joint random variables, for instance, Y. The event $Y = y_j$ for some fixed value y_j is the

Table 5.2 Universe of joint experiment

(x_1, y_1)	.	.	.	(x_1, y_j)	.	.	.	(x_1, y_J)
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
(x_i, y_1)	.	.	.	(x_i, y_j)	.	.	.	(x_i, y_J)
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
(x_I, y_1)	.	.	.	(x_I, y_j)	.	.	.	(x_I, y_J)

union of all events $X = x_i, Y = y_j$, where $i = 1, \ldots, I$, and they are all disjoint. Thus

$$P(Y = y_j) = P(\cup_i (X = x_i, Y = y_j)) = \sum_i P(X = x_i, Y = y_j)$$

for $j = 1, \ldots, J$, since probability is additive over a disjoint union. This probability distribution of Y by itself is called the *marginal* distribution of Y. Putting this relationship in terms of the probability function we get

$$f(y_j) = \sum_i f(x_i, y_j) \tag{5.9}$$

for $j = 1, \ldots J$. So we see that the individual probabilities of Y is found by summing the joint probabilities down the columns. Similarly the individual probabilities of X can be found by summing the joint probabilities across the rows. We can write them on the margins of the table, hence the names *marginal* probability distribution of Y and X respectively. The joint probability distribution and the marginal probability distributions are shown in Table 5.3. The joint probabilities are in the main body of the table, and the marginal probabilities for X and Y are in the right column and bottom row, respectively.

The expected value of a function of the joint random variables is given by

$$E(h(X, Y)) = \sum_i \sum_j h(x_i, y_j) \times f(x_i, y_j)\,.$$

Often we wish to find the expected value of a sum of random variables. In that case

$$\begin{aligned} E(X + Y) &= \sum_i \sum_j (x_i + y_j) \times f(x_i, y_j) \\ &= \sum_i \sum_j x_i \times f(x_i, y_j) + \sum_i \sum_j y_j \times f(x_i, y_j) \end{aligned}$$

Table 5.3 Joint and marginal probability distributions

	y_1	.	.	.	y_j	.	.	.	y_J	
x_1	$f(x_1, y_1)$	.	.	.	$f(x_1, y_j)$	.	.	.	$f(x_1, y_J)$	$f(x_1)$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
x_i	$f(x_i, y_1)$	.	.	.	$f(x_i, y_j)$	.	.	.	$f(x_i, y_J)$	$f(x_i)$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
x_I	$f(x_I, y_1)$	.	.	.	$f(x_I, y_j)$	.	.	.	$f(x_I, y_J)$	$f(x_I)$
	$f(y_1)$	.	.	.	$f(y_j)$	.	.	.	$f(y_J)$	

$$\begin{aligned} &= \sum_i x_i \sum_j f(x_i, y_j) + \sum_j y_j \sum_i f(x_i, y_j) \\ &= \sum_i x_i \times f(x_i) + \sum_j y_j \times f(y_j)\,. \end{aligned}$$

We see the mean of the sum of two random variables is the sum of the means.

$$E(X+Y) = E(X) + E(Y)\,. \tag{5.10}$$

This equation always holds.

Independent Random Variables

Two (discrete) random variables X and Y are independent of each other if and only if every element in the joint distribution table equals the product of the corresponding marginal distributions. In other words,

$$f(x_i, y_j) = f(x_i) \times f(y_j)$$

for all possible x_i and y_j.
The variance of a sum of random variables is given by

$$\begin{aligned} Var(X+Y) &= E(X+Y-E(X+Y))^2 \\ &= \sum_i \sum_j (x_i + y_j - (E(X)+E(Y))^2 \times f(x_i, y_j) \\ &= \sum_i \sum_j [(x_i - E(X)) + (y_j - E(Y))]^2 \times f(x_i, y_j)\,. \end{aligned}$$

Multiplying this out and breaking it into three separate sums gives

$$\begin{aligned} Var(X+Y) &= \sum_i \sum_j (x_i - E(X))^2 \times f(x_i, y_j) \\ &\quad + \sum_i \sum_j 2(x_i - E(X))(y_j - E(Y)) f(x_i, y_j) \\ &\quad + \sum_i \sum_j (y_j - E(Y))^2 \times f(x_i, y_j)\,. \end{aligned}$$

The middle term is 2 $\times$ the covariance of the random variables. For independent random variables the covariance

$$\begin{aligned} Cov(X,Y) &= \sum_i \sum_j (x_i - E(X)) \times (y_j - E(Y)) f(x_i, y_j) \\ &= \sum_i (x_i - E(X)) f(x_i) \times \sum_j (y_j - E(Y)) f(y_j)\,. \end{aligned}$$

This is clearly equal to 0. Hence for independent random variables

$$Var(X+Y) = \sum_i (x_i - E(X)))^2 \times f(x_i) + \sum_j (y_j - E(Y))^2 \times f(y_j)\,.$$

We see the variance of the sum of two independent random variables is the sum of the variances.

$$Var(X+Y) = Var(X) + Var(Y)\,. \tag{5.11}$$

This equation only holds for independent[1] random variables!

Example 7 *Let X and Y be jointly distributed discrete random variables. Their joint probability distribution is given in the following table:*

		Y				$f(x)$
		1	*2*	*3*	*4*	
	1	*.02*	*.04*	*.06*	*.08*	
X	*2*	*.03*	*.01*	*.09*	*.17*	
	3	*.05*	*.15*	*.15*	*.15*	
	$f(y)$					

We find the marginal distributions of X and Y by summing across the rows and summing down the columns respectively. That gives the table

[1] In general, the variance of a sum of two random variables is given by $Var(X+Y) = Var(X) + 2 \times Cov(X,Y) + Var(Y)$.

		Y				$f(x)$
		1	*2*	*3*	*4*	
	1	*.02*	*.04*	*.06*	*.08*	*.2*
X	*2*	*.03*	*.01*	*.09*	*.17*	*.3*
	3	*.05*	*.15*	*.15*	*.15*	*.5*
	$f(y)$	*.1*	*.2*	*.3*	*.4*	

We see that the joint probability $f(x_i, y_j)$ is not always equal to the product of the marginal probabilities $f(x_i) \times f(y_j)$. Therefore the two random variables X and Y are not independent.

Mean and variance of a difference between two independent random variables. When we combine the results of Equations 5.10 and 5.11 with the results of Equations 5.4 and 5.5, we find the that mean of a difference between random variables is

$$E(X - Y) = E(X) - E(Y)\,. \tag{5.12}$$

If the two random variables are independent, we find that the variance of their difference is

$$Var(X - Y) = Var(X) + Var(Y)\,. \tag{5.13}$$

Variability always adds for independent random variables, regardless of whether we are taking the sum or taking the difference.

5.6 CONDITIONAL PROBABILITY FOR JOINT RANDOM VARIABLES

If we are given $Y = y_j$, the reduced universe is the set of ordered pairs where the second element is y_j. This is shown in Table 5.4. It is the only part of the universe that remains, given $Y = y_j$. The only part of the event $X = x_i$ that remains is the part in the reduced universe. This is the intersection of the events $X = x_i$ and $Y = y_j$. Table 5.5 shows the original joint probability function in the reduced universe, and the marginal probability. We see that this is not a probability distribution. The sum of the probabilities in the reduced universe sums to the marginal probability, not to one!

The conditional probability that random variable $X = x_i$, given $Y = y_j$ is the probability of the intersection of the events $X = x_i$ and $Y = y_j$ divided by the probability that $Y = y_j$ from Equation 4.1. Dividing the joint probability by the marginal probability scales it up so the probability of the reduced universe equals 1. The conditional probability is given by

$$f(x_i|y_j) = P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}\,. \tag{5.14}$$

When we put this in terms of the joint and marginal probability functions we get

$$f(x_i|y_j) = \frac{f(x_i, y_j)}{f(y_j)}\,. \tag{5.15}$$

Table 5.4 Reduced universe given $Y = y_j$

.	.	.	.	(x_1, y_j)	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	(x_i, y_j)	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	(x_I, y_j)	.	.	.	.

Table 5.5 Joint probability function values in the reduced universe $Y = y_j$. The marginal probability is found by summing down the column.

.	.	.	.	$f(x_1, y_j)$	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	$f(x_i, y_j)$	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	$f(x_I, y_j)$	.	.	.	.
.	.	.	.	$f(y_j)$	.	.	.	.

The conditional probability distribution. Letting x_i vary across all possible values of X gives us the conditional probability distribution of $X|Y = y_j$. The conditional probability distribution is defined on the reduced universe given $Y = y_j$. The conditional probability distribution is shown in Table 5.6. Each entry was found by dividing the *i,j* entry in the joint probability table by j^{th} element in the marginal probability. The marginal probability $f(y_j) = \sum_i f(x_i, y_j)$, and is found by summing down the j^{th} column of the joint probability table. So the conditional probability of x_i given y_j is the j^{th} column in the joint probability table, divided by the sum of the joint probabilities in the j^{th} column.

Table 5.6 The conditional probability function defined on the reduced universe $Y = y_j$

.	.	.	.	$f(x_1\|y_j)$	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	$f(x_i\|y_j)$	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	$f(x_I\|y_j)$	.	.	.	.

Example 7 (continued) *If we want to determine the conditional probability* $P(X = 2|Y = 2)$ *we plug in the joint and marginal probabilities into Equation 5.14. This gives*

$$P(X = 2|Y = 2) = \frac{P(X = 2, Y = 2)}{P(Y = 2)} = \frac{.01}{.2} = .05\,.$$

Conditional probability as multiplication rule. Using similar arguments, we could find that the conditional probability function of Y given $X = x_i$ is given by

$$f(y_j|x_i) = \frac{f(x_i, y_j)}{f(x_i)}\,.$$

However, we will not use the relationship in this form, since we do not consider the random variables interchangeably. In Bayesian statistics, the random variable X is the unobservable parameter. The random variable Y is an observable random variable that has a probability distribution depending on the parameter. In the next chapter we will use the conditional probability relationship as the multiplication rule

$$f(x_i, y_j) = f(x_i) \times f(y_j|x_i) \tag{5.16}$$

when we develop Bayes' theorem for discrete random variables.

Main Points

- A random variable Y is a number associated with the outcome of a random experiment.
- If the only possible values of the random variable are a finite set of separated values, $y_1, \ldots, y_K$ the random variable is said to be discrete.

- The probability distribution of the discrete random variable gives the probability associated with each possible value.
- The probability of any event associated with the random experiment can be calculated from the probability function of the random variable using the laws of probability.
- The expected value of a discrete random variable is

$$E(Y) \quad = \quad \sum_k y_k f(y_k)\,,$$

  where the sum is over all possible values of the random variable. It is the mean of the distribution of the random variable.
- The variance of a discrete random variable is the expected value of the squared deviation of the random variable from its mean.

$$Var(Y) \;=\; E(Y - E(Y))^2 \;=\; \sum_k (y_k - E(Y))^2 f(y_k)\,.$$

  Another formula for the variance is

$$Var(Y) \;=\; E(Y^2) - [E(Y)]^2\,.$$

- The mean and variance of a linear function of a random variable $aY + b$ are

$$E(aY + b) \;=\; aE(Y) + b$$

  and

$$Var(aY + b) \;=\; a^2 \times Var(Y)\,.$$

- The *binomial* (n, π) distribution models the number of successes in n independent trials where each trial has the same success probability, π.
- The joint probability distribution of two discrete random variables X and Y is written as joint probability function

$$f(x_i, y_j) \quad = \quad P(X = x_i, Y = y_j)\,.$$

  Note: $(X = x_i, Y = y_j)$ is another way of writing the intersection $(X = x_i \cap Y = y_j)$. This joint probability function can be put in a table.
- The marginal probability distribution of one of the random variables can be found by summing the joint probability distribution across rows (for X) or by summing down columns (for Y).
- The mean and variance of a sum of independent random variables are

$$E(X + Y) \;=\; E(X) + E(Y)$$

and

$$Var(X+Y) \;=\; Var(X) + Var(Y)\,.$$

- The mean and variance of a difference between independent random variables are

$$E(X-Y) \;=\; E(X) - E(Y)$$

and

$$Var(X-Y) \;=\; Var(X) + Var(Y)\,.$$

- Conditional probability function of X given $Y = y_j$ is found by

$$f(x_i|y_j) \;=\; \frac{f(x_i, y_j)}{f(y_j)}\,.$$

  This is the joint probability divided by the marginal probability that $Y = y_j$

- The joint probabilities on the reduced universe $Y = y_j$ are not a probability distribution. They sum to the marginal probability $f(y_j)$, not to one.

- Dividing the joint probabilities by the marginal probability scales up the probabilities, so the sum of probabilities in the reduced universe is one.

Exercises

5.1 A discrete random variable Y has discrete distribution given in the following table:

y_i	$f(y_i)$
0	.2
1	.3
2	.3
3	.1
4	.1

(a) Calculate $P(1 < Y \leq 3)$.

(b) Calculate $E(Y)$.

(c) Calculate $Var(Y)$.

(d) Let $W = 2Y + 3$. Calculate $E(W)$.

(e) Calculate $Var(W)$.

5.2 A discrete random variable Y has discrete distribution given in the following table:

y_i	$f(y_i)$
0	.1
1	.2
2	.3
5	.4

(a) Calculate $P(0 < Y < 2)$.

(b) Calculate $E(Y)$.

(c) Calculate $Var(Y)$.

(d) Let $W = 3Y - 1$. Calculate $E(W)$.

(e) Calculate $Var(W)$.

5.3 Let Y be *binomial* $(n = 5, \pi = .6)$.

(a) Calculate the mean and variance by filling in the following table:

y_i	$f(y_i)$	$y_i \times f(y_i)$	$y_i^2 \times f(y_i)$
0			
1			
2			
3			
4			
5			
Sum			

i. $E(Y) =$

ii. $Var(Y) =$

(b) Calculate the mean and variance of Y using Equations 5.7 and 5.8, respectively. Do you get the same results as in part (a)?

5.4 Let Y be *binomial* $(n = 4, \pi = .3)$.

(a) Calculate the mean and variance by filling in the following table:

y_i	$f(y_i)$	$y_i \times f(y_i)$	$y_i^2 \times f(y_i)$
0			
1			
2			
3			
4			
Sum			

i. $E(Y) =$
ii. $Var(Y) =$

(b) Calculate the mean and variance of Y using Equations 5.7 and 5.8, respectively. Do you get the same as you got in part (a)?

5.5 Let X and Y be jointly distributed discrete random variables. Their joint probability distribution is given in the following table:

		Y					$f(x)$
		1	2	3	4	5	
	1	.02	.04	.06	.08	.05	
X	2	.08	.02	.10	.02	.03	
	3	.05	.05	.03	.02	.10	
	4	.10	.04	.05	.03	.03	
	$f(y)$						

(a) Calculate the marginal probability distribution of X.

(b) Calculate the marginal probability distribution of Y.

(c) Are X and Y independent random variables? Explain why or why not.

(d) Calculate the conditional probability $P(X = 3|Y = 1)$.

5.6 Let X and Y be jointly distributed discrete random variables. Their joint probability distribution is given in the following table:

		Y					$f(x)$
		1	2	3	4	5	
	1	.015	.030	.010	.020	.025	
X	2	.030	.060	.020	.040	.050	
	3	.045	.090	.030	.060	.075	
	4	.060	.120	.040	.080	.100	
	$f(y)$						

(a) Calculate the marginal probability distribution of X.

(b) Calculate the marginal probability distribution of Y.

(c) Are X and Y independent random variables? Explain why or why not.

(d) Calculate the conditional probability $P(X = 2|Y = 3)$.

6

Bayesian Inference for Discrete Random Variables

In this chapter we introduce Bayes' theorem for discrete random variables. Then we see how we can use it to revise our beliefs about the parameter, given the sample data that depends on the parameter. This is how we will perform statistical inference in a Bayesian manner.

We will consider the parameter to be random variable X, which has possible values $x_1, \ldots, x_I$. We never observe the parameter random variable. The random variable Y, which depends on the parameter, has possible values $y_1, \ldots, y_J$. We make inferences about the parameter random variable X given the observed value $Y = y_j$ using Bayes' theorem.

The *Bayesian universe* consists of the all possible ordered pairs (x_i, y_j) for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. This is analogous to the universe we used for joint random variables in the last chapter. However, we will not consider the random variables X and Y the same way. The events $(X = x_1), \ldots, (X = x_I)$ partition the universe, but we never observe which one has occurred. The event $Y = y_j$ is observed.

We know that the Bayesian universe has two dimensions, the horizontal dimension which is observable, and the vertical dimension which is unobservable. In the horizontal direction it goes across the sample space which is the set of all possible values, $\{y_1, \ldots, y_J\}$, of the observed random variable Y. In the vertical direction it goes through the parameter space, which is the set of all possible parameter values, $\{x_1, \ldots, x_I\}$. The Bayesian universe for discrete random variables is shown in Table 6.1. This is analogous the Bayesian universe for events described in Chapter 4. The

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

Table 6.1 The Bayesian universe

(x_1, y_1)	(x_1, y_2)	. . .	(x_1, y_j)	. . .	(x_1, y_J)
.	.	. . .	.	. . .	.
.	.	. . .	.	. . .	.
.	.	. . .	.	. . .	.
(x_i, y_1)	(x_i, y_2)	. . .	(x_i, y_j)	. . .	(x_i, y_J)
.	.	. . .	.	. . .	.
.	.	. . .	.	. . .	.
.	.	. . .	.	. . .	.
(x_I, y_1)	(x_1, y_2)	. . .	(x_I, y_j)	. . .	(x_I, y_J)

parameter value is unobserved. Probabilities are defined at all points in the Bayesian universe.

We will change our notation slightly. We will use $f()$ to denote a probability distribution (conditional or unconditional) that contains the observable random variable Y, and $g()$ to denote a probability distribution (conditional or unconditional) that only contains the (unobserved) parameter random variable X. This clarifies the distinction between Y, the random variable that we will observe, and X, the unobserved parameter random variable that we want to make our inference about. Each of the joint probabilities in the Bayesian universe is found using the multiplication rule

$$f(x_i, y_j) = g(x_i) \times f(y_j|x_i)\,.$$

The marginal distribution of Y is found by summing the columns. We show the joint and marginal probability function in Table 6.2. Note that this is similar to how we presented the joint and marginal distribution for two discrete random variables in the previous chapter (Table 5.3). However, now we have moved the marginal probability function of X over to the left-hand side and call it the *prior* probability function of the parameter X to indicate it is known to us at the beginning. We also note the changed notation.

When we observe $Y = y_j$, the reduced Bayesian universe is the set of ordered pairs in the j^{th} column. This is shown in Table 6.3. The posterior probability function of X given $Y = y_j$ is given by

$$g(x_i|y_j) = \frac{g(x_i) \times f(y_j|x_i)}{\sum_{i=1}^{n_i} g(x_i) \times f(y_j|x_i)}\,.$$

Let us look at the parts of the formula.

- The prior distribution of the discrete random variable X is given by the *prior* probability function $g(x_i)$, for $i = 1, \ldots, n$. This is what we believe the probability of each x_i to be before we look at the data. It must come from prior experience, not from the current data.

Table 6.2 The joint and marginal distributions of X and Y

	$prior$	y_1	. . .	y_j	. . .	y_J
x_1	$g(x_1)$	$f(x_1, y_1)$	. . .	$f(x_1, y_j)$	. . .	$f(x_1, y_J)$
.	.	.	. . .	.	. . .	.
.	.	.	. . .	.	. . .	.
.	.	.	. . .	.	. . .	.
x_i	$g(x_i)$	$f(x_i, y_1)$	. . .	$f(x_i, y_j)$	. . .	$f(x_i, y_J)$
.	.	.	. . .	.	. . .	.
.	.	.	. . .	.	. . .	.
.	.	.	. . .	.	. . .	.
x_I	$g(x_I)$	$f(x_I, y_1)$	. . .	$f(x_I, y_j)$	. . .	$f(x_I, y_J)$
		$f(y_1)$	. . .	$f(y_j)$	. . .	$f(y_J)$

Table 6.3 The reduced Bayesian universe given $Y = y_j$

.	.	.	.	(x_1, y_j)	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	(x_i, y_j)	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	(x_I, y_j)	.	.	.	.

- Since we observed $Y = y_j$, the likelihood of the discrete parameter random variable is given by the *likelihood function* $f(y_j|x_i)$ for $i = 1, \ldots, n$. This is the conditional probability function of Y given $X = x_i$ evaluated at y_j, the value that actually occurred and where X is allowed to vary over its whole range for $x_i, \ldots, x_n$. We must know the form of the conditional observation distribution as it shows how the distribution of the observation Y depends on the value of the random variable X, but we see that it only needs to be evaluated at the value that actually occurred, y_j. The likelihood function is the conditional observation distribution evaluated on the reduced universe.

- The posterior probability distribution of the discrete random variable is given by the *posterior* probability function $g(x_i|y_j)$ evaluated at x_i for $i = 1, \ldots, n$, given $Y = y_j$

The formula gives us a method for revising our belief probabilities about the possible values of X given that we observed $Y = y_j$.

Example 8 *There is an urn containing a total of 5 balls, some of which are red and the rest of which are green. We don't know how many of the balls are red. Let the random variable X be the number of red balls in the urn. Possible values of X are $x_i = i$ for $i = 0, \ldots, 5$. Since we don't have any idea about the number of red balls, we will assume all possible values are equally likely. Our prior distribution of X is $g(0) = g(1) = g(2) = g(3) = g(4) = g(5) = 1/6$*

We will draw a ball at random from the urn. The random variable Y=1 if draw is red, 0 otherwise. Conditional observation distribution of $Y|X$ is $P(Y = 1|X = x_i) = i/5$ and $P(Y = 0|X = x_i) = (5 - i)/5$. The joint probabilities are found by multiplying the prior probabilities times the conditional observation probabilities. The marginal probabilities of Y are found by summing the joint probabilities down the columns. These are shown in Table 6.4.

Suppose the selected ball is red, so the reduced universe is in the column labelled $y_j = 1$. The conditional observation probabilities in that column are highlighted. They form the likelihood function. Table 6.5 shows the steps for finding the posterior distribution of X given $Y = 1$.

Notice that the only column that was used to find the posterior probability distribution was the in the reduced universe, the column $Y = 1$. The joint probability came from multiplying the prior probabilities times the likelihood function. The posterior probability equals the prior probability times likelihood divided by the sum of prior probabilities times likelihoods:

$$f(x_i|y_j) = P(X = x_i|Y = y_j) = \frac{g(x_i) \times f(y_j|x_i)}{\sum_{i=1}^{n_i} g(x_i) \times f(y_j|x_i)} .$$

Thus a simpler way of finding the posterior probability is to use only the column in the reduced universe. Its probability is product of the prior times the likelihood. This is shown in Table 6.6.

Steps for Bayes' Theorem Using Table

- Set up a table with columns for *parameter value, prior, likelihood, prior* $\times$ *likelihood* and *posterior*.
- Put in the *parameter values*, the *prior*, and the *likelihood* in their respective columns.
- Multiply each element in the *prior* column by the corresponding element in the *likelihood* column and put the results in the *prior* $\times$ *likelihood* column.
- Sum the *prior* $\times$ *likelihood* column.
- Divide each element of *prior* $\times$ *likelihood* column by the sum.
- Put these posterior probabilities in the *posterior* column!

Table 6.4 The joint and marginal probability distributions

x_i	$prior$ $probability$	$y_j = 0$	$y_j = 1$
0	1/6	$\frac{1}{6} \times \frac{5}{5} = \frac{5}{30}$	$\frac{1}{6} \times \frac{0}{5}$ =0
1	1/6	$\frac{1}{6} \times \frac{4}{5} = \frac{4}{30}$	$\frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$
2	1/6	$\frac{1}{6} \times \frac{3}{5} = \frac{3}{30}$	$\frac{1}{6} \times \frac{2}{5} = \frac{2}{30}$
3	1/6	$\frac{1}{6} \times \frac{2}{5} = \frac{2}{30}$	$\frac{1}{6} \times \frac{3}{5} = \frac{3}{30}$
4	1/6	$\frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$	$\frac{1}{6} \times \frac{4}{5} = \frac{4}{30}$
5	1/6	$\frac{1}{6} \times \frac{0}{5} = \frac{0}{30}$	$\frac{1}{6} \times \frac{5}{5} = \frac{5}{30}$
$f(y_j)$		$\frac{15}{30}$	$\mathbf{\frac{15}{30} = \frac{1}{2}}$

Table 6.5 Finding the posterior probabilities of $X|Y = 1$

x_i	$prior$ $probability$	$y_j = 0$	$y_j = 1$	$posterior$ $probability$
0	1/6	$\frac{1}{6} \times \frac{5}{5} = \frac{5}{30}$	$\frac{1}{6} \times \mathbf{\frac{0}{5}}$ =0	0
1	1/6	$\frac{1}{6} \times \frac{4}{5} = \frac{4}{30}$	$\frac{1}{6} \times \mathbf{\frac{1}{5}} = \frac{1}{30}$	$\frac{1}{30}/\frac{1}{2} = \frac{1}{15}$
2	1/6	$\frac{1}{6} \times \frac{3}{5} = \frac{3}{30}$	$\frac{1}{6} \times \mathbf{\frac{2}{5}} = \frac{2}{30}$	$\frac{2}{30}/\frac{1}{2} = \frac{2}{15}$
3	1/6	$\frac{1}{6} \times \frac{2}{5} = \frac{2}{30}$	$\frac{1}{6} \times \mathbf{\frac{3}{5}} = \frac{3}{30}$	$\frac{3}{30}/\frac{1}{2} = \frac{3}{15}$
4	1/6	$\frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$	$\frac{1}{6} \times \mathbf{\frac{4}{5}} = \frac{4}{30}$	$\frac{4}{30}/\frac{1}{2} = \frac{4}{15}$
5	1/6	$\frac{1}{6} \times \frac{0}{5} = \frac{0}{30}$	$\frac{1}{6} \times \mathbf{\frac{5}{5}} = \frac{5}{30}$	$\frac{5}{30}/\frac{1}{2} = \frac{5}{15}$
$f(y_j)$		$\frac{15}{30}$	$\mathbf{\frac{15}{30} = \frac{1}{2}}$	

Table 6.6 Simplified table for finding the posterior probabilities of $X|Y = 1$

x_i	$prior$	$likelihood$	$prior \times likelihood$	$posterior$
0	1/6	$\frac{0}{5}$	$\frac{1}{6} \times \frac{0}{5}$ =0	0
1	1/6	$\frac{1}{5}$	$\frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$	$\frac{1}{30}/\frac{1}{2} = \frac{1}{15}$
2	1/6	$\frac{2}{5}$	$\frac{1}{6} \times \frac{2}{5} = \frac{2}{30}$	$\frac{2}{30}/\frac{1}{2} = \frac{2}{15}$
3	1/6	$\frac{3}{5}$	$\frac{1}{6} \times \frac{3}{5} = \frac{3}{30}$	$\frac{3}{30}/\frac{1}{2} = \frac{3}{15}$
4	1/6	$\frac{4}{5}$	$\frac{1}{6} \times \frac{4}{5} = \frac{4}{30}$	$\frac{4}{30}/\frac{1}{2} = \frac{4}{15}$
5	1/6	$\frac{5}{5}$	$\frac{1}{6} \times \frac{5}{5} = \frac{5}{30}$	$\frac{5}{30}/\frac{1}{2} = \frac{5}{15}$
$f(y_j)$			$\frac{15}{30} = \frac{1}{2}$	

Table 6.7 The posterior probability distribution after second observation

x_i	$prior$	$likelihood$	$prior \times likelihood$	$posterior$
0	0	??	0	$0/\frac{1}{3} = 0$
1	1/15	$\frac{4}{4}$	$\frac{1}{15}$	$\frac{1}{15}/\frac{1}{3} = \frac{1}{5}$
2	2/15	$\frac{3}{4}$	$\frac{1}{10}$	$\frac{1}{10}/\frac{1}{3} = \frac{6}{20}$
3	3/15	$\frac{2}{4}$	$\frac{1}{10}$	$\frac{1}{10}/\frac{1}{3} = \frac{6}{20}$
4	4/15	$\frac{1}{4}$	$\frac{1}{15}$	$\frac{1}{15}/\frac{1}{3} = \frac{1}{5}$
5	5/15	$\frac{0}{4}$	0	$0/\frac{1}{3} = 0$
			$\frac{1}{3}$	1.00

6.1 TWO EQUIVALENT WAYS OF USING BAYES' THEOREM

We may have more than one data set concerning a parameter. They might not even become available at the same time. Should we wait for the second data set, combine it with the first, and then use Bayes' theorem on the combined data set? This would mean that we have to go back to scratch every time more data became available, which would result in a lot of work. Another approach requiring less work would be to use the posterior probabilities given the first data set, as the prior probabilities for analyzing the second data set. We will find that these two approaches lead to the same posterior probabilities. This is a significant advantage to Bayesian methods. In frequentist statistics, we would have to use the first approach, re-analyzing the combined data set when the second one arrives.

Analyzing the observations in sequence. Suppose that we randomly draw a second ball out of the urn without replacing the first. Suppose the second draw resulted in a green ball, so $Y = 0$. We want to find the posterior probabilities of X given the results of the two observations, red first, green second. We will analyze the observations in sequence using Bayes' theorem each time. We will use the same prior probabilities as before for the first draw. However, we will use the posterior probabilities from the first draw as the prior probabilities for the second draw. The results are shown in Table 6.7.

Analyzing the observations all together. Alternatively, we could consider both draws together, then revise the probabilities using Bayes' theorem only once. Initially, we are in the same state of knowledge as before. So we take the same prior probabilities that we originally used for the first draw when we were analyzing the observations in sequence. All possible values of X are equally likely. The prior probability function is $g(x) = \frac{1}{6}$ for $x = 0, \ldots, 5$.

Let Y_1 and Y_2 be the outcome of the first and second draw, respectively. The probabilities of the second draw depend on the balls left after the first draw. By the multiplication rule, the observation probability conditional on X is

$$f(y_1, y_2|x) = f(y_1|x) \times f(y_2|y_1, x)\,.$$

Table 6.8 The joint distribution of X, Y_1, Y_2 and marginal distribution of Y_1, Y_2

x_i	$prior$	y_{j_1}, y_{j_2} 0,0	y_{j_1}, y_{j_2} 0,1	y_{j_1}, y_{j_2} 1,0	y_{j_1}, y_{j_2} 1,1
0	1/6	$\frac{1}{6} \times \frac{5}{5} \times \frac{4}{4}$	$\frac{1}{6} \times \mathbf{\frac{5}{5}} \times \mathbf{\frac{4}{4}}$	$\frac{1}{6} \times \frac{0}{5} \times \frac{4}{4}$	$\frac{1}{6} \times \frac{0}{5} \times \frac{4}{4}$
1	1/6	$\frac{1}{6} \times \frac{4}{5} \times \frac{3}{4}$	$\frac{1}{6} \times \mathbf{\frac{4}{5}} \times \mathbf{\frac{1}{4}}$	$\frac{1}{6} \times \frac{1}{5} \times \frac{4}{4}$	$\frac{1}{6} \times \frac{1}{5} \times \frac{0}{4}$
2	1/6	$\frac{1}{6} \times \frac{3}{5} \times \frac{2}{4}$	$\frac{1}{6} \times \mathbf{\frac{3}{5}} \times \mathbf{\frac{2}{4}}$	$\frac{1}{6} \times \frac{2}{5} \times \frac{3}{4}$	$\frac{1}{6} \times \frac{2}{5} \times \frac{1}{4}$
3	1/6	$\frac{1}{6} \times \frac{2}{5} \times \frac{1}{4}$	$\frac{1}{6} \times \mathbf{\frac{2}{5}} \times \mathbf{\frac{3}{4}}$	$\frac{1}{6} \times \frac{3}{5} \times \frac{2}{4}$	$\frac{1}{6} \times \frac{3}{5} \times \frac{2}{4}$
4	1/6	$\frac{1}{6} \times \frac{1}{5} \times \frac{0}{4}$	$\frac{1}{6} \times \mathbf{\frac{1}{5}} \times \mathbf{\frac{4}{4}}$	$\frac{1}{6} \times \frac{4}{5} \times \frac{1}{4}$	$\frac{1}{6} \times \frac{4}{5} \times \frac{3}{4}$
5	1/6	$\frac{1}{6} \times \frac{0}{5} \times \frac{0}{4}$	$\frac{1}{6} \times \mathbf{\frac{0}{5}} \times \mathbf{\frac{4}{4}}$	$\frac{1}{6} \times \frac{5}{5} \times \frac{0}{4}$	$\frac{1}{6} \times \frac{5}{5} \times \frac{4}{4}$
	$f(y_1, y_2)$	40/120	20/120	20/120	40/120

Table 6.9 The posterior probability distribution given $Y_1 = 1$ and $Y_2 = 0$

x_i	$prior$	y_{j_1}, y_{j_2} 0,0	y_{j_1}, y_{j_2} 0,1	y_{j_1}, y_{j_2} 1,0	y_{j_1}, y_{j_2} 1,1	$posterior$	
0	1/6	$\frac{20}{120}$	0	$\mathbf{0}$	0	0	=0
1	1/6	$\frac{12}{120}$	$\frac{4}{120}$	$\mathbf{\frac{4}{120}}$	0	$\frac{4}{120} / \frac{20}{120}$	$=\frac{1}{5}$
2	1/6	$\frac{6}{120}$	$\frac{6}{120}$	$\mathbf{\frac{6}{120}}$	$\frac{2}{120}$	$\frac{6}{120} / \frac{20}{120}$	$=\frac{3}{10}$
3	1/6	$\frac{2}{120}$	$\frac{6}{120}$	$\mathbf{\frac{6}{120}}$	$\frac{6}{120}$	$\frac{6}{120} / \frac{20}{120}$	$= \frac{3}{10}$
4	1/6	0	$\frac{4}{120}$	$\mathbf{\frac{4}{120}}$	$\frac{12}{120}$	$\frac{4}{120} / \frac{20}{120}$	$= \frac{1}{5}$
5	1/6	0	0	$\mathbf{0}$	$\frac{20}{120}$	0	$= 0$
	$f(y_1, y_2)$			20/120			1.00

The joint distribution of X and Y_1, Y_2 is given in Table 6.8. The first ball was red, second was green, so the reduced universe probabilities are in column $y_{j_1}, y_{j_2} = 1, 0$. The likelihood function given by the conditional observation probabilities in that column are highlighted.

The first ball was red, second was green, so the reduced universe probabilities are in column $y_{j_1}, y_{j_2} = 1, 0$. The posterior probability of X given $Y_1 = 1$ and $Y_2 = 0$ is found by rescaling the probabilities in the reduced universe so they sum to 1. This is shown in Table 6.9.

We see this is the same as the posterior probabilities we found analyzing the observations sequentially, using the posterior after the first as the prior for the second. This shows that it makes no difference whether you analyze the observations one at a time in sequence using the posterior after the previous step as the prior for the next step, or whether you analyze all observations together in a single step starting with your initial prior!

Table 6.10 The posterior probability distribution after both observations

x_i	$prior$	$likelihood$	$prior \times likelihood$	$posterior$	
0	1/6	$\frac{0}{20}$	$\frac{0}{120}$	$\frac{0}{120}/\frac{1}{6}$	=0
1	1/6	$\frac{4}{20}$	$\frac{4}{120}$	$\frac{4}{120}/\frac{1}{6}$	$=\frac{1}{5}$
2	1/6	$\frac{6}{20}$	$\frac{6}{120}$	$\frac{6}{120}/\frac{1}{6}$	$=\frac{3}{10}$
3	1/6	$\frac{6}{20}$	$\frac{6}{120}$	$\frac{6}{120}/\frac{1}{6}$	$=\frac{3}{10}$
4	1/6	$\frac{4}{20}$	$\frac{4}{120}$	$\frac{4}{120}/\frac{1}{6}$	$=\frac{1}{5}$
5	1/6	$\frac{0}{20}$	$\frac{0}{120}$	$\frac{0}{120}/\frac{1}{6}$	$=0$
			$\frac{1}{6}$		1.00

Since we only use the column corresponding to the reduced universe, it is simpler to finding the posterior by multiplying prior times likelihood and rescaling to make it a probability distribution. This is shown in Table 6.10

6.2 BAYES' THEOREM FOR BINOMIAL WITH DISCRETE PRIOR

We will look at using Bayes' theorem when the observation comes from the binomial distribution, and there are only a few possible values for the parameter. $Y|\pi$ has the binomial n, π distribution. (There are n independent trials, each of which can result in "success" or "failure" and the probability of success π remains the same for all trials. Y is the total number of "successes" over the n trials.) There are I discrete possible values of $\pi_1, \ldots, \pi_I$.

Set up a table for the observation distributions. Row i correspond to the binomial n, π_i probability distribution. Column j corresponds to $Y = j$ (There are $n + 1$ columns corresponding to $0, \ldots, n$.) These binomial probabilities can be found in Table B.1 in Appendix B. The conditional observation probabilities in the reduced universe (column that corresponds to the actual observed value) is called the *likelihood*.

- We decide on our prior probability distribution of the parameter. They give our prior belief about each possible value of the parameter π. If we have no idea beforehand, we can choose the prior distribution that has all values equally likely.
- The joint probability distribution of the parameter π and the observation Y is found by multiplying the conditional probability of $Y|\pi$ by the prior probability of π.
- The marginal distribution of Y is found by summing the joint distribution down the columns.

Now take the observed value of Y. It is the only column that is now relevant. It contains the probabilities of the **reduced universe.** Note that it is the *prior* times the

Table 6.11 The joint probability distribution found by multiplying marginal distribution of π (the $prior$) by the conditional distribution of Y given π (which is binomial). $Y = 3$ was observed, so the binomial probabilities of $Y = 3$ (the $likelihood$) are highlighted.

π	$prior$	0	1	2	3	4
.4	$\frac{1}{3}$	$\frac{1}{3}\times$.1296	$\frac{1}{3}\times$.3456	$\frac{1}{3}\times$.3456	$\frac{1}{3}\times$ **.1536**	$\frac{1}{3}\times$.0256
.5	$\frac{1}{3}$	$\frac{1}{3}\times$.0625	$\frac{1}{3}\times$.2500	$\frac{1}{3}\times$.3750	$\frac{1}{3}\times$ **.2500**	$\frac{1}{3}\times$.0625
.6	$\frac{1}{3}$	$\frac{1}{3}\times$.0256	$\frac{1}{3}\times$.1536	$\frac{1}{3}\times$.3456	$\frac{1}{3}\times$ **.3456**	$\frac{1}{3}\times$.1296

Table 6.12 The joint and marginal probability distributions. $Y = 3$ was observed, so those probabilities are highlighted.

π	$prior$	0	1	2	3	4
.4	$\frac{1}{3}$	.0432	.1152	.1152	**.0512**	.0085
.5	$\frac{1}{3}$	.0208	.0833	.1250	**.0833**	.0208
.6	$\frac{1}{3}$	.0085	.0512	.1152	**.1152**	.0432
$marginal$		.0725	.2497	.3554	**.2497**	.0725

likelihood. The posterior probability of each possible value of π is found by dividing that row's element in the relevant column by the marginal probability of Y in that column.

Example 9 *Let $Y|\pi$ be binomial $(n = 4, \pi)$. Suppose we consider there are only three possible values for π, .4,.5, and .6. We will assume they are equally likely. The prior distribution of π and joint distribution of π and Y are given in Table 6.11. The joint probability distribution $f(\pi_i, y_j)$ is found by multiplying the conditional observation distribution $f(y_j|\pi_i)$ times the prior distribution $g(\pi_i)$. In this case, the conditional observation probabilities come from the binomial $(n = 4, \pi)$ distribution. These binomial probabilities come from Table B.1 in Appendix B. Suppose $Y = 3$ was observed. The reduced universe is the column for $Y = 3$. The conditional observation probabilities in that column is called the likelihood and is highlighted.*

The marginal distribution of Y is found by summing the joint distribution of π and Y down the columns. The prior distribution of π , joint probability distribution of (π, Y) , and marginal probability distribution of Y are shown in Table 6.12.

Given $Y = 3$ was observed, only the column labelled 3 is relevant. The prior distribution of π, joint probability distribution of (π, Y) , marginal probability distribution of Y, and posterior probability distribution of $\pi|Y = 3$ are shown in Table 6.13.

Note that the posterior is proportional to prior times likelihood. We didn't have to set up the whole joint probability table. It is easier to only look at the reduced

Table 6.13 The joint, marginal, and posterior probability distribution of π given $Y = 3$. Note the posterior is found by dividing the joint probabilities in the relevant column by their sum.

π	$prior$	0	1	2	3	4	$posterior$
.4	$\frac{1}{3}$	.0432	.1152	.1152	.0512	. 0085	$\frac{.0512}{.2497}$ =.205
.5	$\frac{1}{3}$	.0208	.0833	.1250	.0833	.0208	$\frac{.0833}{.2497}$ =.334
.6	$\frac{1}{3}$	.0085	.0512	.1152	.1152	.0432	$\frac{.1152}{.2497}$ =.461
$marginal$		.0725	.2497	.3554	.2497	.0725	1.000

Table 6.14 The simplified table for finding posterior distribution given $Y = 3$

π	$prior$	$likelihood$	$prior \times likelihood$	$posterior$
.4	$\frac{1}{3}$	.1536	.0512	$\frac{.0512}{.2497} = .205$
.5	$\frac{1}{3}$	.2500	.0833	$\frac{.0833}{.2497} = .334$
.6	$\frac{1}{3}$	.3456	.1152	$\frac{.1152}{.2497} = .461$
$marginal$			.2497	1.000

universe column. The posterior is equal to prior times likelihood divided by the marginal probability of the observed value. The results are shown in Table 6.14.

Setting up the Table for Bayes' Theorem on Binomial with Discrete Prior

- Set up a table with columns for *parameter value, prior, likelihood, prior* $\times$ *likelihood*, and *posterior*.
- Put in the *parameter values*, the *prior*, and the *likelihood* in their respective columns. The *likelihood* values are *binomial(n,*π_i*)* evaluated at the observed value of y. They can be found in Table B.2, or evaluated from the formula.
- Multiply each element in the *prior* column by the corresponding element in the *likelihood* column and put in the *prior* $\times$ *likelihood* column.
- Sum these *prior* $\times$ *likelihood*.
- Divide each element of *prior* $\times$ *likelihood* column by the sum of *prior* $\times$ *likelihood* column. (This rescales them to sum to 1.)
- Put these in the *posterior* column!

Table 6.15 The simplified table for finding posterior distribution given $Y = 3$. Note we are using the proportional $likelihood$ where we have absorbed that part of the binomial distribution that does not depend on π into the constant.

π	$prior$ $(proportional)$	$likelihood$ $(proportional)$	$posterior$
.4	1	$.4^3 \times .6^1$=.0384	$\frac{.0384}{.1873}$ =.205
.5	1	$.5^3 \times .5^1$=.0625	$\frac{.0625}{.1873}$ =.334
.6	1	$.6^3 \times .4^1$=.0864	$\frac{.0864}{.1873}$ =.461
$marginal\ P(Y = 3)$		.1873	1.000

6.3 IMPORTANT CONSEQUENCES OF BAYES' THEOREM

Multiplying all the prior probabilities by a constant does not change the result of Bayes' theorem. Each of the $prior \times likelihood$ entries in the table would be multiplied by the constant. The marginal entry found by summing down the column would also be multiplied by the same constant. Thus the posterior probabilities would be the same as before, since the constant would cancel out. The *relative* weights we are giving to each parameter value, not the actual weights, are what counts. If there is a formula for the prior, any part of it that does not contain the parameter can be absorbed into the constant. This may make calculations simpler for us!

Multiplying the likelihood by a constant does not change the result of Bayes' theorem. The $prior \times likelihood$ values would also be multiplied by the same constant, which would cancel out in the posterior probabilities. The likelihood can be considered the weights given to the possible values by the data. Again, it is the *relative* weights that are important, not the actual weights. If there is a formula for the likelihood, any part that does not contain the parameter can be absorbed into the constant, simplifying our calculations!

Example 9 (continued) *We used a prior that gave each value equal prior probability. In this example there are three possible values, so each has a prior probability equal to* $\frac{1}{3}$. *Let's multiply each of the 3 prior probabilities by the constant 3 to give prior weights equal to 1. This will simplify our calculations. The observations are* binomial $(n = 4, \pi)$, *and* $y = 3$ *was observed. The formula for the binomial likelihood is*

$$f(y|\pi) = \left(\begin{array}{c} 4 \\ 3 \end{array} \right) \pi^3 (1 - \pi)^1 .$$

The binomial coefficient $\left(\begin{array}{c} 4 \\ 3 \end{array} \right)$ *does not contain the parameter, so it is a constant over the likelihood column. To simplify our calculations, we will absorb it into the constant and use only the part of the likelihood that contains the parameter. In Table 6.15 we see this gives us the same result we obtained before.*

Main Points

- The Bayesian universe has two dimensions. The vertical dimension is the parameter space and is unobservable. The horizontal dimension is the sample space and we observe which value occurs.
- The reduced universe is the column for the observed value.
- For discrete prior and discrete observation, the posterior probabilities are found by multiplying the $prior \times likelihood$, and then dividing by their sum.
- When our data arrives in batches, we can use the posterior from the first batch as the prior for the second batch. This is equivalent to combining both batches and using Bayes' theorem only once, using our initial prior.
- Multiplying the $prior$ by a constant doesn't change the result. Only relative weights are important.
- Multiplying the $likelihood$ by a constant doesn't change the result.
- This means we can absorb any part of formula that doesn't contain the parameter into the constant. This greatly simplifies calculations.

Exercises

6.1 There is an urn containing 9 balls, which can be either green or red. The number of red balls in the urn is not known. One ball is drawn at random from the urn, and its color is observed.

(a) What is the *Bayesian universe* of the experiment.

(b) Let X be the number of red balls in the urn. Assume that all possible values of X from 0 to 9 are equally likely. Let Y_1 =1 if the first ball drawn is red, and Y_1=0 otherwise. Fill in the joint probability table for X and Y_1 given below:

X	*prior*	$Y_1 = 0$	$Y_1 = 1$

(c) Find the marginal distribution of Y_1 and put it in the table.

(d) Suppose a red ball was drawn. What is the reduced Bayesian universe?

(e) Calculate the posterior probability distribution of X.

(f) Find the posterior distribution of X by filling in the simplified table:

X	*prior*	*likelihood*	$prior \times likelihood$	*posterior*
$marginal\ P(Y_1 = 1)$				

6.2 Suppose that a second ball is drawn from the urn, without replacing the first. Let $Y_2 = 1$ if the second ball is red, and let it be 0 otherwise. Use the posterior distribution of X from the previous question as the prior distribution for X. Suppose the second ball is green. Find the posterior distribution of X by filling in the simplified table:

X	*prior*	*likelihood*	$prior \times likelihood$	*posterior*
$marginal\ P(Y_2 = 0)$				

6.3 Suppose we look at the two draws from the urn (without replacement) as a single experiment. The results were first draw red, second draw green. Find the posterior distribution of X by filling in the simplified table.

X	*prior*	*likelihood*	$prior \times likelihood$	*posterior*
$marginal\ P(Y_1 = 1, Y_2 = 0)$				

6.4 In the game of "blackjack" also known as "twenty-one," the player and the dealer are dealt one card face-down, and one card face-up. The object is to get as close as possible to the score 21, without exceeding that. Aces count either 1 or 11, face cards count 10, and all other cards count at their face value. The player can ask for more cards to be dealt to him, provided he hasn't gone bust (exceed 21) and lost. Getting 21 on the deal (an ace and a face card or 10) is called a "blackjack." Suppose 4 decks of cards are shuffled together and dealt from. What is the probability the player gets a "blackjack."

6.5 After the hand, the cards are discarded, and the next hand continues with the remaining cards in the deck. The player has had an opportunity to see some of the cards in the previous hand, those that were dealt face-up. Suppose he saw a total of 4 cards, and none of them were aces, nor were any of them a face card or a ten. What is the probability the player gets a "blackjack" on this hand.

Computer Exercises

6.1 Use the Minitab macro *BinoDP.mac* to find the posterior distribution of the binomial probability π when the observation distribution of $Y|\pi$ is *binomial* (n, π) and we have a discrete prior for π.

Suppose we have 8 independent trials and each has one of two possible either success or failure. The probability of success remains constant for each trial. In that case, $Y|\pi$ is *binomial* $(n = 8, \pi)$. Suppose we only allow that there are 6 possible values of π, 0, .2, .4, .6, .8, and 1.0. In that case we say that we have a *discrete* distribution for π. Initially we have no reason to favor one possible value over another. In that case our we would give all the possible values of π probability equal to $\frac{1}{6}$.

π	$g(\pi)$
0	.166666
.2	.166666
.4	.166666
.6	.166666
.8	.166666
1.0	.166666

Suppose we observe 3 "successes" in the 8 trials. Use *BinoDP.mac* or the equivalent R function to find the posterior distribution $g(\pi|y)$. Details for invoking *BinoDP.mac* are in Appendix 3. The details for the equivalent R function are in Appendix 4.

(a) Identify the matrix of conditional probabilities from the output. Relate these conditional probabilities to the binomial probabilities in Table B.1.

(b) What column in the matrix contains the likelihoods?

(c) Identify the matrix of joint probabilities from the output. How are these joint probabilities found?

(d) Identify the marginal probabilities of Y from the output. How are these found?

(e) How are the posterior probabilities found?

6.2 Suppose we take an additional 7 trials, and achieve 2 successes.

(a) Let the posterior after the 8 trials and 3 successes in the previous problem be the prior and use *BinoDP.mac* or the equivalent R function to find the new posterior distribution for π.

(b) In total, we have taken 15 trials and achieved 5 successes. Go back to the original prior and use *BinoDP.mac* or the equivalent R function to find the posterior after the 15 trials and 5 successes.

(c) What does this show?

7

Continuous Random Variables

When we have a continuous random variable, we believe all values over some range are possible if our measurement device is sufficiently accurate. There are an uncountably infinite number of real numbers in an interval, so the probability of getting any particular value must be zero. This makes it impossible to find the probability function of a continuous random variable the same way we did for a discrete random variable. We will have to find a different way to determine its probability distribution. First we consider a thought experiment similar to those done in Chapter 5 for discrete random variables.

Thought Experiment 1: *Independent trials of a continuous random variable*
We start taking a sequence of independent trials of the random variable. We sketch a graph with a spike at each value in the sample equal to the proportion in the sample having that value. After each draw we update the proportions in the accumulated sample that have each value, and update our graph. The updating of the graph at step n is made by scaling all the existing spikes down by the ratio $\frac{n-1}{n}$ and adding $\frac{1}{n}$ to the spike at the value observed at trial n. This keeps the sum of the spike heights equal to 1. Figure 7.1 shows this after 25 draws. Because there are infinitely many possible numbers, it is almost inevitable that we don't draw any of the previous values, so we get a new spike at each draw. After n draws we will have n spikes, each having height $\frac{1}{n}$. Figure 7.2 shows this after 100 draws. As the sample size, n, approaches infinity, the heights of the spikes shrink to zero. This means the probability of getting any particular value is zero. The output of this thought experiment is not the probability

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

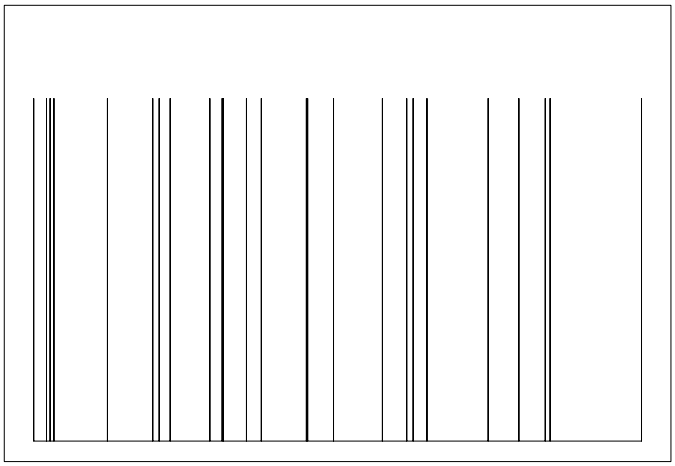


Figure 7.1 Sample probability function after 25 draws.

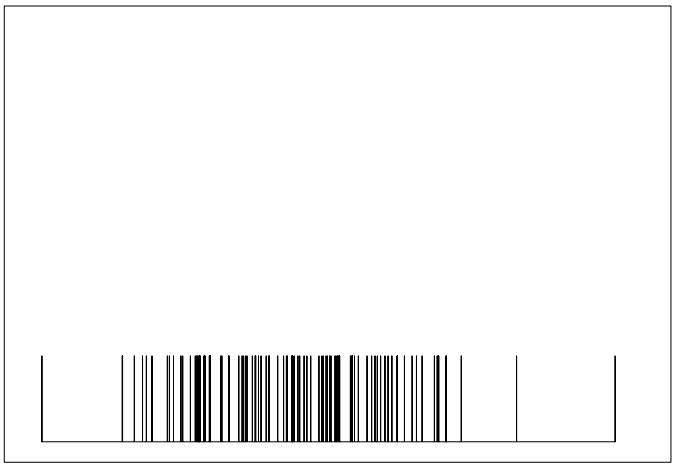


Figure 7.2 Sample probability function after 100 draws.

function, which gives the probability of each possible value. This is not like the output of the thought experiments in Chapter 5 where the random variable was discrete.

What we do notice is that there are some places with many spikes close by, and there are other places with very few spikes close by. In other words, the density of spikes varies. We can think of partitioning the interval into subintervals, and recording the number of observations that fall into each subinterval. We can form

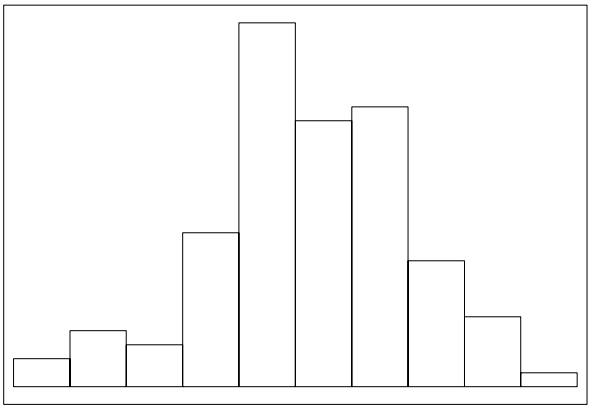


Figure 7.3 Density histogram after 100 draws.

a density histogram by dividing the number in each subinterval by the width of the subinterval. This makes the area under the histogram equal to one. Figure 7.3 shows the density histogram for the first 100 observations.

Now let n increase, and let the width of the subintervals decrease, but at a slower rate than n. Figures 7.4 and 7.5 show the density histogram for the first 1000 and for the first 10,000 observations, respectively. The proportion of observations in a subinterval approaches the probability of being in the subinterval. As n increases, we get a larger number of shorter subintervals. The histograms get closer and closer to a smooth curve.

7.1 PROBABILITY DENSITY FUNCTION

The smooth curve is called the probability density function. It is the limiting shape of the histograms as n goes to infinity, and the width of the bars goes to 0. Its height at a point is not the probability of that point. The thought experiment showed us that probability was equal to zero at every point. Instead, the height of the curve measures how *dense* is the probability at that point.

Since the areas under the histograms all equaled one, the total area under the probability density function must also equal 1:

$$\int_{-\infty}^{\infty} f(y)\, dy = 1\,. \tag{7.1}$$

The proportion of the observations that lie in an interval (a, b) is given by the area of the histogram bars that lie in the interval. In the limit as n increases to infinity, the histograms become the smooth curve, the probability density function. The area of the bars that lie in the interval becomes the area under the curve over that interval. The proportion of observations that lie in the interval becomes the probability that

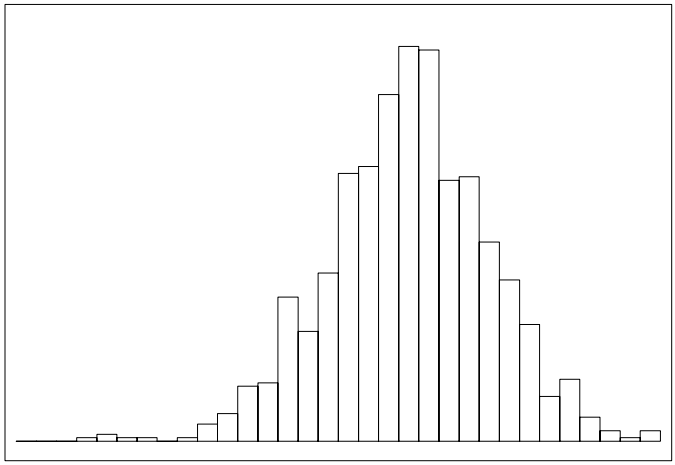


Figure 7.4 Density histogram after 1000 draws.

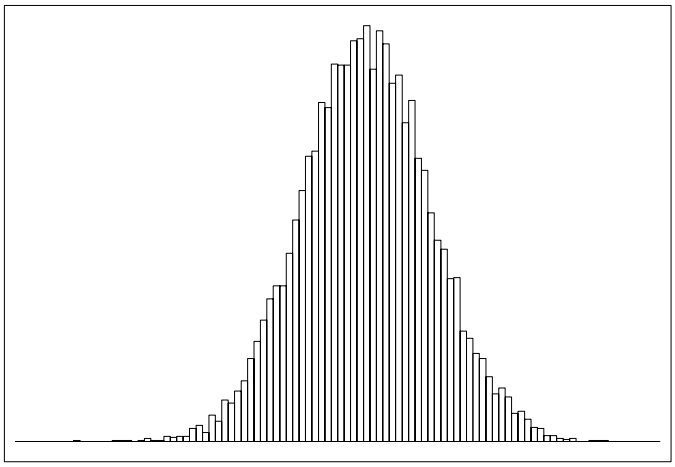


Figure 7.5 Density histogram after 10,000 draws.

the random variable lies in the interval. We know the area under a curve is found by integration, so we can find the probability that the random variable lies in the interval (a, b) by integrating the probability density function over that range:

$$P(a < Y < b) = \int_a^b f(y)\, dy\,. \tag{7.2}$$

Mean of a Continuous Random Variable

In Section 3.3 we defined the mean of the random sample of observations from the random variable to be

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} .$$

Suppose we put the observations in a density histogram where all groups have equal width. The grouped mean of the data is

$$\bar{y} = \sum_{j} m_j \frac{n_j}{n} ,$$

where m_j is the midpoint of the j^{th} bar and $\frac{n_j}{n}$ is its relative frequency. Multiplying and dividing by the width of the bars, we get

$$\bar{y} = \sum_{j} m_j \times width \times \frac{n_j}{n \times width} ,$$

where the relative frequency density $\frac{n_j}{n \times width}$ gives the height of bar j. Multiplying it by *width* gives the area of the bar. Thus the sample mean is the midpoint of each bar times the area of that bar summed over all bars.

Suppose we let n increase without bound, and let the number of bars increase, but at a slower rate. For example, as n increases by a factor of 4, we let the number of bars increase by a factor of 2 so the width of each bar is divided by 2. As n increases without bound, each observation in a group becomes quite close to the midpoint of the group, the number of bars increase without bound, and the width of each bar goes to zero. In the limit, the midpoint of the bar containing the point y approaches y, and the height of the bar containing point y (which is the relative frequency density) approaches $f(y)$. So, in the limit, the relative frequency density approaches the probability density and the sample mean reaches its limit

$$E(Y) = \int_{-\infty}^{\infty} y f(y)\, dy , \tag{7.3}$$

which is called the *expected value* of the random variable. The expected value is like the mean of all possible values of the random variable. Sometimes it is referred to as the mean of the random variable Y and denoted μ.

Variance of a Continuous Random Variable

The expected value $E(Y - E(Y))^2$ is called the variance of the random variable. We can look at the variance of a random sample of numbers, and let the sample size increase.

$$Var(y) = \frac{1}{n} \times \sum_{i=1}^{n} (y_i - \bar{y})^2 \, .$$

As we let n increase, we decrease the width of the bars. This makes each observation become closer to the midpoint of the bar it is in. Now, when we sum over all groups, the variance becomes

$$Var(y) = \sum_{j} \frac{n_j}{n} (m_j - \bar{y})^2 \, .$$

We multiply and divide by the width of the bar to get

$$Var(y) = \sum_{j} \frac{n_j}{n \times width} \times width \times (m_j - \bar{y})^2 \, .$$

This is the square of the midpoint minus the mean times the area of the bar summed over all bars. As n increases to ∞, the relative frequency density approaches the probability density , the midpoint of the bar containing the point y approaches y, and the sample mean $\bar{y}$ approaches the expected value $E(Y)$, so in the limit the variance becomes

$$Var(Y) = E[(Y - E(Y)^2] = \int_{-\infty}^{\infty} (y - \mu)^2 f(y) \, dy \, . \tag{7.4}$$

The variance of the random variable is denoted σ^2. We can square the term in brackets,

$$Var(Y) = \int_{-\infty}^{\infty} (y^2 - 2\mu y + \mu^2) f(y) \, dy$$

and break the integral into three terms,

$$Var(Y) = \int_{-\infty}^{\infty} y^2 \, f(y) \, dy - 2\mu \int_{-\infty}^{\infty} y \, f(y) \, dy + \mu^2 \int_{-\infty}^{\infty} f(y) \, dy \, ,$$

and simplify to get an alternate form for the variance:

$$Var(Y) = E(Y^2) - [E(Y)]^2 \, . \tag{7.5}$$

7.2 SOME CONTINUOUS DISTRIBUTIONS

Uniform Distribution

The random variable has the *uniform* $(0, 1)$ distribution if its probability density function is constant over the interval [0,1], and 0 everywhere else.

$$g(x) = \begin{cases} 1 & \text{for} \quad 0 \leq x \leq 1 \\ 0 & \text{for} \quad \notin [0, 1] \end{cases} \, .$$

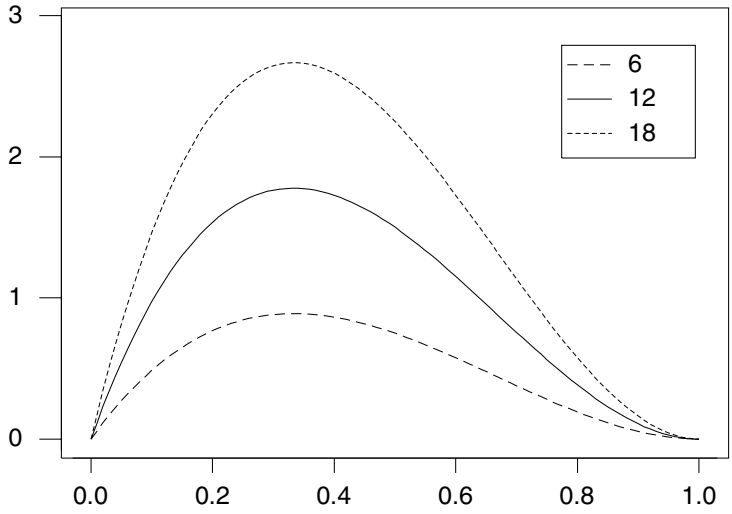


Figure 7.6 The curve $g(x) = kx^1(1-x)^2$ for several values of k.

It is easily shown that the mean and variance of a uniform (0,1) random variable are $\frac{1}{2}$ and $\frac{1}{12}$ respectively.

Beta Family of Distributions

The *Beta(a,b)* distribution is another commonly used distribution for a continuous random variable that can only take on values $0 \leq x \leq 1$. It has the probability density function

$$g(x;a,b) = \begin{cases} k \times x^{a-1}(1-x)^{b-1} & \text{for} \quad 0 \leq x \leq 1 \\ 0 & \text{for} \quad \notin [0,1] \end{cases} .$$

The most important thing is that $x^{a-1}(1-x)^{b-1}$ determines the shape of the curve, and k is only the constant needed to make this a probability density function. Figure 7.6 shows the graphs of this for $a = 2$ and $b = 3$ for a number of values of k. We see that the curves all have the same basic shape but have different areas under the curves. The value of $k = 12$ gives area equal to 1, so that is the one that makes a density function.

The distribution with shape given by $x^{a-1}(1-x)^{b-1}$ is called the *beta* (a,b) distribution. The constant needed to make the curve a density function is given by the formula

$$k = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} ,$$

where $\Gamma(c)$ is the Gamma function, which is a generalization of the factorial function.[1] The probability density function of the *beta* (a, b) distribution is given by

$$g(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}\,. \tag{7.6}$$

All we need remember is that $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the constant needed to make the curve with shape given by $x^{a-1}(1-x)^{b-1}$ a density. a equals one plus the power of x and b equals one plus the power of $(1-x)$.

This curve can have different shapes depending on the values a and b, so the *beta(a,b)* is actually a family of distributions. The *uniform(0,1)* distribution is a special case of the *beta(a,b)* distribution, where $a = 1$ and $b = 1$.

Mean of a beta distribution. The expected value of a continuous random variable x is found by integrating the variable times the density function over the whole range of possible values. (Since the *beta(a,b)* density equals 0 for x outside the interval $[0, 1]$, the integration only has to go from 0 to 1, not $-\infty$ to ∞.) For a random variable having the *beta(a,b)* distribution,

$$E(X) = \int_0^1 x \times g(x; a, b)dx = \int_0^1 x \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}\, dx\,.$$

However, by using our understanding of the *beta* distribution, we can evaluate this integral without having to do the integration. First move the constant out in front of the integral, then combine the x terms by adding exponents:

$$E(X) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x \times x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^a(1-x)^{b-1}\, dx\,.$$

We recognize the part under the integral sign as a curve that has the *beta(a+1,b)* shape. So we must multiply inside the integral by the appropriate constant to make it integrate to 1, and multiply by the reciprocal of the constant outside of the integral to keep the balance:

$$E(X) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} x^a(1-x)^{b-1}\, dx\,.$$

The integral equals 1, and when we use the fact that $\Gamma(c) = (c-1) \times \Gamma(c-1)$ and do some cancellation, we get the simple formula

$$E(X) = \frac{a}{a+b} \tag{7.7}$$

for the mean of a *beta(a,b)* random variable.

[1] When c is an integer, $\Gamma(c) = (c-1)!$. The Gamma function always satisfies the equation $\Gamma(c) = (c-1) \times \Gamma(c-1)$ whether or not c is an integer.

Variance of a beta distribution. For a continuous random variable the expected value of a *function* of a random variable is found by integrating the *function* times the density function over the whole range of possible values.

For a random variable having the *beta(a,b)* distribution,

$$E(X^2) = \int_0^1 x^2 \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}\, dx\,.$$

When we evaluate this integral using the properties of the *beta(a,b)* distribution, we get

$$E(X^2) = \frac{a(a+1)}{(a+b+1)(a+b)}\,.$$

When we substitute this formula and the formula for the mean into Equation 7.5 and simplify, we find the variance of the random variable having the *beta(a,b)* distribution is

$$Var(X) = E(X^2) - [E(X)]^2 \;=\; \frac{ab}{(a+b)^2(a+b+1)}\,. \tag{7.8}$$

Normal Distribution

Very often data appear to have a symmetric bell-shaped distribution. In the early years of statistics, this shape seemed to occur so frequently that it was thought to be normal. The family of distributions with this shape has become known as the *normal* distribution family. It is also known as the *Gaussian* distribution after the mathematician Gauss who studied its properties. It is the most widely used distribution in statistics. We will see that there is a good reason for its frequent occurrence. However, we must remain aware that the term *normal distribution* is only a name, and distributions with other shapes are not abnormal.

The *normal*(μ, σ^2) distribution is the member of the family having mean μ and variance σ^2. The probability density function of a *normal*(μ, σ^2) distribution is given by

$$g(x|\mu, \sigma^2) = ke^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

for $-\infty < x < \infty$ where k is the constant value needed to make this a probability density. The shape of the curve is determined by $e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$. Figure 7.7 shows the curve $ke^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ for several values of k. Changing the value of k only changes the area under the curve, not its basic shape. To be a probability density function, the area under the curve must equal 1. The value of k that makes the curve a probability density is $k = \frac{1}{\sqrt{2\pi}\sigma}$.

Central limit theorem. The central limit theorem says that if you take a random sample $y_1, \ldots, y_n$ from any shape distribution having mean μ and variance σ^2, then the limiting distribution of $\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$ is *normal* $(0, 1)$. The shape of the limiting distribution is *normal* despite the original distribution not necessarily being normal. A linear transformation of a normal distribution is also normal, so the shape of $\bar{y}$

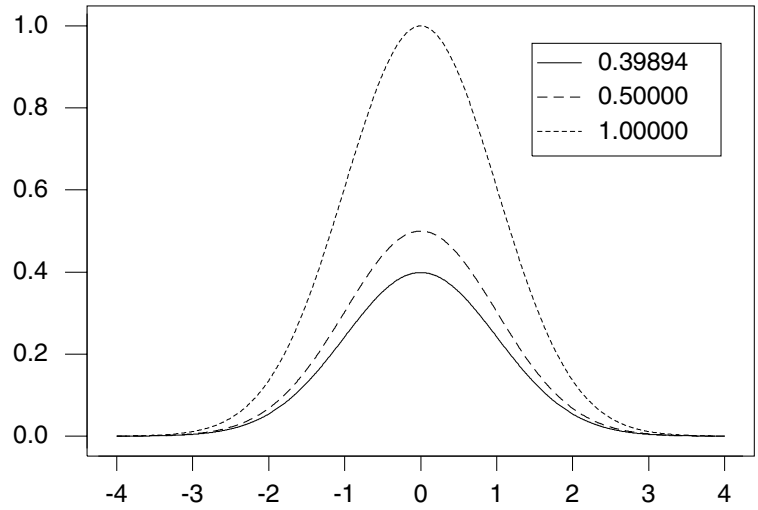


Figure 7.7 The curve $g(x) = ke^{-\frac{1}{2}(x-0)^2}$ for several values of k.

and $\sum y$ are also normal. Amazingly, n doesn't have to be particularly large for the shape to be approximately normal, $n \geq 25$ is sufficient.

The key factor of the central limit distribution is that when we are averaging a large number of independent effects, each of which is small in relation to the sum, the distribution of the sum approaches the *normal* shape regardless of the shapes of the individual distributions. Thus any random variable that arises as the sum of a large number of independent effects will be approximately normal! This explains why the normal distribution is encountered so frequently.

Finding probabilities using standard normal table. The standard normal density has mean $\mu = 0$ and variance $\sigma^2 = 1$. Its probability density function is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

We note that this curve is symmetric about $z = 0$. Unfortunately, Equation 7.2, the general form for finding the probability $P(a \leq z \leq b)$ isn't any practical use here. There is no closed form for integrating the standard normal probability density function. Instead, the area between 0 and z for values of z between 0 and 3.99 has been numerically calculated and tabulated in Table B.2 in Appendix B. We use this table to calculate the probability we need.

Example 10 *Suppose we want to find* $P(-.62 \leq Z \leq 1.37)$. *In Figure 7.8 we see that the shaded area between* $-.62$ *and 1.37 is the sum of the two areas between* $-.62$ *and 0 and between 0 and 1.37 respectively. The area between* $-.62$ *and 0 is the same as the area between 0 and* $+.62$ *because the standard normal density is symmetric*

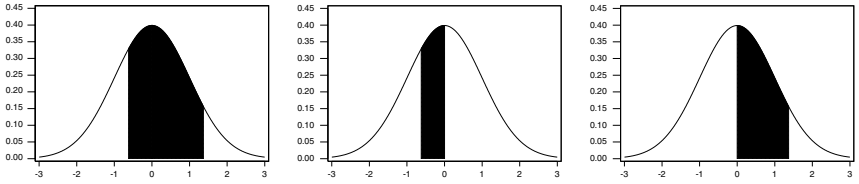


Figure 7.8 The area between $-.62$ and 1.37 split into two parts.

about 0. In Table B.2 we find this area equals .2291. The area between 0 and 1.37 equals .4147 from the table. So

$$\begin{aligned} P(-.62 \le Z \le 1.37) &= .2291 + .4147 \\ &= .6338\,. \end{aligned}$$

Any normal distribution can be transformed into a standard normal by subtracting the mean and then dividing by the standard deviation. This lets us find any normal probability using the areas under the standard normal density found in Table B.2.

Example 11 *Suppose we know Y is normal with mean $\mu = 10.8$ and standard deviation $\sigma = 2.1$, and suppose we want to find the probability $P(Y \ge 9.9)$.*

$$\begin{aligned} P(Y \ge 9.9) &= P(Y - 10.8 \ge 9.9 - 10.8) \\ &= P\left(\frac{Y - 10.8}{2.1} \ge \frac{9.9 - 10.8}{2.1}\right)\,. \end{aligned}$$

The left side is a standard normal. The right side is a number. We find this probability from the standard normal:

$$\begin{aligned} P(Y \ge 9.9) &= P(Z \ge -.429) \\ &= .1659 + .5000 \\ &= .6659\,. \end{aligned}$$

Finding beta probabilities using normal approximation. We can approximate a *beta* (a, b) distribution by the *normal* distribution having the same mean and variance. This approximation is very effective when both a and b are greater than or equal to ten.

Example 12 *Suppose Y has the beta $(12, 25)$ distribution and we wish to find $P(Y > .4)$. The mean and variance of Y are*

$$E(Y) = \frac{12}{37} = .3243 \quad \text{and} \quad Var(Y) = \frac{12 \times 25}{37^2 \times 38} = .005767$$

respectively. We approximate the beta $(12, 25)$ distribution with a normal $(.3243, .005767)$

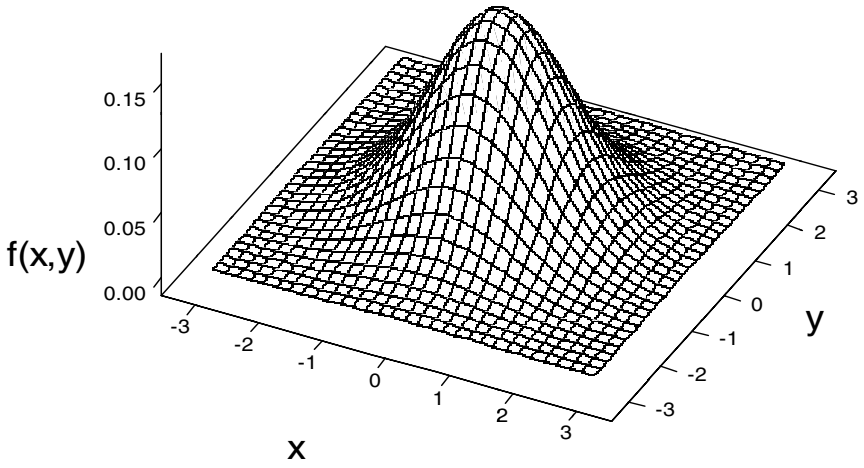


Figure 7.9 A joint density.

distribution. The approximate probability is

$$\begin{aligned} P(Y > .4) &= P\left(\frac{Y - .3243}{\sqrt{.005767}} > \frac{.4 - .3243}{\sqrt{.005767}}\right) \\ &= P(Z > .997) \\ &= .3406\,. \end{aligned}$$

7.3 JOINT CONTINUOUS RANDOM VARIABLES

We consider two (or more) random variables distributed together. If both X and Y are continuous random variables, they have joint density $f(x, y)$, which measures the probability density at the point (x, y). This would be found by dividing the plane into rectangular regions by partitioning both the x axis and y axis. We look at the proportion of the sample that lie in a region. We increase n, the sample size of the joint random variables without bound, and at the same time decrease the width of the regions (in both dimensions) at a slower rate. In the limit, the proportion of the sample lying in the region centered at (x, y) approaches the joint density $f(x, y)$. Figure 7.9 shows a joint density function.

We might be interested in determining the density of one of the joint random variables by itself, its *marginal* density. When X and Y are joint random variables

that are both continuous, the marginal density of Y is found by integrating the joint density over the whole range of X:

$$f(y) = \int_{-\infty}^{\infty} f(x, y)\, dx\,,$$

and vice versa. (Finding the marginal density by integrating the joint density over the whole range of one variable is analogous to finding the marginal probability distribution by summing the joint probability distribution over all possible values of one variable for jointly distributed discrete random variables.)

Conditional Probability Density

The conditional density of X given $Y = y$ is given by

$$f(x|y) = \frac{f(x, y)}{f(y)}\,.$$

We see that the conditional density of X given $Y = y$ is proportional to the joint density where $Y = y$ is held fixed. Dividing by the marginal density $f(y)$ makes the integral of the conditional density over the whole range of x equal 1. This makes it a proper density function.

7.4 JOINT CONTINUOUS AND DISCRETE RANDOM VARIABLES

It may be that one of the variables is continuous, and the other is discrete. For instance, let X be continuous, and let Y be discrete. In that case $f(x, y_j)$ is a joint probability-probability density function. In the x direction it is continuous, and in the y direction it is discrete. This is shown in Figure 7.10. In this case, the marginal density of the continuous random variable X is found by

$$f(x) = \sum_j f(x, y_j)\,,$$

and the marginal probability function of the discrete random variable Y is found by

$$f(y_j) = \int f(x, y_j)\, dx\,.$$

The conditional density of X given $Y = y_j$ is given by

$$f(x|y_j) = \frac{f(x, y_j)}{f(y_j)} = \frac{f(x, y_j)}{\int f(x, y_j)\, dx}\,.$$

We see that this is proportional to the joint probability-probability density function $f(x, y_j)$ where x is allowed to vary over its whole range. Dividing by the marginal probability $f(y_j)$ just scales it to be a proper density function (integrates to 1).

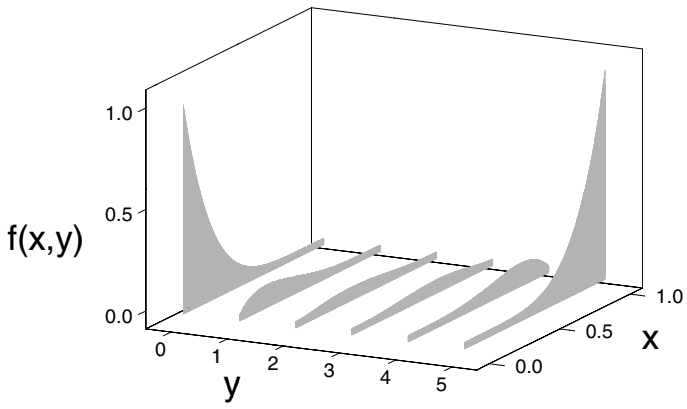


Figure 7.10 A joint continuous and discrete distribution.

Similarly, the conditional distribution of $Y = y_j$ given x is found by

$$f(y_j|x) = \frac{f(x, y_j)}{f(x)} = \frac{f(x, y_j)}{\sum_j f(x, y_j)} .$$

This is also proportional to the joint probability-probability density function $f(x, y_j)$ where x is fixed, and Y is allowed to take on all the possible values $y_1, \ldots, y_J$.

Main Points

- The probability that a continuous random variable equals any particular value is zero!
- The probability density function of a continuous random variable is a smooth curve that measures the *density* of probability at each value. It is found as the limit of density histograms of random samples of the random variable, where the sample size increases to infinity and the width of the bars goes to zero.
- The probability a continuous random variable lies between two values a and b is given by the area under the probability density function between the two values. This is found by the integral

$$P(a < X < b) = \int_a^b f(x)\, dx .$$

- The expected value of a continuous random variable X is found by integrating x times the density function $f(x)$ over the whole range.

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx\,.$$

- A *beta* (a, b) random variable has probability density

$$f(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad \text{for} \quad 0 \le x \le 1\,.$$

- The mean and variance of a *beta* (a, b) random variable are given by

$$E(X) = \frac{a}{a+b} \qquad \text{and} \qquad Var(X) = \frac{a \times b}{(a+b)^2 \times (a+b+1)}\,.$$

- A *normal* (μ, σ^2) random variable has probability density

$$g(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

  where μ is the mean, and σ^2 is the variance.

- The central limit theorem says that for a random sample $y_1, \ldots y_n$ from any distribution $f(y)$ having mean μ and variance σ^2, the distribution of

$$\frac{\bar{y} - \mu}{\sigma/\sqrt{n}}$$

  is approximately *normal* $(0, 1)$ for $n > 25$. This is regardless of the shape of the original density $f(y)$.

- By reasoning similar to that of the central limit theorem, any random variable that is the sum of a large number of independent random variables will be approximately normal. This is the reason why the normal distribution occurs so frequently.

- The marginal distribution of y is found by integrating the joint distribution $f(x, y)$ with respect to x over its whole range.

- The conditional distribution of x given y is proportional to the joint distribution $f(x, y)$ where y fixed and x is allowed to vary over its whole range.

$$f(x|y) = \frac{f(x,y)}{f(y)}\,.$$

  Dividing by the marginal distribution of $f(y)$ scales it properly so that $f(y|x)$ integrates to 1 and is a probability density function.

Exercises

7.1 Let X have a *beta* $(3, 5)$ distribution.

(a) Find $E(X)$.
(b) Find $Var(X)$.

7.2 Let X have a *beta* $(12, 4)$ distribution.

(a) Find $E(X)$.
(b) Find $Var(X)$.

7.3 Let X have the *uniform* distribution.

(a) Find $E(X)$.
(b) Find $Var(X)$.
(c) Find $P(X \leq .25)$.
(d) Find $P(.33 < X < .75)$.

7.4 Let X be a random variable having probability density function

$$f(x) = 2x \quad \text{for} \quad 0 \leq x \leq 1\,.$$

(a) Find $P(X \geq .75)$.
(b) Find $P(.25 \leq X \leq .6)$.

7.5 Let Z have the standard normal distribution.

(a) Find $P(0 \leq Z \leq .65)$.
(b) Find $P(Z \geq .54)$.
(c) Find $P(-.35 \leq Z \leq 1.34)$.

7.6 Let Z have the standard normal distribution.

(a) Find $P(0 \leq Z \leq 1.52)$.
(b) Find $P(Z \geq 2.11)$.
(c) Find $P(-1.45 \leq Z \leq 1.74)$.

7.7 Let Y be normally distributed with mean $\mu = 120$ and variance $\sigma^2 = 64$.

(a) Find $P(Y \leq 130)$.
(b) Find $P(Y \geq 135)$.
(c) Find $P(114 \leq Y \leq 127)$.

7.8 Let Y be normally distributed with mean $\mu = 860$ and variance $\sigma^2 = 576$.

(a) Find $P(Y \leq 900)$.
(b) Find $P(Y \geq 825)$.
(c) Find $P(840 \leq Y \leq 890)$.

7.9 Let Y be distributed according to the *beta* $(10, 12)$ distribution.

(a) Find $E(Y)$.

(b) Find $Var(Y)$.

(c) Find $P(Y > .5)$ using the normal approximation.

7.10 let Y be distributed according to the *beta* $(15, 10)$ distribution.

(a) Find $E(Y)$.

(b) Find $Var(Y)$.

(c) Find $P(Y < .5)$ using the normal approximation.

8

Bayesian Inference for Binomial Proportion

Frequently there is a large population where π, a proportion of the population, has some attribute. For instance, the population could be registered voters living in a city, and the attribute is "plans to vote for candidate A for mayor." We take a random sample from the population and let Y be the observed number in the sample having the attribute, in this case the number who say they plan to vote candidate A for mayor.

We are counting the total number of "successes" in n independent trials where each trial has two possible outcomes, "success" and "failure." Success on trial i means the item drawn on trial i has the attribute. The probability of success on any single trial is π, the proportion in the population having the attribute. This proportion remains constant over all trials because the population is large.

The conditional distribution of the observation Y, the total number of successes in n trials given the parameter π, is *binomial* (n, π). The conditional probability function for y given π is given by

$$f(y|\pi) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y} \quad \text{for} \quad y = 1, \ldots, n\,.$$

Here we are holding π fixed, and looking at the probability distribution of y over its possible values.

If we look at this same relationship between π and y, but hold y fixed at the number of successes we observed, and let π vary over its possible values, we have

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

the likelihood function given by

$$f(y|\pi) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y} \quad \text{for} \quad 0 \leq \pi \leq 1 \,.$$

We see that we are looking at the same relationship as the distribution of the observation y given the parameter π, but the subject of the formula has changed to the parameter, for the observation held at the value that actually occurred.

To use Bayes' theorem, we need a prior distribution $g(\pi)$ that gives our belief about the possible values of the parameter π before taking the data. It is important to realize that the prior must not be constructed from the data. Bayes' theorem is summarized by *posterior is proportional to the prior times the likelihood*. The multiplication in Bayes' theorem can only be justified when the prior is *independent* of the likelihood![1] This means the observed data must not have any influence on the choice of prior! The posterior distribution is proportional to prior distribution times likelihood:

$$g(\pi|y) \propto g(\pi) \times f(y|\pi) \,.$$

This gives us the shape of the posterior density, but not the exact posterior density itself. To get the actual posterior, we need to divide this by some constant k to make sure it is a probability distribution, meaning that the area under the posterior integrates to 1. We find k by integrating $g(\pi) \times f(y|\pi)$ over the whole range. So, in general,

$$g(\pi|y) = \frac{g(\pi) \times f(y|\pi)}{\int_0^1 g(\pi) \times f(y|\pi) \, d\pi} \,, \tag{8.1}$$

which requires an integration. Depending on the prior $g(\pi)$ chosen, there may not necessarily be a closed form for the integral, so it may be necessary to do the integration numerically. We will look at some possible priors.

8.1 USING A UNIFORM PRIOR

If we don't have any idea beforehand what the proportion π is, we might like to choose a prior that does not favor any one value over another. Or, we may want to be as objective as possible, and not put our personal belief into the inference. In that case we should use the uniform prior that gives equal weight to all possible values of the success probability π. Although this does not achieve universal objectivity

[1] We know that for independent events (or random variables) the joint probability (or density) is the product of the marginal probabilities (or density functions). If they are not independent this does not hold. Likelihoods come from probability functions or probability density functions, so the same pattern holds. They can only be multiplied when they are independent.

(which is impossible to achieve), it is objective for this formulation of the problem:[2]

$$g(\pi) = 1 \quad \text{for} \quad 0 \leq \pi \leq 1\,.$$

Clearly, we see that in this case, the posterior density is proportional to the likelihood:

$$g(\pi|y) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y} \quad \text{for} \quad 0 \leq \pi \leq 1\,.$$

We can ignore the part that doesn't depend on π. It is a constant for all values of π, so it doesn't affect the shape of the posterior. When we examine that part of the formula that shows the shape of the posterior as a function of π, we recognize this is a *beta(a,b)* distribution where $a = y + 1$ and $b = n - y + 1$. So in this case, the posterior distribution of π given y is easily obtained. All that is necessary is look at the exponents of π and $(1 - \pi)$. We didn't have to do the integration.

8.2 USING A BETA PRIOR

Suppose a *beta (a,b)* prior density is used for π:

$$g(\pi; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \pi^{a-1}(1-\pi)^{b-1} \quad \text{for} \quad 0 \leq \pi \leq 1\,.$$

The posterior is proportional to prior times likelihood. We can ignore the constants in the prior and likelihood that don't depend on the parameter, since we know multiplying either the prior or the likelihood by a constant won't affect the results of Bayes' theorem. This gives

$$g(\pi|y) \propto \pi^{a+y-1}(1-\pi)^{b+n-y-1} \qquad \text{for} \qquad 0 \leq \pi \leq 1$$

which is the shape of the posterior as a function of π. We recognize that this is the beta distribution with parameters $a' = a + y$ and $b' = b + n - y$. That is, we add the number of successes to a, and add the number of failures to b:

$$g(\pi|y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\, \pi^{y+a-1}(1-\pi)^{n-y+b-1}$$

for $0 \leq \pi \leq 1$. Again, the posterior density of π has been easily obtained without having to go through the integration.

Figure 8.1 shows the shapes of *beta(a,b)* densities for values of $a = .5, 1, 2, 3$ and $b = .5, 1, 2, 3$. This shows the variety of shapes members of the *beta(a,b)* family can take. When $a < b$, the density has more weight in the lower half. The opposite is true when $a > b$. When $a = b$, the *beta(a,b)* density is symmetric. We note that the uniform prior is a special case of the *beta(a,b)* prior where $a = 1$ and $b = 1$.

[2]There are many possible parameterizations of the problem. Any one-to-one function of the parameter would also be a suitable parameter. The prior density for the new parameter could be found from the prior density of the original parameter using the change of variable formula, and would not be flat. In other words, it would favor some values of the new parameter over others. You can be objective in a given parameterization, but it would not be objective in the new formulation. *Universal* objectivity is not attainable.

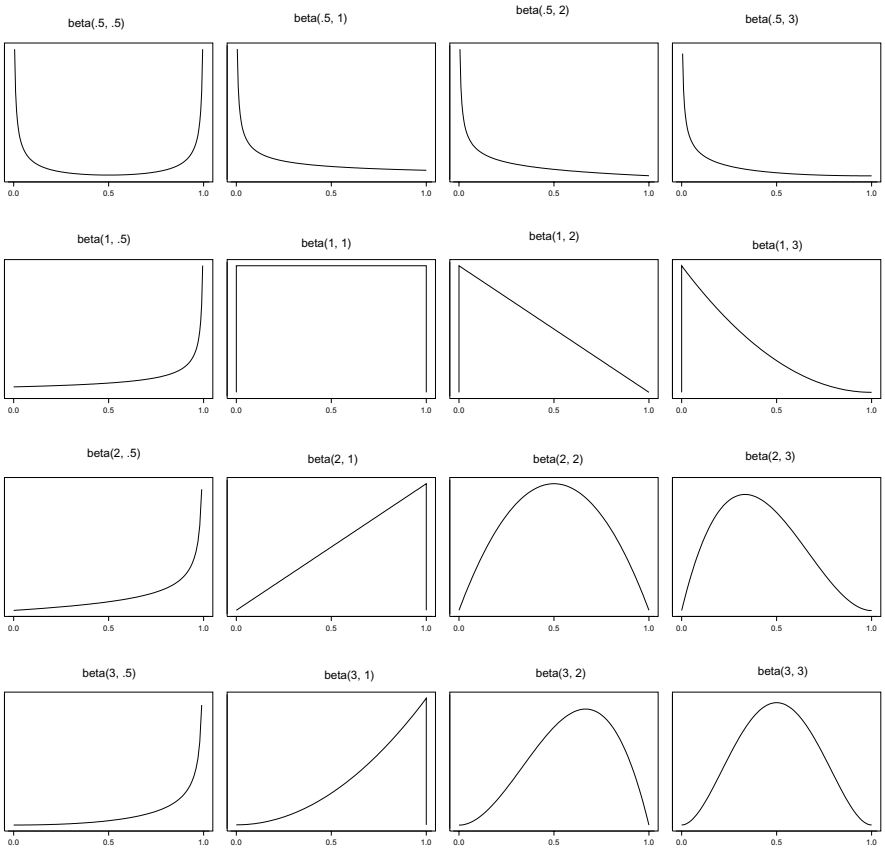


Figure 8.1 Some beta distributions.

Conjugate Family of Priors for Binomial Observation is the Beta Family

When we examine the shape of the binomial likelihood function as a function of π, we see that this is of the same form as the *beta(a,b)* distribution, a product of π to a power times $(1-\pi)$ to another power. When we multiply the beta prior times the binomial likelihood, we add the exponents of π and $(1-\pi)$, respectively. So we start with a beta prior, we get a beta posterior by the simple rule "add successes to a, add failures to b." This makes using *beta(a,b)* priors when we have binomial observations particularly easy. Using Bayes' theorem moves us to another member of the same family.

We say that the *beta* distribution is the conjugate[3] family for the *binomial* observation distribution. When we use a prior from the conjugate family, we don't have to do any integration to find the posterior. All we have to do is use the observations to update the parameters of the conjugate family prior to find the conjugate family posterior. This is a big advantage.

8.3 CHOOSING YOUR PRIOR

Bayes' theorem gives you a method to revise your (belief) distribution about the parameter, given the data. In order to use it, you must have a distribution that represents your belief about the parameter, before we look at the data.[4] This is your prior distribution. In this section we propose some methods to help you choose your prior, and things to consider in prior choice.

Choosing a Conjugate Prior When You Have Vague Prior Knowledge

When you have vague prior knowledge, one of the *beta(a,b)* prior distributions shown in Figure 8.1 would be a suitable prior. For example, if your prior knowledge about π, is that π is very small, then *beta(.5,1)*, *beta(.5,2)*, *beta(.5,3)*, *beta(1,2)*, or *beta(1,3)* would all be satisfactory priors. All of these conjugate priors offer easy computation of the posterior, together with putting most of the prior probability at small values of π. It doesn't much matter very much which one you chose; the resulting posteriors given the data would be very similar.

Choosing a Conjugate Prior When You Have Real Prior Knowledge by Matching Location and Scale

The *beta(a,b)* family of distributions is the conjugate family for *binomial*(n, π) observations. We saw in the previous section that priors from this family have significant advantages computationally. The posterior will be a member of the same family, with the parameters updated by simple rules. We can find the posterior without integration. The beta distribution can have a number of shapes. The prior chosen should correspond to your belief. We suggest choosing a *beta(a,b)* that matches your prior belief about the (location) mean and (scale) standard deviation[5]. Let π_0 be your

[3]Conjugate priors only exists when the observation distribution comes from the *exponential* family. In that case the observation distribution can be written $f(y|\theta) = a(\theta)b(y)e^{c(\theta)\times T(y)}$. The conjugate family of priors will then have the same functional form as the likelihood of the observation distribution.

[4]This could be elicited from your coherent betting strategy about the parameter value. Having a coherent betting strategy means that if someone started offering you bets about the parameter value, you would not take a worse bet than one you already rejected, nor would you refuse to take a better bet than one you already accepted.

[5]Some people would say that you should not use a conjugate prior just because of these advantages. Instead, you should elicit your prior from your coherent betting strategy. I don't think most people carry around a coherent betting strategy in their head. Their prior belief is less structured. They have a belief about the *location* and *scale* of the parameter distribution. Choosing a prior by finding the conjugate

prior mean for the proportion, and let σ_0 be your prior standard deviation for the proportion.

The mean of *beta(a,b)* distribution is $\frac{a}{a+b}$. Set this equal to what your prior belief about the mean of the proportion to give

$$\pi_0 = \frac{a}{a+b}\,.$$

The standard deviation of beta distribution is $\sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$. Set this equal to what your prior belief about the standard deviation for the proportion. Noting that $\frac{a}{a+b} = \pi_0$ and $\frac{b}{a+b} = 1 - \pi_0$, we see

$$\sigma_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{a+b+1}}\,.$$

Solving these two equations for a and b gives your *beta(a,b)* prior.

Precautions Before Using Your Conjugate Prior

1. Graph your *beta(a,b)* prior. If the shape looks reasonably close to what you believe, you will use it. Otherwise, you can adjust π_0 and σ_0 until you find a prior whose graph approximately corresponds to your belief. As long as the prior has reasonable probability over the whole range you think the parameter could possibly be in, it will be a satisfactory prior.

2. Calculate the *equivalent sample size* of the prior. We note that the sample proportion $\hat{\pi} = \frac{y}{n}$ from a *binomial(n,π)* distribution has variance equal to $\frac{\pi(1-\pi)}{n}$. We equate this variance (at π_0, the prior mean) to the prior variance.

$$\frac{\pi_0(1-\pi_0)}{n_{eq}} = \frac{ab}{(a+b)^2 \times (a+b+1)}\,.$$

   Since $\pi_0 = \frac{a}{a+b}$ and $(1-\pi_0) = \frac{b}{a+b}$, the equivalent sample size is $n_{eq} = a+b+1$. It says that the amount of information about the parameter from your prior is equivalent to the amount from a random sample of that size. You should always check if this is unrealistically high. Ask yourself, "Is my prior knowledge about π really equal to the knowledge about π that I would obtain if I checked a random sample of size n_{eq}? If it is not, you should increase your prior standard deviation and recalculate your prior. Otherwise, you would be putting too much prior information about the parameter relative to the amount of information that will come from the data.

family member that matches these beliefs will give a prior on which a coherent betting strategy could be based!

Table 8.1 Chris's prior weights. His continuous prior is found by linearly interpolating between them.

Value	Weight
0	0
.05	1
.1	2
.3	2
.4	1
.5	0

Constructing a General Continuous Prior

Your prior shows the *relative* weights you give each possible value before you see the data. The shape of your prior belief may not match the *beta* shape. You can construct a discrete prior that matches your belief weights at several values over the range you believe possible, and then interpolate between them to make the continuous prior. You can ignore the constant needed to make this a density, because when you multiply the prior by a constant, the constant gets cancelled out by Bayes' theorem. However, if you do construct your prior this way, you will have to evaluate the integral of the prior times likelihood numerically to get the posterior. This will be shown in Example 13.

Example 13 *Three students are constructing their prior belief about π, the proportion of Hamilton residents who support building a casino in Hamilton. Anna thinks that her prior mean is .2, and her prior standard deviation is .08. The beta(a,b) prior that satisfies her prior belief is found by*

$$\frac{.2 \times .8}{a+b+1} = .08^2 .$$

Therefore her equivalent sample size is $a+b+1 = 25$. For Anna's prior, $a = 4.8$ and $b = 19.2$.

Bart is a newcomer to Hamilton, so he is not aware of the local feeling for or against the proposed casino. He decides to use a uniform prior. For him, $a = b = 1$. His equivalent sample size is $a+b+1 = 3$.

Chris can't fit a beta(a,b) prior to match his belief. He believes his prior probability has a trapezoidal shape. He gives heights of his prior in the Table 8.1, and linearly interpolates between them to get his continuous prior. When we interpolate between these points, we see that Chris's prior is given by

$$g(\pi) = \begin{cases} 20\pi & \text{for} \quad 0 \le \pi \le .20 \\ .2 & \text{for} \quad .20 \le \pi \le .30 \\ 5 - 10\pi & \text{for} \quad .30 \le \pi \le .50 \end{cases} .$$

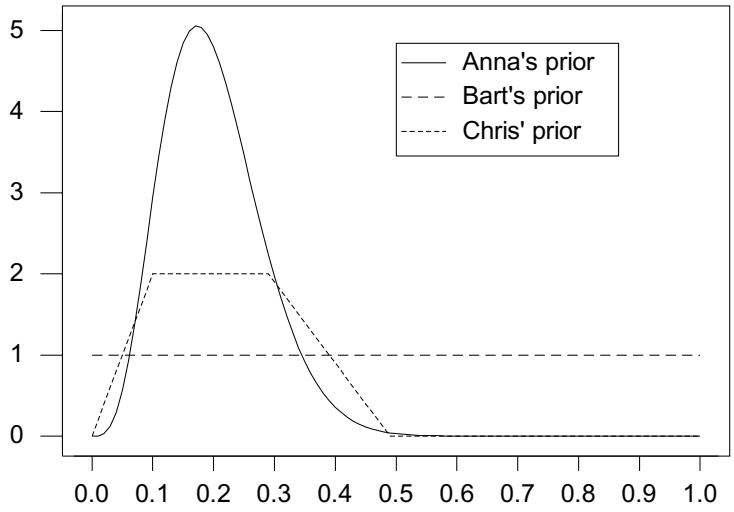


Figure 8.2 Anna's, Bart's, and Chris' prior distributions.

The three priors are shown in the Figure 8.2. Note that Chris's prior is not actually a density since it doesn't have area equal to one. However, remember this is not a problem since it is only the relative *weights that are important.*

Effect of the Prior

When we have enough data, the effect of the prior we choose will be small compared to the data. In that case we will find that we can get very similar posteriors despite starting from quite different priors. All that is necessary is that they give reasonable weight over the range that is indicated by the likelihood. The exact shape of the prior doesn't much matter. The data are said to "swamp the prior."

Example 13 (continued) *The three students take a random sample of* $n = 100$ *Hamilton residents and find their views on the casino. Out of the random sample,* $y = 26$ *said they support building a casino in Hamilton. Anna's posterior is beta(4.8 + 26, 19.2 + 74). Bart's posterior is beta(1 + 26, 1 + 74). Chris' posterior is found using Equation 8.1. We need to evaluate Chris' prior numerically. To do this, we integrate Chris'* $prior \times likelihood$ *using the Minitab macro tintegral.mac. The three posteriors are shown in Figure 8.3. We see that the three students end up with very similar posteriors, despite starting with priors having quite different shapes.*

8.4 SUMMARIZING THE POSTERIOR DISTRIBUTION

The posterior distribution summarizes our belief about the parameter *after* seeing the data. It takes into account our prior belief (the prior distribution) and the data

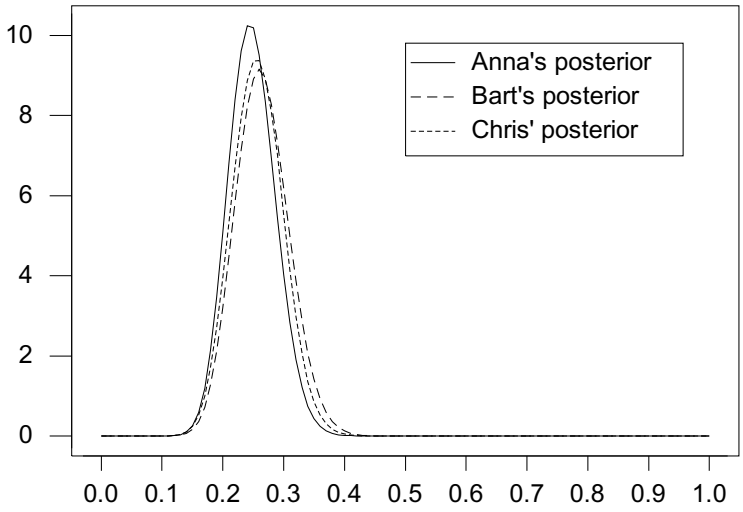


Figure 8.3 Anna's, Bart's, and Chris' posterior distributions.

(likelihood). A graph of the posterior shows us all we can know about the parameter, after the data. A distribution is hard to interpret. Often we want to find a few numbers that characterize it. These include measures of location that determine where most of the probability is on the number line, and measures of spread that determine how widely the probability is spread. They could also include percentiles of the distribution. We may want to determine an interval that has a high probability of containing the parameter. These are known as *Bayesian credible intervals* and are somewhat analogous to confidence intervals. However, they have the direct probability interpretation that confidence intervals lack.

Measures of Location

First, we want to know where the posterior distribution is located on the number line. There are three possible measures of location we will consider: posterior mode, posterior median, and posterior mean.

Posterior mode. This is the value that maximizes the posterior distribution. If the posterior distribution is continuous, it can be found by setting the derivative of the posterior density equal to zero. When the posterior $g(\pi|y)$ is $beta(a', b')$, its derivative

$$g'(\pi|y) = (a'-1)\pi^{a'-2} \times (1-\pi)^{b'-1} + \pi^{a'-1} \times (-1)(b'-1)(1-\pi)^{b'-2}\,.$$

(Note: the prime $'$ has two meanings in this equation; $g'(\pi|y)$ is the derivative of the posterior, a' and b' are the constants of the *beta* posterior found by the updating

rules.) Setting $g'(\pi|y)$ equal to 0 and solving gives the posterior mode

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

The posterior mode has some potential disadvantages as a measure of location. First, it may lie at or near one end of the distribution, and thus not be representative of the distribution as a whole. Second, there may be multiple local maximums. When we set the derivative function equal to zero and solve, we will find all of them and the local minimums as well.

Posterior median. This is the value that has 50% of posterior distribution below it, 50% above it. If $g(\pi|y)$ is *beta*(a', b'), it is the solution of

$$\int_0^{median} g(\pi|y)\, d\pi = .5 .$$

The only disadvantage of the posterior median is that it has to be found numerically. It is an excellent measure of location.

Posterior mean. The posterior mean is a very frequently used measure of location. It is the expected value, or mean, of the posterior distribution.

$$m' = \int_0^1 \pi g(\pi|y)\, d\pi . \tag{8.2}$$

The posterior mean is strongly affected when the distribution has a heavy tail. For a skewed distribution with one heavy tail, the posterior mean may be quite a distance away from most of the probability. When the posterior $g(\pi|y)$ is *beta*(a', b') the posterior mean equals

$$m' = \frac{a'}{a' + b'} . \tag{8.3}$$

The *beta(a,b)* distribution is bounded between 0 and 1, so it does not have heavy tails. The posterior mean will be a good measure of location for a *beta* posterior.

Measures of Spread

The second thing we want to know about the posterior distribution is how spread out it is. If it has large spread, then our knowledge about the parameter, even after analyzing the observed data, is still imprecise.

Posterior variance. This is the variance of posterior distribution.

$$Var(\pi|y) = \int_0^1 (\pi - m')^2 g(\pi|y)\, d\pi . \tag{8.4}$$

When we have a $\beta(a', b')$ posterior the posterior variance is

$$Var(\pi|y) = \frac{a' \times b'}{(a' + b')^2 \times (a' + b' + 1)} . \tag{8.5}$$

The posterior variance is very affected for heavy tailed distributions. For a heavy tailed distribution, the variance will be very large, yet most of the probability is very concentrated quite close the middle of the distribution. It is also in squared units, which makes it hard to interpret its size in relation to the size of the mean. We overcome these disadvantages of the posterior variance by using the posterior standard deviation.

Posterior standard deviation. This is the square root of posterior variance. It is in terms of units, so its size can be compared to the size of the mean, and it will be less affected by heavy tails.

Percentiles of the posterior distribution. The k^{th} percentile of the posterior distribution is the value π_k, which has $k\%$ of the area below it. It is the found numerically by solving

$$k = 100 \times \int_{-\infty}^{\pi_k} g(\pi|y)\, d\pi .$$

Some percentiles are particularly important. The first (or lower) quartile Q_1 is the 25^{th} percentile. The second quartile , Q_2 (or median) is the 50^{th} percentile, and the third (or upper) quartile Q_3 is the 75^{th} percentile.

The interquartile range. The interquartile range

$$IQR = Q_3 - Q_1$$

is a useful measure of spread that is not affected by heavy tails.

Example 13 (continued) *Anna, Bart, and Chris computed some measures of location and spread for their posterior distributions. Anna and Bart used Equations 8.3 and 8.5 to find their posterior mean and variance, respectively, since they had beta posteriors. Chris used Equations 8.2 and 8.4 to find his posterior mean and variance since his posterior did not have the beta distribution. He evaluated the integrals numerically using the Minitab macro tintegral.mac. Their posterior means, medians, standard deviations, and interquartile ranges are shown in Table 8.2. We see clearly that the posterior distributions have similar summary statistics, despite the different priors used.*

8.5 ESTIMATING THE PROPORTION

A point estimate $\hat{\pi}$ is a statistic calculated from the data used as an estimate of the parameter π. Suitable Bayesian point estimates are single values such as measures of location calculated from the posterior distribution. The posterior mean and posterior median are often used as point estimates.

Table 8.2 Measures of location and spread of posterior distributions

Person	Posterior	Mean	Median	Std. Dev.	IQR
Anna	*beta(30.8,93.2)*	.248	.247	.039	.053
Bart	*beta(27,75)*	.270	.263	.044	.059
Chris	numerical	.261	.255	.041	.057

The posterior mean square of an estimate. The posterior mean square of an estimator $\hat{\pi}$ of the proportion π is

$$PMS(\hat{\pi}) = \int_0^1 (\pi - \hat{\pi})^2 \, g(\pi|y) \, d\pi \, . \tag{8.6}$$

It measures the average squared distance (with respect to the posterior) that the estimate is away from the true value. Adding and subtracting the posterior mean m' we get

$$PMS(\hat{\pi}) = \int_0^1 (\pi - m' + m' - \hat{\pi})^2 \, g(\pi|y) \, d\pi \, .$$

Multiplying out the square we get

$$PMS(\hat{\pi}) = \int_0^1 [(\pi - m')^2 + 2(\pi - m')(m' - \hat{\pi}) + (m' - \hat{\pi})^2] \, g(\pi|y) \, d\pi \, .$$

We split the integral into three integrals. Since both m' and $\hat{\pi}$ are constants with respect to the posterior distribution when we evaluate the integrals we get

$$PMS(\hat{\pi}) = Var(\pi|y) + 0 + (m' - \hat{\pi})^2 \, . \tag{8.7}$$

This is the posterior variance of π plus the square of the distance $\hat{\pi}$ is away from the posterior mean m'.

The last term is a square, and always greater than or equal to zero. We see that on average, the squared distance the true value is away from the posterior mean m' is less than that for any other possible estimate $\hat{\pi}$, given our prior belief and the observed data. The posterior mean is the optimum estimator *post-data*. That's a good reason to use the posterior mean as the estimate, and explains why the posterior mean is the most widely used Bayesian estimate. We will use the posterior mean as our estimate for π.

8.6 BAYESIAN CREDIBLE INTERVAL

Often we wish to find a high probability interval for the parameter. A range of values that has a known high *posterior* probability, $(1 - \alpha)$, of containing the parameter is known as a Bayesian credible interval. It is sometimes called *Bayesian confidence interval*. In the next chapter we will see that credible intervals answer a more

relevant question than do ordinary frequentist confidence intervals, because of the direct probability interpretation.

There are many possible intervals with same (posterior) probability. The shortest interval with given probability is preferred. It would be found by having the equal heights of the posterior density at the lower and upper endpoints, and total tail area of $1-\alpha$. The upper and lower tails would not necessarily have equal tail areas. However, it is often easier to split the total tail area into equal parts and find the interval with equal tail areas.

Bayesian Credible Interval for π

If we used a *beta*(a, b) prior, the posterior distribution of $\pi|y$ is *beta(*a', b'*)*. An equal tail area 95% Bayesian credible interval for π can be found by obtaining the difference between the 97.5 th and the 2.5 th percentiles. Using Minitab, pull down *calc* menu to *probability distributions* over to *beta* and fill in the dialog box. Without Minitab, we approximate the *beta(*a', b'*)* posterior distribution by the normal distribution having the same mean and variance,

$$(\pi|y) \text{ is approximately } N[m'; (s')^2]$$

where the posterior mean

$$m' = \frac{a'}{a'+b'},$$

and the posterior variance

$$(s')^2 = \frac{a'b'}{(a'+b')^2(a'+b'+1)}.$$

The $(1-\alpha)\times 100\%$ credible region for π is approximately

$$m' \pm z_{\frac{\alpha}{2}} \times s', \tag{8.8}$$

where $z_{\frac{\alpha}{2}}$ is the value found from the standard normal table. For a 95% credible interval, $z_{.025} = 1.96$. The approximation works very well if both $a' \geq 10$ and $b' \geq 10$

Example 13 (continued) *Anna, Bart, and Chris calculated 95% credible intervals for π having equal tail areas two ways; using the exact (*beta*) density function, and using the normal approximation. These are shown in Table 8.3. Anna, Bart, and Chris have slightly different credible intervals because they started with different prior beliefs. But the effect of the data was much greater than the effect of their priors and they end up with very similar credible intervals. We see that in each case, the 95% credible interval for π calculated using the normal approximation is nearly identical to the corresponding exact 95% credible interval.*

Table 8.3 Exact and approximate 95% credible intervals

Person	Posterior Distribution	Credible Interval *Exact*		Credible Interval *Normal Approximation*	
		Lower	Upper	Lower	Upper
Anna	*beta(30.8,93.2)*	.177	.328	.172	.324
Bart	*beta(27,75)*	.184	.354	.183	.355
Chris	*numerical*	.181	.340	.181	.341

Main Points

- The key relationship is $posterior \propto prior \times likelihood$. This gives us the shape of the posterior density. We must find the constant to divide this by to make it a density, eg. integrate to 1 over its whole range.

- The constant we need is $k = \int_0^1 g(\pi) \times f(y|\pi) d\pi$. In general, this integral does not have a closed form, so we have to evaluate it numerically.

- If the prior is *beta(a,b)*, then the posterior is *beta*(a', b') where the constants are updated by simple rules $a' = a + y$ (add number of successes to a) and $b' = b + n - y$ (add number of failures to b).

- The *beta* family of priors is called the *conjugate* family for *binomial* observation distribution. This means that the posterior is also a member of the same family, and it can easily be found without the need for any integration.

- It makes sense to choose a prior from the conjugate family, which makes finding the posterior easier. Find the *beta(a,b)* prior that has mean and standard deviation that correspond to your prior belief. Then graph it to make sure that it looks similar to your belief. If so, use it. If you have no prior knowledge about π at all, you can use the *uniform* prior which gives equal weight to all values. The *uniform* is actually the *beta(1,1)* prior.

- If you have some prior knowledge, and you can't find a member of the conjugate family that matches it, you can construct a discrete prior at several values over the range, and interpolate between them to make the prior continuous. Of course, you may ignore the constant needed to make this a density, since any constant gets cancelled out by when you divide by $\int prior \times likelihood$ to find the exact posterior.

- The main thing is that your prior must have reasonable probability over all values that realistically are possible. If that is the case, the actual shape doesn't matter very much. If there is a reasonable amount of data, different people will get similar posteriors, despite starting from quite different shaped priors.

- The posterior mean is the estimate that has the smallest posterior mean square. This means that, on average (with respect to posterior), it is closer to the parameter than any other estimate. In other words, given our prior belief and the observed data, the posterior mean will be, on average, closer to the parameter than any other estimate. It is the most widely used Bayesian estimate because it is optimal *post-data*.
- A $(1-\alpha)\times 100\%$ Bayesian credible interval is an interval that has a posterior probability of $1-\alpha$ of containing the parameter.
- The shortest $(1-\alpha)\times 100\%$ Bayesian credible interval would have equal posterior density heights at the lower and upper endpoints, however, the areas of the two tails would not necessarily be equal.
- Equal tail area Bayesian credible intervals are often used instead, because they are easier to find.

Exercises

8.1 In order to determine how effective a magazine is at reaching its target audience, a market research company selects a random sample of people from the target audience and interviews them. Out of the 150 people in the sample, 29 had seen the latest issue.

(a) What is the distribution of y, the number who have seen the latest issue?

(b) Use a uniform prior for π, the proportion of the target audience that has seen the latest issue. What is the posterior distribution of π?

8.2 A city is considering building a new museum. The local paper wishes to determine the level of support for this project, and is going to conduct a poll of city residents. Out of the sample of 120 people, 74 support the city building the museum.

(a) What is the distribution of y, the number who support the building the museum?

(b) Use a uniform prior for π, the proportion of the target audience that support the museum. What is the posterior distribution of π?

8.3 Sophie, the editor of the student newspaper, is going to conduct a survey of students to determine the level of support for the current president of the students association. She needs to determine her prior distribution for π, the proportion of students who support the president. She decides her prior mean is .5, and her prior standard deviation is .15.

(a) Determine the *beta* (a, b) prior that matches her prior belief.

(b) What is the equivalent sample size of her prior?

(c) Out of the 68 students that she polls, $y = 21$ support the current president. Determine her posterior distribution.

8.4 You are going to take a random sample of voters in a city in order to estimate the proportion π who support stopping the fluoridation of the municipal water supply. Before you analyze the data, you need a prior distribution for π. You decide that your prior mean is .4, and your prior standard deviation is .1.

(a) Determine the *beta* (a, b) prior that matches your prior belief.

(b) What is the equivalent sample size of your prior?

(c) Out of the 100 city voters polled, $y = 21$ support the removal of fluoridation from the municipal water supply. Determine your posterior distribution.

8.5 In a research program on human health risk from recreational contact with water contaminated with pathogenic microbiological material, the National Institute of Water and Atmosphere (NIWA) instituted a study to determine the quality of New Zealand stream water at a variety of catchment types. This study is documented in McBride et al. (2002) where $n = 116$ one-liter water samples from sites identified as having a heavy environmental impact from birds (seagulls) and waterfowl. Out of these samples, $y = 17$ samples contained Giardia cysts.

(a) What is the distribution of y, the number of samples containing Giardia cysts?

(b) Let π be the true probability that a one-liter water sample from this type of site contains Giardia cysts. Use a *beta* $(1, 4)$ prior for π. Find the posterior distribution of π given y.

(c) Summarize the posterior distribution by its first two moments.

(d) Find the *normal* approximation to the posterior distribution $g(\pi|y)$.

(e) Compute a 95% credible interval for π using the normal approximation in part (c).

8.6 The same study found that $y = 12$ out of $n = 145$ samples identified as having a heavy environmental impact from dairy farms contained Giardia cysts.

(a) What is the distribution of y, the number of samples containing Giardia cysts?

(b) Let π be the true probability that a one-liter water sample from this type of site contains Giardia cysts. Use a *beta* $(1, 4)$ prior for π. Find the posterior distribution of π given y.

(c) Summarize the posterior distribution by its first two moments.

(d) Find the *normal* approximation to the posterior distribution $g(\pi|y)$.

(e) Compute a 95% credible interval for π using the normal approximation in part (c).

8.7 The same study found that $y = 10$ out of $n = 174$ samples identified as having a heavy environmental impact from pastoral (sheep) farms contained Giardia cysts.

(a) What is the distribution of y, the number of samples containing Giardia cysts?

(b) Let π be the true probability that a one-liter water sample from this type of site contains Giardia cysts. Use a *beta* $(1, 4)$ prior for π. Find the posterior distribution of π given y.

(c) Summarize the posterior distribution by its first two moments.

(d) Find the *normal* approximation to the posterior distribution $g(\pi|y)$.

(e) Compute a 95% credible interval for π using the normal approximation in part (c).

8.8 The same study found that $y = 6$ out of $n = 87$ samples within municipal catchments contained Giardia cysts.

(a) What is the distribution of y, the number of samples containing Giardia cysts?

(b) Let π be the true probability that a one-liter water sample from a site within a municipal catchment contains Giardia cysts. Use a *beta* $(1, 4)$ prior for π. Find the posterior distribution of π given y.

(c) Summarize the posterior distribution by its first two moments.

(d) Find the *normal* approximation to the posterior distribution $g(\pi|y)$.

(e) Calculate a 95% credible interval for π using the normal approximation in part (c).

Computer Exercises

8.1 We will use the Minitab macro *BinoBP.mac* or the equivalent R function to find the posterior distribution of the binomial probability π when the observation distribution of $Y|\pi$ is *binomial* (n, π) and we have a *beta* (a, b) prior for π. The *beta* family of priors is the conjugate family for *binomial* observations. That means that if we start with one member of the family as the prior distribution, we will get another member of the family as the posterior distribution. It is especially easy, for when we start with a *beta* (a, b) prior, we get a *beta* (a', b') posterior where $a' = a + y$ and $b' = b + n - y$.

Suppose we have 15 independent trials and each trial results in one of two possible outcomes, success or failure. The probability of success remains constant for each trial. In that case, $Y|\pi$ is *binomial* $(n = 15, \pi)$. Suppose that

we observed $y = 6$ successes. Let us start with a *beta* $(1, 1)$ prior. The details for invoking *BinoBP.mac* and the equivalent R function are given in Appendix 3 and Appendix 4, respectively. Store π, the prior $g(\pi)$, the likelihood $f(y|\pi)$, and the posterior $g(\pi|y)$ in columns c1-c4 respectively.

(a) What are the posterior mean and standard deviation?

(b) Find a 95% credible interval for π.

8.2 Repeat part (a) with a *beta* $(2, 4)$ prior, storing the likelihood and posterior in c5 and c6.

8.3 Graph both posteriors on the same graph. What do you notice? What do you notice about the two posterior means and standard deviations? What do you notice about the two credible intervals for π?

8.4 We will use the Minitab macro *BinoGCP.mac* or the equivalent R function to find the posterior distribution of the binomial probability π when the observation distribution of $Y|\pi$ is *binomial* (n, π) and we have a general continuous prior for π. Suppose the prior has the shape given by

$$g(\pi) = \begin{cases} \pi & \text{for} \quad \pi \leq .2 \\ .2 & \text{for} \quad .2 < \pi < .3 \\ .5 - \pi & \text{for} \quad .3 < \pi \leq .5 \\ 0 & \text{for} \quad .5 < \pi \end{cases} .$$

Store the values of π and prior $g(\pi)$ in columns c1 and c2, respectively. Suppose out of $n = 20$ independent trials, $y = 7$ successes were observed.

(a) Use *BinoGCP.mac* or the equivalent R function to determine the posterior distribution $g(\pi|y)$. Details for invoking *BinoGCP.mac* and the equivalent R function are in Appendix 3 and Appendix 4, respectively.

(b) Use *tintegral.mac* ind the posterior mean and posterior standard deviation of π. Details for invoking *tintegral.mac* and the equivalent R function are in Appendix 3 and Appendix 4, respectively.

(c) Find a 95% credible interval for π by using *tintegral.mac* or the equivalent R function.

8.5 Repeat the previous question with a *uniform* prior for π.

8.6 Graph the two posterior distributions on the same graph. What do you notice? What do you notice about the two posterior means and standard deviations? What do you notice about the two credible intervals for π?

9

Comparing Bayesian and Frequentist Inferences for Proportion

The posterior distribution of the parameter given the data gives the complete inference from the Bayesian point of view. It summarizes our belief about the parameter after we have analyzed the data. However, from the frequentist point of view there are several different types of inference that can be made about the parameter. These include point estimation, interval estimation, and hypothesis testing. These frequentist inferences about the parameter require probabilities calculated from the sampling distribution of the data, given the fixed but unknown parameter. These probabilities are based on all possible random samples that could have occurred. These probabilities are not conditional on the actual sample that did occur!

In this chapter we will see how we can do these types of inferences using the Bayesian viewpoint. These Bayesian inferences will use probabilities calculated from the posterior distribution. That makes them conditional on the sample that actually did occur.

9.1 FREQUENTIST INTERPRETATION OF PROBABILITY AND PARAMETERS

Most statistical work is done using the frequentist paradigm. A random sample of observations is drawn from a distribution with an unknown parameter. The parameter is assumed to be a fixed but unknown constant. This doesn't allow any probability distribution to be associated with it. The only probability considered is the probability

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

distribution of the random sample of size *n* given the parameter. This explains how the random sample varies over all possible random samples, given the fixed but unknown parameter value. The probability is interpreted as long run relative frequency.

Sampling Distribution of Statistic

Let $Y_1, \ldots, Y_n$ be a random sample from a distribution that depends on a parameter θ. Suppose a statistic S is calculated from the random sample. This statistic can be interpreted as a random variable, since the random sample can vary over all possible samples. Calculate the statistic for each possible random sample of size *n*. The distribution of these values is called the *sampling distribution of the statistic*. It explains how the statistic varies over all possible random samples of size *n*. Of course, the sampling distribution also depends on the unknown value of the parameter θ. We will write this sampling distribution as

$$f(s|\theta)\,.$$

However, we must remember that in frequentist statistics, the parameter θ is a fixed but unknown constant, not a random variable. The sampling distribution measures how the statistic varies over all possible samples given the unknown fixed parameter value. This distribution does not have anything to do with the actual data that occurred. It is the distribution of values of the statistic that could have occurred, given that specific parameter value. Frequentist statistics uses the sampling distribution of the statistic to perform inference on the parameter. From a Bayesian perspective, this is a backwards form of inference.[1]

This contrasts with Bayesian statistics where the complete inference is the posterior distribution of the parameter given the actual data that occurred:

$$g(\theta|data)\,.$$

Any subsequent Bayesian inference such as a Bayesian estimate or a Bayesian credible interval is calculated from the posterior distribution. Thus the estimate or the credible interval depends on the data that actually occurred. Bayesian inference is straightforward.[2]

[1] Frequentist statistics performs inferences in the parameter space, which is the unobservable dimension of the Bayesian universe, based on a probability distribution in the sample space, which is the observable dimension.

[2] Bayesian statistics performs inference in the parameter space based on a probability distribution in the parameter space.

9.2 POINT ESTIMATION

The first type of inference we consider is point estimation, where a single statistic is calculated from the sample data and used to estimate the unknown parameter. The statistic depends on the random sample, so it is a random variable, and its distribution is its sampling distribution. If its sampling distribution is centered close to the true but unknown parameter value θ, and the sampling distribution does not have much spread, the statistic could be used to estimate the parameter. We would call the statistic an *estimator* of the parameter and the value it takes for the actual sample data an *estimate*. There are several theoretical approaches for finding frequentist estimators, such as maximum likelihood estimation (MLE)[3] and uniformly minimum variance unbiased estimation (UMVUE). We will not go into them here. Instead, we will use the sample statistic that corresponds to the population parameter we wish to estimate, such as the sample proportion as the frequentist estimator for the population proportion. This turns out to be the same estimator that would be found using either of the main theoretical approaches (MLE and UMVUE) for estimating the binomial parameter π.

From a Bayesian perspective, point estimation means that we would use a single statistic to summarize the posterior distribution. The most important number summarizing a distribution would be its location. The posterior mean, or the posterior median would be good candidates here. We will use the posterior mean as the Bayesian estimate because it minimizes the posterior mean squared error, as we saw in the previous chapter. This means it will be the optimal estimator, given our prior belief and this sample data (i.e., *post-data*).

Frequentist Criteria for Evaluating Estimators

We don't know the true value of the parameter, so we can't judge an estimator from the value it gives for the random sample. Instead, we will use a criterion based on the sampling distribution of the estimator that is the distribution of the estimator over all possible random samples. We compare possible estimators by looking at how concentrated their sampling distributions are around the parameter value for a range of fixed possible values. When we use the sampling distribution, we are still thinking of the estimator as a random variable because we haven't yet obtained the sample data and calculated the estimate. This is a *pre-data* analysis.

Although this "what if the parameter has this value" type of analysis comes from a frequentist point of view, it can be used to evaluate Bayesian estimators as well. It can be done before we obtain the data and in Bayesian statistics it is called a *pre-posterior* analysis. The procedure is used to evaluate how the estimator performs over all possible random samples, given that parameter value. We often find that Bayesian estimators perform very well when evaluated this way, sometimes even better than frequentist estimators.

[3] Maximum likelihood estimation was pioneered by R. A. Fisher

Unbiased Estimators

The expected value of an estimator is a measure of the center of its distribution. This is the average value that the estimator would have averaged over all possible samples. An estimator is said to be *unbiased* if the mean of its sampling distribution is the true parameter value. That is, an estimator $\hat{\theta}$ is unbiased if and only if

$$E(\hat{\theta}) = \int \hat{\theta} f(\hat{\theta}|\theta) d\hat{\theta} = \theta \,,$$

where $f(\hat{\theta}|\theta)$ is the sampling distribution of the estimator $\hat{\theta}$ given the parameter θ. Frequentist statistics emphasizes unbiased estimators because averaged over all possible random samples, an unbiased estimator gives the true value. The bias of an estimator $\hat{\theta}$ is the difference between its expected value and the true parameter value.

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta \,. \tag{9.1}$$

Unbiased estimators have bias equal to zero.

In contrast, Bayesian statistics does not place any emphasis on being unbiased. In fact Bayesian estimators are usually biased.

Minimum Variance Unbiased Estimator

An estimator is said to be a minimum variance unbiased estimator if no other unbiased estimator has a smaller variance. Minimum variance unbiased estimators are often considered the *best* estimators in frequentist statistics. The sampling distribution of a minimum variance unbiased estimator has the smallest spread (as measured by the variance) of all sampling distributions that have mean equal to the parameter value.

However, it is possible that there may be biased estimators that, on average, are closer to the true value than the best unbiased estimator. We need to look at a possible trade-off between bias and variance. Figure 9.1 shows the sampling distributions of three possible estimators of θ. Estimator 1 and estimator 2 are seen to be unbiased estimators. Estimator 1 is the *best unbiased* estimator, since it has the smallest variance among the unbiased estimators. Estimator 3 is seen to be a biased estimator, but it has a smaller variance than estimator 1. We need some way of comparison that includes biased estimators, to find which one will be closest, on average, to the parameter value.

Mean Squared Error of an Estimator

The (frequentist) mean squared error of an estimator $\hat{\theta}$ is the average squared distance the estimator is away from the true value:

$$\begin{aligned} MS(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= \int (\hat{\theta} - \theta)^2 \, f(\hat{\theta}|\theta) \, d\hat{\theta} \,. \end{aligned} \tag{9.2}$$

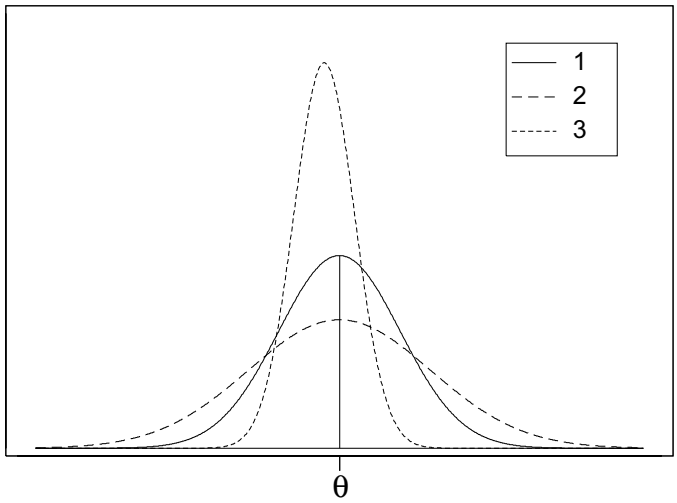


Figure 9.1 Sampling distributions of three estimators.

The frequentist mean squared error is calculated from the sampling distribution of the estimator, which means the averaging is over all possible samples given that fixed parameter value. It is *not* the posterior mean square calculated from the posterior distribution that we introduced in the previous chapter. It turns out that the mean squared error of an estimator is the square of the bias plus the variance of the estimator:

$$MS(\hat{\theta}) = bias(\hat{\theta})^2 + Var(\hat{\theta})\,. \tag{9.3}$$

Thus it gives a better frequentist criterion for judging estimators than the bias or the variance alone. An estimator that has a smaller mean squared error is closer to the true value averaged over all possible samples.

9.3 COMPARING ESTIMATORS FOR PROPORTION

Bayesian estimators often have smaller mean squared errors than frequentist estimators. In other words, on average, they are closer to the true value. Thus Bayesian estimators can be better than frequentist estimators, even when judged by the frequentist criterion of mean squared error.

The frequentist estimator for π is

$$\hat{\pi}_f = \frac{y}{n}\,,$$

where y, the number of successes in the n trials, has the binomial (n, π) distribution. $\hat{\pi}_f$ is unbiased, and $Var(\hat{\pi}_f) = \frac{\pi \times (1-\pi)}{n}$. Hence the mean squared error of $\hat{\pi}_f$ equals

$$MS(\hat{\pi}_f) \quad = \quad 0^2 + Var(\hat{\pi}_f)$$

$$= \frac{\pi \times (1-\pi)}{n}.$$

Suppose we use the posterior mean as the Bayesian estimate for π, where we use the Beta(1,1) prior (uniform prior). The estimator is the posterior mean, so

$$\hat{\pi}_B = m' = \frac{a'}{a'+b'},$$

where $a' = 1 + y$ and $b' = 1 + n - y$. We can rewrite this as a linear function of y, the number of successes in the n trials:

$$\hat{\pi}_B = \frac{y+1}{n+2} = \frac{y}{n+2} + \frac{1}{n+2}.$$

Thus, the mean of its sampling distribution is

$$\frac{n\pi}{n+2} + \frac{1}{n+2},$$

and the variance of its sampling distribution is

$$\left[\frac{1}{n+2}\right]^2 \times n\pi(1-\pi).$$

Hence from Equation 9.3, the mean squared error is

$$\begin{aligned} MS(\hat{\pi}_B) &= \left[\frac{n\pi}{n+2} \times \pi + \frac{1}{n+2} - \pi\right]^2 + \left[\frac{1}{n+2}\right]^2 \times n\pi(1-\pi) \\ &= \left[\frac{1-2\pi}{n+2}\right]^2 + \left[\frac{1}{n+2}\right]^2 \times n\pi(1-\pi). \end{aligned}$$

For example, suppose $\pi = .4$ and the sample size is $n = 10$. Then

$$\begin{aligned} MS(\hat{\pi}_f) &= \frac{.4 \times .6}{10} \\ &= .024 \end{aligned}$$

and

$$\begin{aligned} MS(\hat{\pi}_B) &= \left[\frac{1 - 2 \times .4}{12}\right]^2 + \left[\frac{1}{12}\right]^2 \times 10 \times .4 \times .6 \\ &= .0169. \end{aligned}$$

Next, suppose $\pi = .5$ and $n = 10$. Then

$$\begin{aligned} MS(\hat{\pi}_f) &= \frac{.5 \times .5}{10} \\ &= .025 \end{aligned}$$

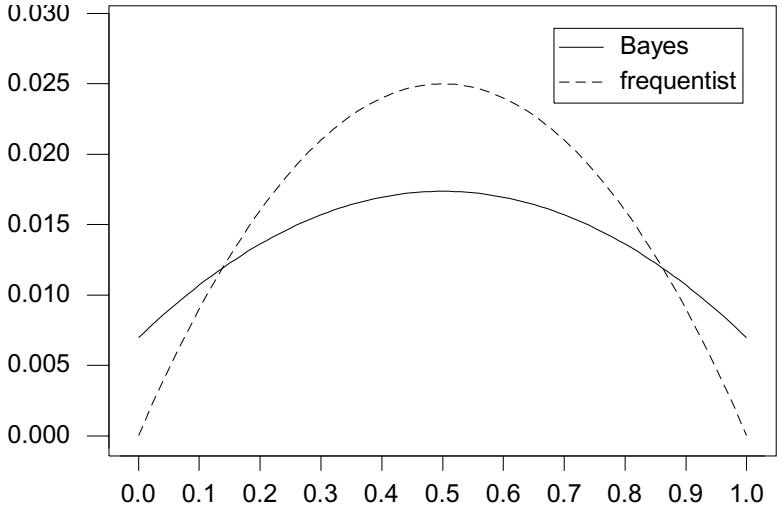


Figure 9.2 Mean squared error for the two estimators.

and

$$\begin{aligned} MS(\hat{\pi}_B) &= \left[\frac{1-2\times.5}{12}\right]^2 + \left[\frac{1}{12}\right]^2 \times 10 \times .5 \times .5 \\ &= .01736\,. \end{aligned}$$

We see that, on average (for these two values of π), the Bayesian posterior estimator is closer to the true value than the frequentist estimator. Figure 9.2 shows the mean squared error for the Bayesian estimator and the frequentist estimator as a function of π. We see that over most (but not all) of the range, the Bayesian estimator (using uniform prior) is better than the frequentist estimator.[4]

9.4 INTERVAL ESTIMATION

The second type of inference we consider is interval estimation. We wish to find an interval (l, u) that has a predetermined probability of containing the parameter. In the frequentist interpretation, the parameter is fixed but unknown, and before the sample is taken, the interval endpoints are random because they depend on the data. After the sample is taken, and the endpoints are calculated, there is nothing random,

[4]The frequentist estimator, $\hat{\pi}_f = \frac{y}{n}$, would be Bayesian posterior mean if we used the prior $g(\pi) \propto \pi^{-1}(1-\pi)^{-1}$. This prior is improper since it does not integrate to 1. An estimator is said to be admissible if no other estimator has smaller mean squared error over the whole range of possible values. Wald (1950) showed that Bayesian posterior mean estimators that arose from proper priors are always admissible. Bayesian posterior mean estimators from improper priors sometimes are admissible, as in this case.

so the interval is said to be a *confidence interval* for the parameter. We know that a predetermined proportion of intervals calculated for random samples using this method will contain the true parameter. But it doesn't say anything at all about the specific interval we calculate from our data.

In Chapter 8, we found a *Bayesian credible interval* for the parameter π that has the probability that we want. Because it is found from the posterior distribution, it has the coverage probability we want for this specific data.

Confidence Intervals

Confidence intervals are how frequentist statistics tries to find an interval has a high probability of containing the true value of the parameter θ. A $(1-\alpha)\times 100\%$ confidence interval for a parameter θ is an interval (l, u) such that

$$P(l \leq \theta \leq u) = 1 - \alpha\,.$$

This probability is found using the sampling distribution of an estimator for the parameter. There are many possible values of l and u that satisfy this. The most commonly used criteria for choosing them are (1) equal ordinates (heights) on the sampling distribution and (2) equal tail area on the sampling distribution. Equal ordinates will find the shortest confidence interval. However, the equal tail area intervals are often used because they are easier to find. When the sampling distribution of the estimator is symmetric, the two criteria will coincide.

The parameter θ is regarded as a fixed but unknown constant. The endpoints l and u are random variables since they depend on the random sample. When we plug in the actual sample data that occurred for our random sample and calculate the values for l and u, there is nothing left that is random. The interval either contains the unknown fixed parameter or it doesn't, and we don't know which is true. The interval can no longer be regarded as a probability interval.

Under the frequentist paradigm, the correct interpretation is that $(1-\alpha)\times 100\%$ of the *random* intervals calculated this way will contain the true value. Therefore we have $(1-\alpha)\times 100\%$ *confidence* that our interval does. It is a misinterpretation to make a probability statement about the parameter θ from the calculated confidence interval.

Often, the sampling distribution of the estimator used is approximately normal, with mean equal to the true value. In this case, the confidence interval has the form

$$\textit{estimator} \pm \textit{critical value} \times \textit{standard deviation of the estimator,}$$

where the critical value comes from the *standard normal table.* For example if n is large, then the sample proportion

$$\pi_f = \frac{y}{n}$$

is approximately normal with mean π and standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$. This gives an approximate $(1-\alpha)\times 100\%$ equal tail area confidence interval for π:

$$\pi_f \pm z_{\frac{\alpha}{2}} \times \sqrt{\frac{\pi_f(1-\pi_f)}{n}}\,. \tag{9.4}$$

Comparing Confidence and Credible Intervals for π

The probability calculations for the confidence interval are based on the sampling distribution of the statistic. In other words, how it varies over all possible samples. Hence the probabilities are *pre-data*. They do not depend on the particular sample that occurred. This is in contrast to the Bayesian credible interval calculated from the posterior distribution that has a direct (degree of belief) probability interpretation conditional on the observed sample data. The Bayesian credible interval is more useful to the scientist whose data we are analyzing. It summarizes our beliefs about the parameter values that could credibly be believed given the observed data that occurred. In other words, it is *post-data*. He/she is not concerned about data that could have occurred but did not.

Example 13 (continued from Chapter 8) *Out of a random sample of 100 Hamilton residents, y=26 said they support building a casino in Hamilton. A frequentist 95 % confidence interval for π is*

$$.26 \pm 1.96\sqrt{\frac{.26\times .74}{100}}$$

$$= (.174, .346)\,.$$

Compare this with the 95% credible intervals for π calculated by the three students in Chapter 8 and shown in Table 8.3.

9.5 HYPOTHESIS TESTING

The third type of inference we consider is hypothesis testing. Scientists do not like to claim the existence of an effect where the discrepancy in the data could be due to chance alone. If they make their claims too quickly, later studies would show their claim was wrong, and their scientific reputation would suffer.

Hypothesis testing, sometimes called significance testing[5], is the frequentist statistical method widely used by scientists to guard against making claims unjustified by the data. The nonexistence of the treatment effect is set up as the *null hypothesis* that "the shift in the parameter value caused by the treatment is zero." The competing

[5]Significance testing was developed by R. A. Fisher as an inferential tool to weigh the evidence against a particular hypothesis. Hypothesis testing was developed by Neymann and Pearson as a method to control the error rate in deciding between two competing hypotheses. These days, the two terms are used almost interchangeably, despite their differing goals and interpretations. This continues to cause confusion.

hypothesis that there is a nonzero shift in the parameter value caused by the treatment is called the *alternative hypothesis*. Two possible explanations for the discrepancy between the observed data and what would be expected under the null hypothesis are proposed.

1. The null hypothesis is true, and the discrepancy is due to random chance alone.

2. The null hypothesis is false. This causes at least part of the discrepancy.

To be consistent with Ockham's razor, we will stick with explanation (1), which has the null hypothesis being true and the discrepancy being due to chance alone, unless the discrepancy is so large that it is very unlikely to be due to chance alone. This means that when we accept the null hypothesis as true, it doesn't mean that we believe it is literally true. Rather, it means that chance alone remains a reasonable explanation for the observed discrepancy, so we can't discard chance as the sole explanation.

When the discrepancy is too large, we are forced to discard explanation (1) leaving us with explanation (2), that the null hypothesis is false. This gives us a backward way to establish the existence of an effect. We conclude the effect exists (the null hypothesis is false) whenever the probability of the discrepancy between what occurred and what would be expected under the null hypothesis is too small to be attributed to chance alone.

Because hypothesis testing is very well established in science, we will show how it can be done in a Bayesian manner. There are two situations we will look at. The first is testing a one-sided hypothesis where we are only interested in detecting the effect in one direction. We will see that in this case, Bayesian hypothesis testing works extremely well, without the contradictions required in frequentist tests. The Bayesian test of a one-sided null hypothesis is evaluated from the posterior probability of the null hypothesis.

The second situation is where we want to detect a shift in either direction. This is a two-sided hypothesis test, where we test a point hypothesis (that the effect is zero) against a two-sided alternative. The prior density of a continuous parameter measures probability density, not probability. The prior probability of the null hypothesis (shift equal to zero) must be equal to 0. So its posterior probability must also be zero,[6] and we cannot test a two-sided hypothesis using the posterior probability of the null hypothesis. Rather, we will test the *credibility* of the null hypothesis by seeing if the null value lies in the credible interval. If the null value does lie within the credible interval, we cannot reject the null hypothesis, because the null value remains a credible value.

[6]We are also warned that frequentist hypothesis tests of a point null hypothesis never "accept" the null hypothesis, rather, they "can't reject the null hypothesis."

9.6 TESTING A ONE-SIDED HYPOTHESIS

The effect of the treatment is included as a parameter in the model. The hypothesis that the treatment has no effect becomes the *null hypothesis* the parameter representing the treatment effect has the *null* value that corresponds to no effect of the treatment.

Frequentist Test of One-Sided Hypothesis

The probability of the data (or results even more extreme) given that the null hypothesis is true is calculated. If this is below a threshold called the *level of significance*, the results are deemed to be incompatible with the null hypothesis, and the null hypothesis is rejected at that level of significance. This establishes the existence of the treatment effect. This is similar to a "proof by contradiction." However, because of sampling variation, complete contradiction is impossible. Even very unlikely data are possible when there is no treatment effect. So hypothesis tests are actually more like "proof by low probability." The probability is calculated from the sampling distribution given the null hypothesis is true. This makes it a *pre-data* probability.

Example 14 *Suppose we wish to determine if a new treatment is better than the standard treatment. If so, π, the proportion of patients who benefit from the new treatment, should be better than π_0, the proportion who benefit from the standard treatment. It is known from historical records that $\pi_0 = .6$. A random group of 10 patients are given the new treatment. Y, the number who benefit from the treatment will be* binomial(n, π). *We observe $y = 8$ patients benefit. This is better than we would expect if $\pi = .6$. But, is it enough better for us to conclude that $\pi > .6$ at the 10% level of significance?*

The steps are:

1. *Set up a null hypothesis about the (fixed but unknown) parameter. For example, $H_0 : \pi \leq .6$. (The proportion who would benefit from the new treatment is less than or equal to the proportion who benefit from the standard treatment.) We include all π values less than the null value .6 in with the null hypothesis because we are trying to determine if the new treatment is better. We have no interest in determining if the new treatment is worse. We won't recommend it unless it is demonstrably better than the standard treatment.*

2. *The alternative hypothesis is $H_1 : \pi > .6$. (The proportion who would benefit from the new treatment is greater than the proportion who benefit from the standard treatment.)*

3. *The null distribution of the test statistic is the sampling distribution of the test statistic, given the null hypothesis is true. In this case, it will be binomial$(n, .6)$ where $n = 10$ is the number of patients given the new treatment.*

4. *We choose level of significance for the test to be as close as possible to $\alpha = 5\%$. Since y has a discrete distribution, only some values of α are possible, so we will have to choose a value either just above or just below 5%.*

Table 9.1 Null distribution of Y with a rejection region for a one-sided hypothesis test

Value	$f(y\|\pi = .6)$	Region
0	.0001	accept
1	.0016	accept
2	.0106	accept
3	.0425	accept
4	.1115	accept
5	.2007	accept
6	.2508	accept
7	.2150	accept
8	.1209	accept
9	.0403	reject
10	.0060	reject

5. *The rejection region is chosen so that it has a probability of α under the null distribution.*[7] *If we choose the rejection region $y \geq 9$, then the $\alpha = .0463$. The null distribution with the rejection region for the one-sided hypothesis test is shown in Table 9.1.*

6. *If the value of the test statistic for the given sample lies in the rejection region, then reject the null hypothesis H_0 at level α. Otherwise, we can't reject H_0. In this case, $y = 8$ was observed. This lies in the acceptance region.*

7. *The p-value is the probability of getting what we observed, or something even more unlikely, given the null hypothesis is true. The p-value is put forward as measuring the strength of evidence against the null hypothesis*[8]*. In this case, the p-value $= .1672$.*

8. *If the p-value $< \alpha$ the test statistic lies in the rejection region, and vice versa. So an equivalent way of testing the hypothesis is to reject if p-value $< \alpha$*[9] *Looking at it either way, we cannot reject the null hypothesis $H_0 : \pi \leq .6$. $y = .8$ lies in the acceptance region, and the p-value $> .05$. The evidence is not strong enough to conclude that $\pi > .6$.*

There is much confusion about the *p-value* of a test. It is *not* the posterior probability of the null hypothesis being true given the data. Instead, it is the tail

[7] This approach is from Neyman and Pearson

[8] This approach is from R. A. Fisher.

[9] Both α and p-value are tail areas calculated from the null distribution. However, α represents the long run rate of rejecting a true null hypothesis, and *p-value* is looked at as the evidence against *this particular null hypothesis by this particular data set*. Using tail areas as simultaneously representing both the long run and a particular result is inherently contradictory.

probability calculated using the null distribution. In the binomial case

$$\textit{p-value} = \sum_{y_{obs}}^{n} f(y|\pi_0)\,,$$

where y_{obs} is the observed value of y. Frequentist hypothesis tests use a probability calculated on all possible data sets that could have occurred (for the fixed parameter value), but the hypothesis is about the parameter value being in some range of values.

Bayesian Tests of a One-Sided Hypothesis

We wish to test

$$H_0 : \pi \leq \pi_0 \quad \text{versus} \quad H_1 : \pi > \pi_0$$

at the level of significance α using Bayesian methods. We can calculate the posterior probability of the null hypothesis being true by integrating the posterior density over the correct region:

$$P(H_0 : \pi \leq \pi_0|y) = \int_0^{\pi_0} g(\pi|y)\, d\pi\,. \tag{9.5}$$

We reject the null hypothesis if that posterior probability is less than the level of significance α. Thus a Bayesian one-sided hypothesis test is a "test by low probability" using the probability calculated directly from the posterior distribution of π. We are testing a hypothesis about the parameter using the posterior distribution of the parameter. Bayesian one-sided tests use *post-data* probability.

Example 14 (continued) *Suppose we use a* beta $(1, 1)$ *prior for* π. *Then given* $y = 8$, *the posterior density is* beta $(9, 3)$. *The posterior probability of the null hypothesis is*

$$\begin{aligned} P(\pi \leq .6|y = 8) &= \int_0^{.6} \frac{\Gamma(12)}{\Gamma(3)\Gamma(9)}\pi^2(1-\pi)^8 d\pi \\ &= .1189 \end{aligned}$$

when we evaluate it numerically. This is not less than .05, *so we cannot reject the null hypothesis at the 5% level of significance* 5%. *Figure 9.3 shows the posterior density. The probability of the null hypothesis is the area under the curve to the right of* $\pi = .6$.

9.7 TESTING A TWO-SIDED HYPOTHESIS

Sometimes we might want to detect a change in the parameter value in either direction. This is known as a two-sided test since we are wanting to detect any changes from the value π_0. We set this up as testing the point null hypothesis $H_0 : \pi = \pi_0$ against the alternative hypothesis $H_1 : \pi \neq \pi_0$.

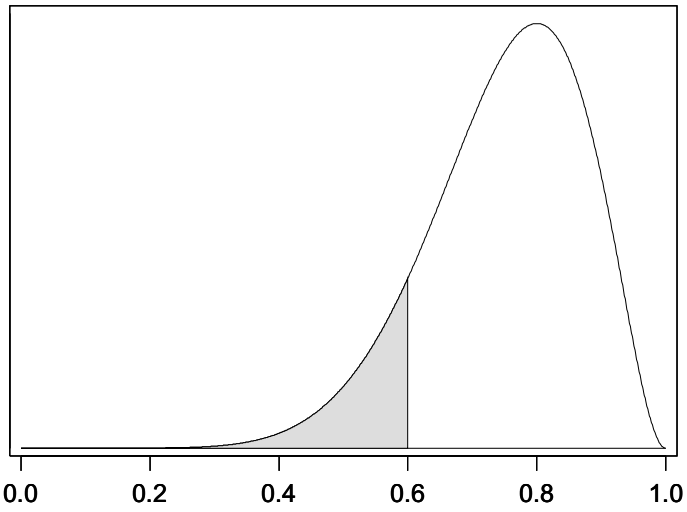


Figure 9.3 Posterior probability of null hypothesis.

Frequentist Test of a Two-Sided Hypothesis

The null distribution is evaluated at π_0, and the rejection region is two-sided, as are *p-values* calculated for this test.

Example 15 *A coin is tossed 15 times, and we observe 10 heads. Are 10 heads out of 15 tosses enough to determine that the coin is not fair? In other words, is π the probability of getting a head different than $\frac{1}{2}$?*

The steps are:

1. *Set up the null hypothesis about the fixed but unknown parameter π. It is $H_0 : \pi = .5$.*
2. *The alternative hypothesis is $H_1 : \pi \neq .5$. We are interested in determining a difference in either direction, so we will have a two-sided rejection region.*
3. *The null distribution is the sampling distribution of Y when the null hypothesis is true. It is binomial$(n = 15, \pi = .5)$.*
4. *Since Y has a discrete distribution, we choose the level of significance for the test to be as close to 5% as possible.*
5. *The rejection region is chosen so that it has a probability of α under the null distribution. If we choose rejection region $\{Y \leq 3\} \cup \{Y \geq 12\}$, then $\alpha = .0352$. The null distribution and rejection region for the two-sided hypothesis are shown in Table 9.2.*
6. *If the value of the test statistic lies in the rejection region, then we reject the null hypothesis H_0 at level α. Otherwise, we can't reject H_0. In this case,*

Table 9.2 Null distribution of Y with the rejection region for two-sided hypothesis test

Value	$f(y\|\pi = .5)$	Region
0	.0000	reject
1	.0005	reject
2	.0032	reject
3	.0139	reject
4	.0417	accept
5	.0916	accept
6	.1527	accept
7	.1964	accept
8	.1964	accept
9	.1527	accept
10	.0916	accept
11	.0417	accept
12	.0139	reject
13	.0032	reject
14	.0005	reject
15	.0000	reject

$y = 10$ *was observed. This lies in the region where we can't reject the null hypothesis. We must conclude that chance alone is sufficient to explain the discrepancy, so* $\pi = .5$ *remains a reasonable possibility.*

7. *The p-value is the probability of getting what we got (10) or something more unlikely, given the null hypothesis* H_0 *is true. In this case we have a two-sided alternative, so the p-value is the* $P(Y \geq 10) + P(Y \leq 5) = .274$. *This is larger than* α*, so we can't reject the null hypothesis.*

Relationship between two-sided hypothesis tests and confidence intervals. While the null value of the parameter usually comes from the idea of no treatment effect, it is possible to test other parameter values. There is a close relationship between two-sided hypothesis tests and confidence intervals. If you are testing a two-sided hypothesis at level α, there is a corresponding $(1 - \alpha) \times 100\%$ confidence interval for the parameter. If the null hypothesis

$$H_0 : \pi = \pi_0$$

is rejected, then the value π_0 lies outside the confidence interval, and vice versa. If the null hypothesis is accepted (can't be rejected), then π_0 lies inside the confidence

interval, and vice versa. The confidence interval "summarizes" all possible null hypotheses that would be accepted if they were tested.

Bayesian Test of a Two-Sided Hypothesis

From the Bayesian perspective, the posterior distribution of the parameter given the data sums up our entire belief after the data. However, the idea of hypothesis testing as a protector of scientific credibility is well established in science. So we look at using the posterior distribution to test a point null hypothesis versus a two-sided alternative in a Bayesian way.

If we use a continuous prior, we will get a continuous posterior. The probability of the exact value represented by the point null hypothesis will be zero. We can't use posterior probability to test the hypothesis. Instead, we use a correspondence similar to the one between confidence intervals and hypothesis tests, but with credible interval instead.

Compute a $(1-\alpha)\times 100\%$ credible interval for π. If π_0 lies inside the credible interval, accept (do not reject) the null hypothesis $H_0 : \pi = \pi_0$, and if π_0 lies outside the credible interval, then reject the null hypothesis.

Example 15 (continued) *If we use a* uniform *prior distribution, the posterior is the* beta*(10+1,5+1) distribution. A 95% Bayesian credible interval for π found using the normal approximation is*

$$\frac{11}{17} + 1.96 \times \sqrt{\frac{11\times 6}{((11+6)^2\times(11+6+1))}}$$

$$= .647 \pm .221 \;=\; (.426, .868)\,.$$

The null value $\pi = .5$ lies within the credible interval, so we cannot reject the null hypothesis. It remains a credible value.

Main Points

- The posterior distribution of the parameter given the data is the entire inference from a Bayesian perspective. Probabilities calculated from the posterior distribution are *post-data* because the posterior distribution is found after the observed data has been taken into the analysis.
- Under the frequentist perspective there are specific inferences about the parameter: point estimation, confidence intervals, and hypothesis tests.
- Frequentist statistics considers the parameter a fixed but unknown constant. The only kind of probability allowed is long run relative frequency.
- The sampling distribution of a statistic is its distribution over all possible random samples given the fixed parameter value. Frequentist statistics is based on the sampling distribution.

- Probabilities calculated using the sampling distribution are *pre-data* because they are based on all possible random samples, not the specific random sample we obtained.
- An estimator of a parameter is unbiased if its expected value calculated from the sampling distribution is the true value of the parameter.
- Frequentist statistics often call the minimum variance unbiased estimator the *best* estimator.
- The mean squared error of an estimator measures its average squared distance from the true parameter value. It is the square of the bias plus the variance.
- Bayesian estimators are often better than frequentist estimators even when judged by the frequentist criteria such as mean squared error.
- Seeing how a Bayesian estimator performs using frequentist criteria for a range of possible parameter values is called a *pre-posterior* analysis, because it can be done before we obtain the data.
- A $(1-\alpha)\times 100\%$ confidence interval for a parameter θ is an interval (l,u) such that
$$P(l \leq \theta \leq u) = 1-\alpha\,,$$
where the probability is found using the sampling distribution of an estimator for θ. The correct interpretation is that $(1-\alpha)\times 100\%$ of the random intervals calculated this way do contain the true value. When the actual data are put in and the endpoints calculated, there is nothing left to be random. The endpoints are numbers; the parameter is fixed but unknown. We say that we are $(1-\alpha)\times 100\%$ *confident* that the calculated interval covers the true parameter. The confidence comes from our belief in the method used to calculate the interval. It does not say anything about the actual interval we got for that particular data set.
- A $(1-\alpha)\times 100\%$ Bayesian credible interval for θ is a range of parameter values that has posterior probability $(1-\alpha)$.
- Frequentist hypothesis testing is used to determine whether the actual parameter could be a specific value. The sample space is divided into a rejection region and an acceptance region such that the probability the test statistic lies in the rejection region if the null hypothesis is true is less than the level of significance α. If the test statistic falls into the rejection region, we reject the null hypothesis at level of significance α.
- Or we could calculate the *p-value*. If the *p-value*$<\alpha$, we reject the null hypothesis at level α.
- The *p-value* is not the probability the null hypothesis is true. Rather, it is the probability of observing what we observed, or even something more extreme, given that the null hypothesis is true.

- We can test a one-sided hypothesis in a Bayesian manner by computing the posterior probability of the null hypothesis. This probability is found by integrating the posterior density over the null region. If this probability is less than the level of significance α, then we reject the null hypothesis.
- We cannot test a two-sided hypothesis by integrating the posterior probability over the null region because, with a continuous prior, the prior probability of a point null hypothesis is zero, so the posterior probability will also be zero. Instead, we test the credibility of the null value by observing whether or not it lies within the Bayesian credible interval. If it does, the null value remains credible and we can't reject it.

Exercises

9.1 Let π be the proportion of students at a university who approve the governments policy on students allowances. The students newspaper is going to take a random sample of $n = 30$ students at a university and ask if they approve of the governments policy on student allowances.

(a) What is the distribution of y, the number who answer "yes"?

(b) Suppose out of the 30 students, 8 answered yes. What is the *frequentist* estimate of π.

(c) Find the posterior distribution $g(\pi|y)$ if we use a uniform prior.

(d) What would be the Bayesian estimate of π?

9.2 The standard method of screening for a disease fails to detect the presence of the disease in 15% of the patients who actually do have the disease. A new method of screening for the presence of the disease has been developed. A random sample of $n = 75$ patients who are known to have the disease is screened using the new method. Let π be the probability the new screening method fails to detect the disease.

(a) What is the distribution of y, the number of times the new screening method fails to detect the disease?

(b) Of these $n = 75$ patients, the new method failed to detect the disease in $y = 6$ cases. What is the frequentist estimator of π?

(c) Use a *beta* $(1, 6)$ prior for π. Find $g(\pi|y)$, the posterior distribution of π.

(d) Find the posterior mean and variance.

(e) If $\pi \geq .15$, then the new screening method is no better than the standard method. Test

$$H_0 : \pi \geq .15 \quad \text{versus} \quad H_1 : \pi < .15$$

at the 5% level of significance in a Bayesian manner.

9.3 In the study of water quality in New Zealand streams documented in McBride et al. (2002) a high level of Campylobacter was defined as a level greater than 100 per 100 ml of stream water. $n = 116$ samples were taken from streams having a high environmental impact from birds. Out of these $y = 11$ had a high Campylobacter level. Let π be the true probability that a sample of water from this type of stream has a high Campylobacter level.

(a) Find the frequentist estimator for π.

(b) Use a *beta* $(1, 10)$ prior for π. Calculate the posterior distribution $g(\pi|y)$.

(c) Find the posterior mean and variance. What is the Bayesian estimator for π?

(d) Find a 95% Credible interval for π.

(e) Test the hypothesis

$$H_0 : \pi = .10 \quad \text{versus} \quad H_1 : \pi \neq .10$$

at the 5% level of significance.

9.4 In the same study of water quality, $n = 145$ samples were taken from streams having a high environmental impact from dairying. Out of these $y = 9$ had a high Campylobacter level. Let π be the true probability that a sample of water from this type of stream has a high Campylobacter level.

(a) Find the frequentist estimator for π.

(b) Use a *beta* $(1, 10)$ prior for π. Calculate the posterior distribution $g(\pi|y)$.

(c) Find the posterior mean and variance. What is the Bayesian estimator for π?

(d) Find a 95% Credible interval for π.

(e) Test the hypothesis

$$H_0 : \pi = .10 \qquad \text{versus} \qquad H_1 : \pi \neq .10$$

at the 5% level of significance.

9.5 In the same study of water quality, $n = 176$ samples were taken from streams having a high environmental impact from sheep farming. Out of these $y = 24$ had a high Campylobacter level. Let π be the true probability that a sample of water from this type of stream has a high Campylobacter level.

(a) Find the frequentist estimator for π.

(b) Use a *beta* $(1, 10)$ prior for π. Calculate the posterior distribution $g(\pi|y)$.

(c) Find the posterior mean and variance. What is the Bayesian estimator for π?

(d) Test the hypothesis

$$H_0 : \pi \geq .15 \quad \text{versus} \quad H_1 : \pi < .15$$

at the 5% level of significance.

9.6 In the same study of water quality, $n = 87$ samples were taken from streams in municipal catchments. Out of these $y = 8$ had a high Campylobacter level. Let π be the true probability that a sample of water from this type of stream has a high Campylobacter level.

(a) Find the frequentist estimator for π.

(b) Use a *beta* $(1, 10)$ prior for π. Calculate the posterior distribution $g(\pi|y)$.

(c) Find the posterior mean and variance. What is the Bayesian estimator for π?

(d) Test the hypothesis

$$H_0 : \pi \geq .10 \quad \text{versus} \quad H_1 : \pi < .10$$

at the 5% level of significance.

Monte Carlo Exercises

9.1 **Comparing Bayesian and frequentist estimators for** π. In Chapter 1 we learned that the frequentist procedure for evaluating a statistical procedure, namely looking at how it performs in the long run, for a (range of) fixed but unknown parameter values can also be used to evaluate a Bayesian statistical procedure. This "what if the parameter has this value " type of analysis would be done before we obtained the data and is called a *pre-posterior* analysis. It evaluates the procedure by seeing how it performs over all possible random samples, given that parameter value. In Chapter 8 we found that the posterior mean used as a Bayesian estimator minimizes the posterior mean squared error. Thus it has optimal post-data properties, in other words after making use of the actual data. We will see that Bayesian estimators have excellent pre-data (frequentist) properties as well, often better than the corresponding frequentist estimators.

We will perform a Monte Carlo study approximating the sampling distribution of two estimators of π. The frequentist estimator we will use is $\hat{\pi}_f = \frac{y}{n}$, the sample proportion. The Bayesian estimator we will use is $\hat{\pi}_B = \frac{y+1}{n+1}$ which equals the posterior mean when we used a uniform prior for π. We will compare the sampling distributions (in terms of bias, variance, and mean squared error) of the two estimators over a range of π values from 0 to 1. However, unlike the exact analysis we did in Section 9.3, here we will do a Monte Carlo study. For each of the parameter values, we will approximate the sampling distribution

of the estimator by an empirical distribution based on 5000 samples drawn when that is the parameter value. The true characteristics of the sampling distribution (mean, variance, mean squared error) are approximated by the sample equivalent from the empirical distribution. You can use either Minitab or R for your analysis.

(a) For $\pi = .1, .2, \ldots, .9$

i. Draw 5000 random samples from *binomial* $(n = 10, \pi)$.

ii. Calculate the frequentist estimator $\hat{\pi}_f = \frac{y}{n}$ for each of the 5000 samples.

iii. Calculate the Bayesian estimator $\hat{\pi}_B = \frac{y+1}{n+2}$ for each of the 5000 samples.

iv. Calculate the means of these estimators over the 5000 samples, and subtract π to give the biases of the two estimators. Note that this is a function of π.

v. Calculate the variances of these estimators over the 5000 samples. Note that this is also a function of π.

vi. Calculate the mean squared error of these estimators over the 5000 samples. The first way is

$$MS(\hat{\pi}) = (bias(\hat{\pi}))^2 + Var(\hat{\pi})\,.$$

The second way is to take the sample mean of the squared distance the estimator is away from the true value over all 5000 samples. Do it both ways, and see that they give the same result.

(b) Plot the biases of the two estimators versus π at those values and connect the adjacent points. (Put both estimators on the same graph.)

i. Does the frequentist estimator appear to be unbiased over the range of π values?

ii. Does the Bayesian estimator appear to be unbiased over the range of the π values?

(c) Plot the mean squared errors of the two estimators versus π over the range of π values, connecting adjacent points. (Put both estimators on the same graph.)

i. Does your graph resemble Figure 9.2?

ii. Over what range of π values does the Bayesian estimator have smaller mean squared error than that of the frequentist estimator?

10

Bayesian Inference for Normal Mean

Many random variables seem to follow the normal distribution, at least approximately. The reasoning behind the central limit theorem suggests why this is so. Any random variable that is the sum of a large number of similar sized random variables from independent causes will be approximately normal. The shapes of the individual random variables "average out" to the normal shape. Sample data from the sum distribution will be well approximated by a normal. The most widely used statistical methods are those that have been developed for random samples from a normal distribution. In this chapter we show how Bayesian inference on a random sample from a normal distribution is done.

10.1 BAYES' THEOREM FOR NORMAL MEAN WITH A DISCRETE PRIOR

For a Single Normal Observation

We are going to take a single observation from the conditional density $f(y|\mu)$ that is known to be normal with known variance σ^2. The standard deviation, σ, is the square root of the variance. There are only m possible values $\mu_1, \cdots, \mu_m$ for the mean. We choose a discrete prior probability distribution over these values, which summarizes our prior belief about the parameter, before we take the observation. If we really don't have any prior information, we would give all values equal prior probability.

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

We only need to choose the prior probabilities up to a multiplicative constant, since is only the relative weights we give to the possible values that are important.

The likelihood gives relative weights to all the possible parameter values according to how likely the observed value was given each parameter value. It looks like the conditional observation distribution given the parameter, μ, but instead of the parameter being fixed and the observation varying, we fix the observation at the one that actually occurred, and vary the parameter over all possible values. We only need to know it up to a multiplicative constant since the *relative weights* are all that is needed to apply Bayes' theorem. The posterior is proportional to prior times likelihood, so it equals

$$g(\mu|y) = \frac{prior \times likelihood}{\sum prior \times likelihood}\,.$$

Any multiplicative constant in either the prior or likelihood would cancel out.

Likelihood of Single Observation

The conditional observation distribution of $y|\mu$ is normal with mean μ and variance σ^2, which is known. Its density is

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}\,.$$

The likelihood of each parameter value is the value of the observation distribution at the observed value. The part that doesn't depend on the parameter μ is the same for all parameter values, so it can be absorbed into the proportionality constant. The part that gives the shape as a function of the parameter μ is the important part. Thus the likelihood shape is given by

$$f(y|\mu) \propto e^{-\frac{1}{2\sigma^2}(y-\mu)^2}\,, \tag{10.1}$$

where y is held constant at the observed value and μ is allowed to vary over all possible values.

Table for Performing Bayes' Theorem

We set up a table to help us find the posterior distribution using Bayes' theorem. The first and second columns contain the possible values of the parameter μ and their prior probabilities. The third column contains the likelihood, which is the observation distribution evaluated for each of the possible values μ_i where y is held at the observed value. This puts a weight on each possible value μ_i proportional to the probability of getting the value actually observed if μ_i is the parameter value. There are two methods we can use to evaluate the likelihood.

Table 10.1 Method 1: Finding posterior using likelihood from Table B.3 "ordinates of normal distribution"

μ	Prior	z	Likelihood	Prior $\times$ Likelihood	Posterior
2.0	.2	-1.2	.1942	.03884	.1238
2.5	.2	-.7	.3123	.06246	.1991
3.0	.2	-.2	.3910	.0782	.2493
3.5	.2	.3	.3814	.07628	.2431
4.0	.2	.8	.2897	.05794	.1847
				.31372	1.00

Finding likelihood from the "ordinates of normal distribution" table. The first method is to find the likelihood from the "ordinates of the normal distribution" table. Let

$$z = \frac{y - \mu}{\sigma}$$

for each possible value of μ. Z has a standardized *normal* $(0, 1)$ distribution. The likelihood can be found by looking up $f(z)$ in the "ordinates of the standard normal distribution" given in Table B.3 in Appendix B. Note that $f(-z) = f(z)$ because of standard normal distribution is symmetric about 0 .

Finding the likelihood from the normal density function. The second method is to use the normal density formula given in Equation 10.1, holding y fixed at the observed value and varying μ over all possible values.

Example 16 *Suppose $y|\mu$ is normal with mean μ and known variance $\sigma^2 = 1$. We know there are only five possible values for μ. They are 2.0, 2.5, 3.0, 3.5, and 4. We let them be equally likely for our prior. We take a single observation of y and obtain the value $y = 3.2$. Let*

$$z = \frac{y - \mu}{\sigma}\,.$$

The values for the likelihood $f(z)$ are found in Table B.3, "ordinates of normal distribution," in Appendix B. Note that $f(-z) = f(z)$ because of standard normal density is symmetric about 0. The posterior probability is the prior $\times$ likelihood divided by sum of prior $\times$ likelihood. The results are shown in Table 10.1.

If we evaluate the likelihood using the normal density formula, the likelihood is proportional to

$$e^{-\frac{1}{2\sigma^2}(y-\mu)^2}\,,$$

where y is held at 3.2 and μ varies over all possible values. Note, we are absorbing everything that doesn't depend on μ into the proportionality constant. The posterior probability is the prior $\times$ likelihood divided by sum of prior $\times$ likelihood. The results are shown in Table 10.2. We note that the results agree with what we found before except for small round-off errors.

Table 10.2 Method 2: Finding posterior using likelihood from normal density formula

μ	Prior	Likelihood *(ignoring constant)*	Prior × Likelihood	Posterior
2.0	.2	$e^{-\frac{1}{2}(3.2-2.0)^2}$ =.4868	.0974	.1239
2.5	.2	$e^{-\frac{1}{2}(3.2-2.5)^2}$ =.7827	.1565	.1990
3.0	.2	$e^{-\frac{1}{2}(3.2-3.0)^2}$ =.9802	.1960	.2493
3.5	.2	$e^{-\frac{1}{2}(3.2-3.5)^2}$ =.9560	.1912	.2432
4.0	.2	$e^{-\frac{1}{2}(3.2-4.0)^2}$ =.7261	.1452	.1846
			.7863	1.00

For a Random Sample of Normal Observations

Usually we have a random sample $y_1, \ldots, y_n$ of observations instead of a single observation. The posterior is always proportional to the prior × likelihood. The observations in a random sample are all independent of each other, so the joint likelihood of the sample is the product of the individual observation likelihoods. This gives

$$f(y_1, \ldots, y_n|\mu) = f(y_1|\mu) \times f(y_2|\mu) \times \cdots \times f(y_n|\mu)\,.$$

Thus given a random sample[1], Bayes' theorem with a discrete prior is given by

$$g(\mu|y_1, \ldots, y_n) \propto g(\mu) \times f(y_1|\mu) \times \ldots \times f(y_n|\mu)$$

We are considering the case where the distribution of each observation $y_j|\mu$ is normal with mean μ and variance σ^2, which is known.

Finding the posterior probabilities analyzing observations one at a time. We could analyze the observations one at a time, in sequence $y_1, \ldots, y_n$, letting the posterior from the previous observation become the prior for the next observation. The likelihood of a single observation y_j is the column of values of the observation distribution at each possible parameter value at that observed value. The posterior is proportional to prior times likelihood.

Example 17 *Suppose we take a random sample of four observations from a normal distribution having mean μ and known variance $\sigma^2 = 1$. The observations are 3.2, 2.2, 3.6, and 4.1.*

[1]De Finetti introduced a condition weaker than independence called exchangeability. Observations are exchangeable if the conditional density of the sample $f(y_1, \ldots, y_n)$ is the unchanged for any permutation of the subscripts. In other words, the order the observations were taken has no useful information. De Finetti (1991) shows that when the observations are exchangeable, $f(y_1, \ldots, y_n) = \int v(\theta) \times w(y_1|\theta) \times w(y_n|\theta)\, d\theta$, for some parameter θ where $v(\theta)$ is some prior distribution and $w(y|\theta)$ is some conditional distribution. The observations are conditionally independent given θ. The posterior $g(\theta) \propto v(\theta) \times w(y_1|\theta) \times w(y_n|\theta)$. This allows us to treat the exchangeable observations as if they come from a random sample.

The possible values of μ are 2.0, 2.5, 3.0, 3.5, and 4.0. Again, we will use the prior that gives them all equal weight. We want to use Bayes' theorem to find our posterior belief about μ given the whole random sample. The posterior equals

$$g(\mu|y) = \frac{prior \times likelihood}{\sum prior \times likelihood}.$$

The results of analyzing the observations one at a time are shown in Table 10.3. This is clearly a lot of work for a large sample. We will see that it is much easier to use the whole sample together.

Finding the posterior probabilities analyzing the sample all at once. The posterior is proportional to the prior $\times$ likelihood, and the joint likelihood of the sample is the product of the individual observation likelihoods. Each observation is normal, so it has a normal likelihood. This gives the joint likelihood

$$f(y_1, \ldots, y_n|\mu) \propto e^{-\frac{1}{2\sigma^2}(y_1-\mu)^2} \times e^{-\frac{1}{2\sigma^2}(y_2-\mu)^2} \times \cdots \times e^{-\frac{1}{2\sigma^2}(y_n-\mu)^2}.$$

Adding the exponents gives

$$f(y_1, \ldots, y_n|\mu) \propto e^{-\frac{1}{2\sigma^2}\left[(y_1-\mu)^2+(y_2-\mu)^2+\cdots+(y_n-\mu)^2\right]}.$$

We look at the term in brackets

$$[(y_1-\mu)^2 + \cdots + (y_n-\mu)^2] = y_1^2 - 2y_1\mu + \mu^2 + \cdots + y_n^2 - 2y_n\mu + \mu^2$$

and combine similar terms to get

$$= (y_1^2 + \cdots + y_n^2) - 2\mu(y_1 + \cdots + y_n) + n\mu^2.$$

When we substitute this back in, factor out n, and complete the square we get

$$\begin{aligned} f(y_1, \ldots, y_n|\mu) &\propto e^{-\frac{n}{2\sigma^2}\left[\mu^2-2\mu\bar{y}+\bar{y}^2-\bar{y}^2+\frac{y_1^2+\cdots+y_n^2}{n}\right]} \\ &\propto e^{-\frac{n}{2\sigma^2}\left[\mu^2-2\mu\bar{y}+\bar{y}^2\right]} \times e^{-\frac{n}{2\sigma^2}\left[\frac{y_1^2+\cdots+y_n^2}{n}-\bar{y}^2\right]}. \end{aligned}$$

The likelihood of the normal random sample $y_1, \ldots, y_n$ is proportional to the likelihood of the sample mean $\bar{y}$. When we absorb the part that doesn't involve μ into the proportionality constant we get

$$f(y_1, \ldots, y_n|\mu) \propto e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2}.$$

We recognize that this likelihood has the shape of a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. We know $\bar{y}$, the sample mean, is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. So the joint likelihood of the random sample is proportional to the likelihood of the sample mean, which is

$$f(\bar{y}|\mu) \propto e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2}. \tag{10.2}$$

Table 10.3 Analyzing observations one at a time [2]

μ	Prior_1	Likelihood_1 *(ignoring constant)*	$\text{prior}_1 \times \text{likelihood}_1$	posterior_1
2.0	.2	$e^{-\frac{1}{2}(3.2-2.0)^2}$=.4868	.0974	.1239
2.5	.2	$e^{-\frac{1}{2}(3.2-2.5)^2}$=.7827	.1565	.1990
3.0	.2	$e^{-\frac{1}{2}(3.2-3.0)^2}$=.9802	.1960	.2493
3.5	.2	$e^{-\frac{1}{2}(3.2-3.5)^2}$=.9560	.1912	.2432
4.0	.2	$e^{-\frac{1}{2}(3.2-4.0)^2}$=.7261	.1452	.1846
			.7863	
μ	Prior_2	Likelihood_2 *(ignoring constant)*	$\text{prior}_2 \times \text{likelihood}_2$	posterior_2
2.0	.1239	$e^{-\frac{1}{2}(2.2-2.0)^2}$=.9802	.1214	.1916
2.5	.1990	$e^{-\frac{1}{2}(2.2-2.5)^2}$=.9560	.1902	.3002
3.0	.2493	$e^{-\frac{1}{2}(2.2-3.0)^2}$=.7261	.1810	.2857
3.5	.2432	$e^{-\frac{1}{2}(2.2-3.5)^2}$=.4296	.1045	.1649
4.0	.1846	$e^{-\frac{1}{2}(2.2-4.0)^2}$=.1979	.0365	.0576
			.6336	
μ	Prior_3	Likelihood_3 *(ignoring constant)*	$\text{prior}_3 \times \text{likelihood}_3$	posterior_3
2.0	.1916	$e^{-\frac{1}{2}(3.6-2.0)^2}$=.2780	.0533	.0792
2.5	.3002	$e^{-\frac{1}{2}(3.6-2.5)^2}$=.5461	.1639	.2573
3.0	.2857	$e^{-\frac{1}{2}(3.6-3.0)^2}$=.8353	.2386	.3745
3.5	.1649	$e^{-\frac{1}{2}(3.6-3.5)^2}$=.9950	.1641	.2576
4.0	.0576	$e^{-\frac{1}{2}(3.6-4.0)^2}$=.9231	.0532	.0835
			.6731	
μ	Prior_4	Likelihood_4 *(ignoring constant)*	$\text{prior}_4 \times \text{likelihood}_4$	posterior_4
2.0	.0792	$e^{-\frac{1}{2}(4.1-2.0)^2}$=.1103	.0087	.0149
2.5	.2573	$e^{-\frac{1}{2}(4.1-2.5)^2}$=.2780	.0715	.1226
3.0	.3745	$e^{-\frac{1}{2}(4.1-3.0)^2}$=.5461	.2045	.3508
3.5	.2576	$e^{-\frac{1}{2}(4.1-3.5)^2}$=.8352	.2152	.3691
4.0	.0835	$e^{-\frac{1}{2}(4.1-4.0)^2}$=.9950	.0838	.1425
			.5830	1.0000

[2]Note: the prior for observation i is the posterior after previous observation $i-1$.

Table 10.4 Analyze the observations all together using likelihood of sample mean

μ	Prior_1	$\text{Likelihood}_{\bar{y}}$	$\text{Prior}_1 \times \text{Likelihood}_{\bar{y}}$	$\text{Posterior}_{\bar{y}}$
2.0	.2	$e^{-\frac{1}{2\times 1/4}(3.275-2.0)^2}$=.0387	.0077	.0157
2.5	.2	$e^{-\frac{1}{2\times 1/4}(3.275-2.5)^2}$=.3008	.0602	.1228
3.0	.2	$e^{-\frac{1}{2\times 1/4}(3.275-3.0)^2}$=.8596	.1719	.3505
3.5	.2	$e^{-\frac{1}{2\times 1/4}(3.275-3.5)^2}$=.9037	.1807	.3685
4.0	.2	$e^{-\frac{1}{2\times 1/4}(3.275-4.0)^2}$=.3495	.0699	.1425
			.4904	1.000

We can think of this as drawing a single value, $\bar{y}$, the sample mean, from the normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. This will make analyzing the random sample much easier.

We substitute in the observed value of $\bar{y}$, the sample mean, and calculate its likelihood. Then we just find the posterior probabilities using Bayes' theorem in only one table. This is much less work !

Example 17 (continued) *In the preceding sample the mean $\bar{y}$ = 3.275. We use the likelihood of $\bar{y}$ which is proportional to the likelihood of the whole sample. The results are shown in Table 10.4. We see that they agree with the previous results to three figures. The slight discrepancy in the fourth decimal place is due to the accumulation of round off errors when we analyze the observations one at a time. It is clearly easier to use $\bar{y}$ to summarize the sample, and perform the calculations for Bayes' theorem only once.*[3]

10.2 BAYES' THEOREM FOR NORMAL MEAN WITH A CONTINUOUS PRIOR

We have a random sample $y_1, \cdots, y_n$ from a normal distribution with mean μ and known variance σ^2. It is more realistic to believe that all values of μ are possible, at least all those in an interval. This means we should use a continuous prior. We know that Bayes' theorem can be summarized as *posterior proportional to prior times likelihood*

$$g(\mu|y_1, \cdots, y_n) \propto g(\mu) \times f(y_1, \cdots, y_n|\mu)\,.$$

Here we allow $g(\mu)$ to be a continuous prior density. When the prior was discrete, we evaluated the posterior by dividing the *prior* $\times$ *likelihood* by the *sum* of the *prior* $\times$ *likelihood* over all possible parameter values. Integration for continuous variables

[3] $\bar{y}$ is said to be a sufficient statistic for the parameter μ. The likelihood of a random sample $y_1, \cdots, y_n$ can be replaced by the likelihood of a single statistic only if the statistic is sufficient for the parameter. One-dimensional sufficient statistics only exist for some distributions, notably those that come from the one-dimensional exponential family.

is analogous to summing for discrete variables. Hence we can evaluate the posterior by dividing the *prior* $\times$ *likelihood* by the *integral* of the *prior* $\times$ *likelihood* over the whole range of possible parameter values. Thus

$$g(\mu|y_1, \cdots, y_n) = \frac{g(\mu) \times f(y_1, \cdots, y_n|\mu)}{\int g(\mu) \times f(y_1, \cdots, y_n|\mu)\, d\mu}\,. \tag{10.3}$$

For a normal distribution, the likelihood of the random sample is proportional to the likelihood of the sample mean, $\bar{y}$. So

$$g(\mu|y_1, \cdots, y_n) = \frac{g(\mu) \times e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2}}{\int g(\mu) \times e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2} d\mu}\,.$$

This works for any continuous prior density $g(\mu)$. However, it requires an integration, which may have to be done numerically. We will look at some special cases where we can find the posterior without having to do the integration. For these cases, we have to be able to recognize when a density must be normal from the shape given in Equation 10.1.

Flat Prior Density for μ

We know that the actual values the prior gives to each possible value is not important. Multiplying all the values of the prior by the same constant would multiply the integral of the prior times likelihood by the same constant, so it would cancel out, and we would obtain the same posterior. What is important is that the prior gives the *relative* weights to all possible values that we believe before looking at the data.

The flat prior gives each possible value of μ equal weight. It does not favor any value over any other value, $g(\mu) = 1$. The flat prior is not really a proper prior distribution since $-\infty < \mu < \infty$, so it can't integrate to 1. Nevertheless, this *improper* prior works out all right. Even though the prior is improper, the posterior will integrate to 1, so it is proper.

A single normal observation y. Let y be a normally distributed observation with mean μ and known variance σ^2. The likelihood

$$f(y|\mu) \propto e^{-\frac{1}{2\sigma^2}(y-\mu)^2}\,,$$

if we ignore the constant of proportionality. Since the prior always equals 1, the posterior is proportional to this. Rewrite it as

$$g(\mu|y) \propto e^{-\frac{1}{2\sigma^2}(\mu-y)^2}\,.$$

We recognize from this shape that the posterior is a normal distribution with mean y and variance σ^2.

A normal random sample $y_1, \ldots y_n$. In the previous section we showed that the likelihood of a random sample from a normal distribution is proportional to likelihood of the sample mean $\bar{y}$. We know that $\bar{y}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. Hence the likelihood has shape given by

$$f(\bar{y}|\mu) \propto e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2},$$

where we are ignoring the constant of proportionality. Since the prior always equals 1, the posterior is proportional to this. Rewrite it as

$$g(\mu|\bar{y}) \propto e^{-\frac{1}{2\sigma^2/n}(\mu-\bar{y})^2}.$$

We recognize from this shape that the posterior distribution is normal with mean $\bar{y}$ and variance $\frac{\sigma^2}{n}$.

Normal Prior Density for μ

Single observation. The observation y is a random variable taken from a normal distribution with mean μ and variance σ^2 which is assumed known. We have a prior distribution that is normal with mean m and variance s^2. The shape of the prior density is given by

$$g(\mu) \propto e^{-\frac{1}{2s^2}(\mu-m)^2},$$

where we are ignoring the part that doesn't involve μ because multiplying the prior by any constant of proportionality will cancel out in the posterior. The shape of the likelihood is

$$f(y|\mu) \propto e^{-\frac{1}{2\sigma^2}(y-\mu)^2},$$

where we ignore the part that doesn't depend on μ because multiplying the likelihood by any constant will cancel out in the posterior. The prior times likelihood is

$$g(\mu) \times f(y|\mu) \propto e^{-\frac{1}{2}\left[\frac{(\mu-m)^2}{s^2}+\frac{(y-\mu)^2}{\sigma^2}\right]}.$$

Putting the terms in exponent over the common denominator and expanding them out gives

$$\propto e^{-\frac{1}{2}\left[\frac{\sigma^2(\mu^2-2\mu m+m^2)+s^2(y^2-2y\mu+\mu^2)}{\sigma^2 s^2}\right]}.$$

We combine the like terms

$$\propto e^{-\frac{1}{2}\left[\frac{(\sigma^2+s^2)\mu^2-2(\sigma^2 m+s^2 y)\mu+m^2\sigma^2+y^2 s^2}{\sigma^2 s^2}\right]}$$

and factor out $(\sigma^2+s^2)/(\sigma^2 s^2)$. Completing the square and absorbing the part that doesn't depend on μ into the proportionality constant, we have

$$\propto e^{-\frac{1}{2\sigma^2 s^2/(\sigma^2+s^2)}\left[\mu^2-2\frac{(\sigma^2 m+s^2 y)}{\sigma^2+s^2}\mu+(\frac{(\sigma^2 m+s^2 y)}{\sigma^2+s^2})^2\right]}$$

$$\propto e^{-\frac{1}{2\sigma^2 s^2/(\sigma^2+s^2)}\left[\mu-\frac{(\sigma^2 m+s^2 y)}{\sigma^2+s^2}\right]^2}.$$

We recognize from this shape that the posterior is a normal distribution having mean and variance given by

$$m' = \frac{(\sigma^2 m + s^2 y)}{\sigma^2 + s^2} \quad \text{and} \quad (s')^2 = \frac{\sigma^2 s^2}{(\sigma^2 + s^2)} \tag{10.4}$$

respectively. We started with a *normal*(m, s^2) prior, and ended up with a *normal*$[m', (s')^2]$ posterior. This shows that the *normal*(m, s^2) distribution is the conjugate family for the normal observation distribution with known variance. Bayes' theorem moves from one member of the conjugate family to another member. Because of this we don't need to perform the integration in order to evaluate the posterior. All that is necessary is to determine the rule for updating the parameters.

Simple updating rule for normal family. The updating rules given in Equation 10.4 can be simplified. First we introduce the *precision* of a distribution that is the reciprocal of the variance. Precisions are additive. The posterior precision

$$\frac{1}{(s')^2} = \left(\frac{\sigma^2 s^2}{\sigma^2 + s^2}\right)^{-1} = \frac{\sigma^2 + s^2}{\sigma^2 s^2} = \frac{1}{s^2} + \frac{1}{\sigma^2}\,.$$

Thus the posterior precision equals prior precision plus the observation precision. The posterior mean is given by

$$m' = \frac{(\sigma^2 m + s^2 y)}{\sigma^2 + s^2} = \frac{\sigma^2}{\sigma^2 + s^2} \times a + \frac{s^2}{\sigma^2 + s^2} \times y\,.$$

This can be simplified to

$$m' = \frac{1/s^2}{1/\sigma^2 + 1/s^2} \times a + \frac{1/\sigma^2}{1/\sigma^2 + 1/s^2} \times y\,.$$

Thus the posterior mean is the weighted average of the prior mean and the observation, where the weights are the proportions of the precisions to the posterior precision.

This updating rule also holds for the flat prior. The flat prior has infinite variance, so it has zero precision. The posterior precision will equal the prior precision

$$1/\sigma^2 = 0 + 1/\sigma^2\,,$$

and the posterior variance equals the observation variance σ^2. The flat prior doesn't have a well-defined prior mean. It could be anything. We note that

$$\frac{0}{1/\sigma^2} \times \text{anything} + \frac{1/\sigma^2}{1/\sigma^2} \times y = y\,,$$

so the posterior mean using flat prior equals the observation y

A random sample $y_1, \ldots, y_n$. A random sample $y_1, \ldots, y_n$ is taken from a normal distribution with mean μ and variance σ^2, which is assumed known. We have a prior distribution that is normal with mean m and variance s^2 given by

$$g(\mu) \propto e^{-\frac{1}{2s^2}(\mu - m)^2},$$

where we are ignoring the part that doesn't involve μ because multiplying the prior by any constant will cancel out in the posterior.

We use the likelihood of the sample mean, $\bar{y}$ which is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. The precision of $\bar{y}$ is $(\frac{n}{\sigma^2})$. We see that this is the sum of all the observation precisions for the random sample.

We have reduced the problem to updating given a single normal observation of $\bar{y}$, which we have already solved. Posterior precision equals the prior precision plus the precision of $\bar{y}$.

$$\frac{1}{(s')^2} = \frac{1}{s^2} + \frac{n}{\sigma^2} = \frac{\sigma^2 + ns^2}{\sigma^2 s^2}. \tag{10.5}$$

The posterior variance equals the reciprocal of posterior precision. The posterior mean equals the weighted average of the prior mean and $\bar{y}$ where the weights are the proportions of the posterior precision:

$$m' = \frac{1/s^2}{n/\sigma^2 + 1/s^2} \times m + \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} \times \bar{y}. \tag{10.6}$$

10.3 CHOOSING YOUR NORMAL PRIOR

The prior distribution you choose should match your prior belief. When the observation is from a normal distribution with known variance, the conjugate family of priors for μ is the *normal*(m, s^2). If you can find a member of this family that matches your prior belief, it will make finding the posterior using Bayes' theorem very easy. The posterior will also be a member of the same family where the parameters have been updated by the simple rules given in Equations 10.5 and 10.6. You won't need to do any numerical integration.

First, decide on your prior mean m. This is the value your prior belief is centered on. Then decide on your prior standard deviation s. Think of the points above and below m that you consider not to be reasonably possible. Divide the distance between them by 6 to get your prior standard deviation m. This way you will get reasonable probability over all the region you believe possible.

A useful check on your prior is to consider the "equivalent sample size". Set your prior variance $s^2 = \sigma^2/n_{eq}$ and solve for n_{eq}. This relates your prior precision to the precision from a sample. Your belief is of equal importance to a sample of size n_{eq}. If n_{eq} is large, it shows you have very strong prior belief about μ. It will take a lot of sample data to move your posterior belief far from your prior belief. If it is small, your prior belief is not strong, and your posterior belief will be strongly influenced by a more modest amount of sample data.

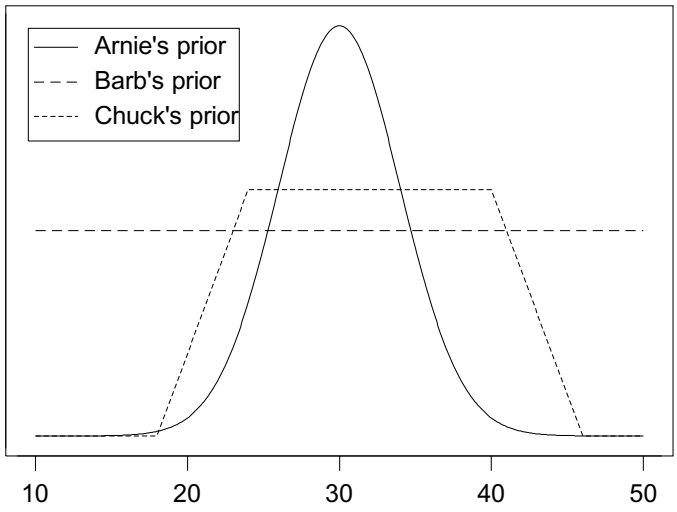


Figure 10.1 Arnie's, Barb's, and Chuck's priors.

If you can't find a prior distribution from the conjugate family that corresponds to your prior belief, then you should determine your prior belief for a selection of points over the range you believe possible, and linearly interpolate between them. Then you can determine your posterior distribution by

$$g(\mu|y_1, \cdots, y_n) = \frac{f(y_1, \cdots, y_n|\mu) \times g(\mu)}{\int f(y_1, \cdots, y_n|\mu) \times g(\mu) d\mu}.$$

Example 18 *Arnie, Barb, and Chuck are going to estimate the mean length of one-year-old rainbow trout in a stream. Previous studies in other streams have shown the length of yearling rainbow trout to be normally distributed with known standard deviation of 2 cm. Arnie decides his prior mean is 30 cm. He decides that he doesn't believe it is possible for a yearling rainbow to be less than 18 cm or greater than 42 cm. Thus his prior standard deviation is 4 cm. Thus he will use a normal*$(30, 4^2)$ *prior. Barb doesn't know anything about trout, so she decides to use the "flat" prior. Chuck decides his prior belief is not normal. His prior has a trapezoidal shape. His prior gives zero weight at 18 cm. It gives weight one at 24 cm, and is level up to 40 cm, and then goes down to zero at 46 cm. He linearly interpolates between those values. The three priors are shown in Figure 10.1.*

They take a random sample of 12 yearling trout from the stream and find the sample mean $\bar{y} = 32$*cm. Arnie and Barb find their posterior distributions using the simple updating rules for the normal conjugate family given by Equations 10.5 and 10.6. For Arnie*

$$\frac{1}{(s')^2} = \frac{1}{4^2} + \frac{12}{2^2}.$$

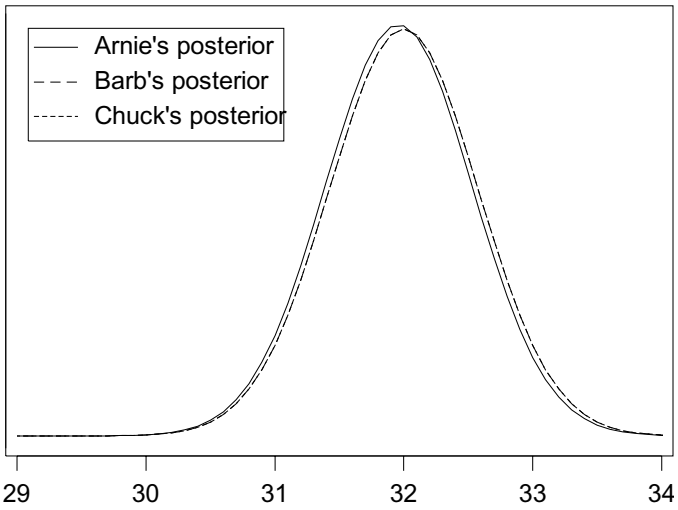


Figure 10.2 Arnie's, Barb's, and Chuck's posteriors. (Barb and Chuck have nearly identical posteriors.)

Solving for this gives his posterior variance $(s')^2 = .3265$. *His posterior standard deviation is* $s' = .5714$. *His posterior mean is found by*

$$m' = \frac{\frac{1}{4^2}}{1.5714^2} \times 30 + \frac{\frac{12}{2^2}}{1.5714^2} = 31.96\,.$$

Barb is using the "flat" prior, so her posterior variance is

$$(s')^2 = \frac{12}{2^2} = .3333$$

and her posterior standard deviation is $s' = .5774$. *Her posterior mean* $m' = 32$, *the sample mean. Both Arnie and Barb have normal posterior distributions.*

Chuck finds his posterior using Equation 10.3 which requires numerical integration. The three posteriors are shown in Figure 10.2. Since Chuck used a prior that was flat over the whole region where the likelihood was appreciable, his posterior is virtually indistinguishable from Barb's who used the flat improper prior. Arnie who used an informative prior has a posterior that is also close to Barb's. This shows that given the data, the posteriors are similar despite starting from quite different priors.

10.4 BAYESIAN CREDIBLE INTERVAL FOR NORMAL MEAN

The posterior distribution $g(\mu|y_1, \cdots, y_n)$ is the inference we make for μ given the observations. It summarizes our entire belief about the parameter given the data. Sometimes we want to summarize our posterior belief into a range of values that we believe cannot be ruled out at some probability level, given the sample data.

An interval like this is called a Bayesian credible interval. It summarizes the range of possible values that are credible at that level. There are many possible credible intervals for a given probability level. Generally, the shortest one is preferred. However, in some cases it is easier to find the credible interval with equal tail probabilities.

Known Variance

When $y_1, \ldots, y_n$ is a random sample from a *normal* (μ, σ^2) distribution, the sampling distribution of $\bar{y}$, the sample mean, is *normal* $(\mu, \sigma^2/n)$. Its mean equals that for a single observation from the distribution, and its variance equals the variance of single observation divided by sample size. Using either a "flat" prior, or a *normal* (m, s^2) prior, the posterior distribution of μ given $\bar{y}$ is *normal* $[m', (s')^2]$, where we update according to the rules:

1. Precision is the reciprocal of the variance.
2. Posterior precision equals prior precision plus the precision of sample mean.
3. Posterior mean is weighted sum of prior mean and sample mean, where the weights are the proportions of the precisions to the posterior precision.

Our $(1 - \alpha) \times 100\%$ Bayesian credible interval for μ is

$$m' \pm z_{\frac{\alpha}{2}} \times s' \,, \tag{10.7}$$

which is the *posterior mean* plus or minus the *z-value* times the *posterior standard deviation*, where the *z-value* is found in the standard normal table. Our posterior *probability* that the true mean μ lies outside the credible interval is α. Since the posterior distribution is *normal* and thus symmetric, the credible interval found using Equation 10.7 is the shortest, as well as having equal tail probabilities.

Unknown Variance

If we don't know the variance, we don't know the precision, so we can't use the updating rules directly. The obvious thing to do is to calculate the sample variance

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

from the data. Then we use Equations 10.5 and 10.6 to find $(s')^2$ and m' where we use the sample variance $\hat{\sigma}^2$ in place of the unknown variance σ^2.

There is extra uncertainty here, the uncertainty in estimating σ^2. We should widen the credible interval to account for this added uncertainty. We do this by taking the values from the *Student's t* table instead of the *standard normal* table. The correct Bayesian credible interval is

$$m' \pm t_{\frac{\alpha}{2}} \times s' \,. \tag{10.8}$$

Table 10.5 95% credible intervals

Person	Posterior distribution	Credible interval lower	upper
Arnie	*Normal(31.96,.3265)*	30.84	33.08
Barb	*Normal(32.00,.3333)*	30.87	33.13
Chuck	numerical	30.8	33.1

The *t* value is taken from the row labelled $df = n - 1$ (degrees of freedom equals number of observations minus 1)[4].

Nonnormal Prior

When we started with a nonnormal prior, we find the posterior distribution for μ using Bayes' theorem where we have to integrate numerically. The posterior distribution will be nonnormal. We can find a $(1 - \alpha) \times 100\%$ credible interval by finding a lower value μ_l and an upper value μ_u such that

$$\int_{\mu_l}^{\mu_u} g(\mu|y_1, \cdots, y_n)\, d\mu = 1 - \alpha\,.$$

There are many such values. The best choice μ_l and μ_u would give us the shortest possible credible interval. These values also satisfy

$$g(\mu_l|y_1, \cdots, y_n) = g(\mu_u|y_1, \cdots, y_n)\,.$$

Sometimes it is easier to find the credible interval with lower and upper tail areas that are equal.

Example 18 (continued) *Arnie, Barb, and Chuck each calculated their 95% credible interval from their respective posterior distributions using Equation 10.7. Chuck had to calculate his numerically from his numerical posterior using the Minitab macro tintegral.mac. The credible intervals are shown in Table 10.5. Arnie, Barb, and Chuck end up with slightly different credible intervals because they started with different prior beliefs. But the effect of the data was much greater than the effect of their priors and their credible intervals are quite similar.*

[4]The resulting Bayesian credible interval is exactly the same one that we would find if we did the full Bayesian analysis with σ^2 as a nuisance parameter, using the joint prior distribution for μ and σ^2 made up of the same prior for $\mu|\sigma^2$ that we used before ["flat" or *normal*(m, s^2)]times the prior for σ^2 given by $g(\sigma^2) \propto (\sigma^2)^{-1}$. We would find the joint posterior by Bayes' theorem. We would find the marginal posterior distribution of μ by marginalizing out σ^2. We would get the same Bayesian credible interval using *Student's t* critical values.

10.5 PREDICTIVE DENSITY FOR NEXT OBSERVATION

Bayesian statistics has a general method for developing the conditional distribution of the next random observation, given the previous random sample. This is called the predictive distribution. This is a clear advantage over frequentist statistics, which can only determine the predictive distribution for some situations. The problem is how to combine the uncertainty from the previous sample with the uncertainty in the observation distribution. The Bayesian approach is called *marginalization*. It entails finding the joint posterior for the next observation and the parameter, given the random sample. The parameter is treated as a *nuisance parameter*, and the marginal distribution of the next observation given the random sample is found by integrating the parameter out of the joint posterior distribution.

Let y_{n+1} be the next random variable drawn after the random sample $y_1, \ldots, y_n$. The predictive density of $y_{n+1}|y_1, \ldots, y_n$ is the conditional density

$$f(y_{n+1}|y_1, \ldots, y_n)\,.$$

This can be found by Bayes' theorem. $y_1, \ldots, y_n, y_{n+1}$ is a random sample from $f(y|\mu)$, which is a normal distribution with mean μ and known variance σ^2 . The conditional distribution of the random sample $y_1, \cdots, y_n$ and the next random observation y_{n+1} given the parameter μ is

$$f(y_1, \ldots, y_n, y_{n+1}|\mu) = f(y_1|\mu) \times \cdots \times f(y_n|\mu) \times f(y_{n+1}|\mu)\,.$$

Let the prior distribution be $g(\mu)$ (either flat prior or *normal*(m, s^2) prior). The joint distribution of the observations and the parameter μ is

$$g(\mu) \times f(y_1|\mu) \times \ldots \times f(y_n|\mu) \times f(y_{n+1}|\mu)\,.$$

The conditional density of y_{n+1} and μ given $y_1, \ldots, y_n$ is

$$f(y_{n+1}, \mu|y_1, \ldots, y_n) = f(y_{n+1}|\mu, y_1, \ldots, y_n) \times g(\mu|y_1, \ldots, y_n)\,.$$

We have already found that the posterior $g(\mu|y_1, \ldots, y_n,)$ is normal with posterior precision equal to prior precision plus the precision of $\bar{y}$ and mean equal to the weighted average of the prior mean and $\bar{y}$ where the weights are proportions of the precisions to the posterior precision. Say it is normal with mean m_n and variance s_n^2. The distribution of y_{n+1} given μ and $y_1, \ldots, y_n$ only depends on μ, because y_{n+1} is another random draw from the distribution $g(y|\mu)$. Thus the joint posterior (to first n observations) distribution is

$$f(y_{n+1}, \mu|y_1, \ldots, y_n) = f(y_{n+1}|\mu) \times g(\mu|y_1, \ldots, y_n)\,.$$

The conditional distribution we want is found by integrating μ out of the joint posterior distribution. This is the marginal posterior distribution

$$\begin{aligned} f(y_{n+1}|y_1, \ldots, y_n) &= \int f(y_{n+1}, \mu|y_1, \ldots, y_n)\, d\mu \\ &= \int f(y_{n+1}|\mu) \times g(\mu|y_1, \ldots, y_n)\, d\mu\,. \end{aligned}$$

These are both normal under our assumed model, so

$$f(y_{n+1}|y_1,\ldots,y_n) \propto \int e^{-\frac{1}{2\sigma^2}(y_{n+1}-\mu)^2} e^{-\frac{1}{2s_n^2}(\mu-m_n)^2}\, d\mu\,.$$

Adding the exponents and combining like terms.

$$\begin{aligned} f(y_{n+1}|y_1,\ldots,y_n) &\propto \int e^{-\frac{1}{2}\left[\frac{(\mu^2-2\mu y_{n+1}+y_{n+1}^2)}{\sigma^2}+\frac{(\mu^2-2\mu m_n+m_n^2)}{s_n^2}\right]} d\mu \\ &\propto \int e^{-\frac{1}{2}\left[(\frac{1}{\sigma^2}+\frac{1}{s_n^2})\mu^2-2(\frac{y_{n+1}}{\sigma^2}+\frac{m_n}{s_n^2})\mu+\frac{y_{n+1}^2}{\sigma^2}+\frac{m_n^2}{s_n^2}\right]} d\mu\,. \end{aligned}$$

Factoring out $(\frac{1}{\sigma^2}+\frac{1}{s_n^2})$ of the exponent and completing the square

$$\propto \int e^{-\frac{1}{2(\sigma^2 s_n^2)/(\sigma^2+s_n^2)}\left[\mu-\frac{(s_n^2 y_{n+1}+\sigma^2 m_n)}{\sigma^2+s_n^2}\right]^2}$$

$$\times e^{-\frac{1}{2(\sigma^2 s_n^2)/(\sigma^2+s_n^2)}\left[-\left(\frac{s_n^2 y_{n+1}+\sigma^2 m_n}{\sigma^2+s_n^2}\right)^2+\frac{s_n^2 y_{n+1}^2+\sigma^2 m_n^2}{s_n^2+\sigma^2}\right]} d\mu\,.$$

The first line is the only part that depends on μ, and we recognize that it is proportional to a normal density, so integrating it over its whole range gives a constant. Reorganizing the second part gives

$$\propto e^{-\frac{1}{2(\sigma^2 s_n^2)/(\sigma^2+s_n^2)}\left[\frac{(s_n^2 y_{n+1}^2+\sigma^2 m_n^2)(\sigma^2+s_n^2)-(s_n^4 y_{n+1}^2+2s_n^2\sigma^2 y_{n+1}m_n+\sigma^4 m_n^2)}{(\sigma^2+s_n^2)^2}\right]},$$

which simplifies to

$$\propto e^{-\frac{1}{2(\sigma^2+s_n^2)}(y_{n+1}-m_n)^2}. \tag{10.9}$$

We recognize this as a normal density with mean m_n and variance $\sigma^2+s_n^2$. The predictive mean for the observation y_{n+1} is the posterior mean of μ given the observations $y_1,\ldots,y_n$. The predictive variance is the observation variance σ^2 plus the posterior variance of μ given the observations $y_1,\ldots,y_n$. (Part of the uncertainty in the prediction is due to the uncertainty in estimating the posterior mean.)

This is one of the advantages of the Bayesian approach. It has a single clear approach (marginalization) that is always used to construct the predictive distribution. There is no single clear cut way this can be done in frequentist statistics, although in many problems such as the normal case we just did, they can come up with similar results.

Main Points

- Analyzing the observations sequentially one at a time, using the posterior from the previous observation as the next prior gives the same results as analyzing all the observations at once using the initial prior.

- The likelihood of a random sample of normal observations is proportional to the likelihood of the sample mean.

- The conjugate family of priors for *normal* observations with known variance is the *normal(*m, s^2*)* family.

- If we have a random sample of normal observations and use a *normal(*m, s^2*)* prior the posterior is *normal[*$m', (s')^2$*]*, where m' and $(s')^2$ are found by the simple updating rules:

    - The precision is the reciprocal of the variance.
    - Posterior precision is the sum of the prior precision and the precision of the sample.
    - The posterior mean is the weighted average of the prior mean and the sample mean, where the weights are the proportions of their precisions to the posterior precision.

- The same updating rules work for the flat prior, remembering the flat prior has precision equal to zero.

- A Bayesian credible interval for μ can be found using the posterior distribution.

- If the variance σ^2 is not known, we use the estimate of the variance calculated from the sample, $\hat{\sigma}^2$, and use the critical values from the *Student's t* table where the degrees of freedom is $n - 1$, the sample size minus 1. Using the *Student's t* critical values compensates for the extra uncertainty due to not knowing σ^2. (This actually gives the correct credible interval if we used a prior $g(\sigma^2) \propto \frac{1}{\sigma^2}$, and marginalized σ^2 out of the joint posterior.)

- The predictive distribution of the next observation is *normal*$(m', s'2 + \sigma^2)$. Its mean is the same as the posterior mean, and its variance is the posterior variance plus the observation variance. (The posterior variance s'^2 allows for the uncertainty in estimating μ.) The predictive distribution is found by marginalizing μ out of the joint distribution $f(y_{n+1}, \mu | y_1, \ldots, y_n)$.

Exercises

10.1 You are the statistician responsible for quality standards at a cheese factory. You want the probability that a randomly chosen block of cheese labelled "1 kg" is actually less than 1 kilogram (1000 grams) to be 1% or less. The weight (in grams) of blocks of cheese produced by the machine is *normal*

(μ, σ^2) where $\sigma^2 = 3^2$. The weights (in grams) of 20 blocks of cheese are:

994	997	999	1003	994
998	1001	998	996	1002
1004	995	994	995	998
1001	995	1006	997	998

You decide to use a discrete prior distribution for μ with the following probabilities:

Value	Prior Probability
991	.05
992	.05
993	.05
994	.05
995	.05
996	.05
997	.05
998	.05
999	.05
1000	.05
1001	.05
1002	.05
1003	.05
1004	.05
1005	.05
1006	.05
1007	.05
1008	.05
1009	.05
1010	.05

(a) Calculate your posterior probability distribution.

(b) Calculate your posterior probability that $\mu < 1000$.

(c) Should you adjust the machine?

10.2 The city health inspector wishes to determine the mean bacteria count per liter of water at a popular city beach. Assume the number of bacteria per liter of water is *normal* with mean μ and standard deviation known to be $\sigma = 15$. She collects 10 water samples and found the bacteria counts to be:

175	190	215	198	184
207	210	193	196	180

She decides that she will use a discrete prior distribution for μ with the following probabilities:

Value	Prior Probability
160	.125
170	.125
180	.125
190	.125
200	.125
210	.125
220	.125
230	.125

(a) Calculate her posterior distribution.

10.3 The standard process for making a polymer has mean yield 35%. A chemical engineer has developed a modified process. He runs the process on 10 batches and measures the yield (in percent) for each batch. They are:

38.7	40.4	37.2	36.6	35.9
34.7	37.6	35.1	37.5	35.6

Assume that yield is *normal* (μ, σ^2) where the standard deviation $\sigma = 3$ is known.

(a) Use a *normal* $(30, 10^2)$ prior for μ. Find the posterior distribution.

(b) The engineer wants to know if the modified process increases the mean yield. Set this up as a hypothesis test stating clearly the null and alternative hypotheses.

(c) Perform the test at the 5% level of significance.

10.4 An engineer takes a sample of 5 steel I beams from a batch, and measures the amount they sag under a standard load. The amounts in mm are:

5.19	4.72	4.81	4.87	4.88

It is known that the sag is *normal* (μ, σ^2) where the standard deviation $\sigma = .25$ is known.

(a) Use a *normal* $(5, .5^2)$ prior for μ. Find the posterior distribution.

(b) For a batch of I beams to be acceptable, the mean sag under the standard load must be less than 5.20. ($\mu < 5.20$). Set this up as a hypothesis test stating clearly the null and alternative hypotheses.

(c) Perform the test at the 5% level of significance.

10.5 New Zealand was the last major land mass to be settled by human beings. The Shag River Mouth in Otago (lower South Island), New Zealand, is one of the sites of early human inhabitation that New Zealand archeologists have investigated, in trying to determine when the Polynesian migration to New Zealand occurred and documenting local adaptations to New Zealand conditions. Petchey and Higham (2000) describe the Radiocarbon dating of well-preserved barracouta *thyrsites atun* bones found at the Shag River Mouth site. They obtained four acceptable samples, which were analyzed by the Waikato University Carbon Dating Unit. Assume that the conventional radiocarbon age (CRA) of a sample follows the *normal* (μ, σ^2) distribution, where the standard deviation $\sigma = 40$ is known. The observations are:

Observation	1	2	3	4
CRA	940	1040	910	990

(a) Use a *normal* $(1000, 200^2)$ prior for μ. Find the posterior distribution $g(\mu|y_1, \ldots, y_4)$.

(b) Find a 95% credible interval for μ.

(c) To find the θ, the calibrated date, the Stuiver, Reimer, and Braziunas marine curve was used. We will approximate this curve with the linear function

$$\theta = 2203 - .835 \times \mu\,.$$

Find the posterior distribution of θ given $y_1, \ldots, y_4$.

(d) Find a 95% credible interval for θ, the calibrated date.

10.6 The Houhora site in Northland (top of North Island) New Zealand is one of the sites of early human inhabitation that New Zealand archeologists have investigated, in trying to determine when the Polynesian migration to New Zealand occurred and documenting local adaptations to New Zealand conditions. Petchey (2000) describe the Radiocarbon dating of well-preserved snapper *Pagrus auratus* bones found at the Houhora site. They obtained four acceptable samples which were analyzed by the Waikato University Carbon Dating Unit. Assume that the conventional radiocarbon age (CRA) of a sample follows the *normal* (μ, σ^2) distribution where the standard deviation $\sigma = 40$ is known. The observations are:

Observation	1	2	3	4
CRA	1010	1000	950	1050

(a) Use a *normal* $(1000, 200^2)$ prior for μ. Find the posterior distribution $g(\mu|y_1, \ldots, y_4)$.

(b) Find a 95% credible interval for μ.

(c) To find the θ, the calibrated date, the Stuiver, Reimer, Braziunas marine curve was used. We will approximate this curve with the linear function

$$\theta = 2203 - .835 \times \mu\,.$$

Find the posterior distribution of θ given $y_1, \ldots, y_4$.

(d) Find a 95% credible interval for θ, the calibrated date.

Computer Exercises

10.1 Use the Minitab macro *NormDP.mac* to find the posterior distribution of the mean μ when we have a random sample of observations from a *normal* (μ, σ^2), where σ^2 is known, and we have a *discrete* prior for μ.

Suppose we have a random sample of $n = 10$ observations from a *normal* (μ, σ^2) distribution where it is known $\sigma^2 = 4$. The random sample of observations are:

3.07	7.51	5.95	6.83	8.80	4.19	7.44	7.06	9.67	6.89

We only allow that there are 12 possible values for μ, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, and 9.5. If we don't favor any possible value over another, so we give all possible values of μ probability equal to $\frac{1}{12}$. The prior distribution is:

μ	$g(\mu)$
4.0	.083333
4.5	.083333
5.0	.083333
5.5	.083333
6.0	.083333
6.5	.083333
7.0	.083333
7.5	.083333
8.0	.083333
8.5	.083333
9.0	.083333
9.5	.083333

Use *NormDP.mac* to find the posterior distribution $g(\mu|y_1, \ldots, y_{10})$. Details for invoking *NormDP.mac* are in Appendix 3.

10.2 Suppose another 6 random observations come later. They are:

6.22	3.99	3.67	6.35	7.89	6.13

Use *NormDP.mac* to find the posterior distribution, where we will use the posterior after the first ten observations $y_1, \ldots, y_{10}$, as the prior for the next six observations $y_{11}, \ldots, y_{16}$.

10.3 Instead, combine all the observations together to give a random sample of size $n = 16$, and use *NormDP.mac* to find the posterior distribution where we go back the original prior that had all the possible values equally likely. What do the results of the last two problems show us?

10.4 Instead of thinking of a random sample of size $n = 16$, let's think of the sample mean as a single observation from its distribution.

(a) What is the distribution of $\bar{y}$? Calculate the observed value of $\bar{y}$?

(b) Use *NormDP.mac* to find the posterior distribution $g(\mu|\bar{y})$.

(c) What does this show us?

10.5 We will use the Minitab macro *NormNP.mac* to find the posterior distribution of the normal mean μ when we have a random sample of size n from a *normal* (μ, σ^2) distribution with known σ^2, and we use a *normal* (m, s^2) prior for μ. The *normal* family of priors is the conjugate family for *normal* observations. That means that if we start with one member of the family as the prior distribution, we will get another member of the family as the posterior distribution. It is especially easy; if we start with a *normal* (m, s^2) prior, we get a *normal* $(m', (s')^2)$ posterior where $(s')^2$ and m' are given by

$$\frac{1}{(s')^2} = \frac{1}{s^2} + \frac{n}{\sigma^2}$$

and

$$m' = \frac{1/s^2}{1/(s')^2} \times m + \frac{n/\sigma^2}{1/(s')^2} \times \bar{y}$$

respectively. Suppose the $n = 15$ observations from a *normal* $(\mu, \sigma^2 = 4^2)$ are:

26.8	26.3	28.03	28.5	26.3
31.9	28.5	27.2	20.9	27.5
28.0	18.6	22.3	25.0	31.5

Use *NormNP.mac* to find the posterior distribution $g(\mu|y_1, \ldots, y_{15}$, where we choose a *normal* $(m = 20, s^2 = 5^2)$ prior for μ. The details for invoking

NormNP.mac are in Appendix 3. Store the likelihood and posterior in c3 and c4, respectively.

(a) What are the posterior mean and standard deviation?

(b) Find a 95% credible interval for μ.

10.6 Repeat part (a) with a *normal* $(30, 4^2)$ prior, storing the likelihood and posterior in c5 and c6.

10.7 Graph both posteriors on the same graph. What do you notice? What do you notice about the two posterior means and standard deviations? What do you notice about the two credible intervals for π?

10.8 We will use the Minitab macro *NormGCP.mac* to find the posterior distribution of the normal mean μ when we have a random samples of size n of *normal* (μ, σ^2) observations with known $\sigma^2 = 2^2$, and we have a general continuous prior for μ. Suppose the prior has the shape given by

$$g(\mu) = \begin{cases} \mu & \text{for} \quad 0 < \mu \leq 3 \\ 3 & \text{for} \quad 3 < \mu < 5 \\ 8 - \mu & \text{for} \quad 5 < \mu \leq 8 \\ 0 & \text{for} \quad 8 < \mu \end{cases}$$

Store the values of μ and prior $g(\mu)$ in column c1 and c2, respectively. Suppose the random sample of size $n = 16$ is:

4.09	4.68	1.87	2.62	5.58	8.68	4.07	4.78
4.79	4.49 5.85	5.90	2.40	6.27	6.30	4.47	

(a) Use *NormGCP.mac* to determine the posterior distribution $g(\mu|y_1, \ldots, y_{16})$. Details for invoking *NormGCP.mac* are in Appendix 3.

(b) Use *tintegral.mac* to find the posterior mean and posterior standard deviation of μ. Details for invoking *tintegral.mac* are in Appendix 3.

(c) Find a 95% credible interval for μ by using *tintegral.mac*.

11

Comparing Bayesian and Frequentist Inferences for Mean

Making inferences about the population mean when we have a random sample from a normally distributed population is one of the most widely encountered situations in statistics. From the Bayesian point of view, the posterior distribution sums up our entire belief about the parameter given the sample data. It really is the complete inference. However, from the frequentist perspective, there are several distinct types of inference that can be done: point estimation, interval estimation, and hypothesis testing. Each of these types of inference can be performed in a Bayesian manner, where they would be considered summaries of the complete inference, the posterior. In Chapter 9 we compared the Bayesian and frequentist inferences about the population proportion π. In this chapter we look at the frequentist methods for point estimation, interval estimation, and hypothesis testing about μ, the mean of a normal distribution, and compare them with their Bayesian counterparts using frequentist criteria.

11.1 COMPARING FREQUENTIST AND BAYESIAN POINT ESTIMATORS

A frequentist point estimator for a parameter is a statistic that we use to estimate the parameter. The simple rule we use to determine a frequentist estimator for μ is to use

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

the statistic that is the sample analog of the parameter to be estimated. So we use the sample mean $\bar{y}$ to estimate the population mean μ.[1]

In Chapter 9 we learned that frequentist estimators for unknown parameters are evaluated by considering their sampling distribution. In other words, we look at the distribution of the estimator over all possible samples. A commonly used criterion is that the estimator be *unbiased*. That is, the mean of its sampling distribution is the true unknown parameter value. The second criterion is that the estimator have small variance in the class of all possible unbiased estimators. The estimator that has the smallest variance in the class of unbiased estimators is called the *minimum variance unbiased estimator* and is generally preferred over other estimators from the frequentist point of view.

When we have a random sample from a normal distribution, we know that the sampling distribution of $\bar{y}$ is normal with mean μ and variance $\frac{\sigma^2}{n}$. The sample mean, $\bar{y}$, turns out to be the *minimum variance unbiased estimator* of μ.

We take the mean of the posterior distribution to be the Bayesian estimator for μ:

$$\hat{\mu}_B = E(\mu|y_1, \cdots, y_n) = \frac{1/s^2}{n/\sigma^2 + 1/s^2} \times m + \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} \times \bar{y}\,.$$

We know that the posterior mean minimizes the posterior mean square. This means that $\hat{\mu}_B$ is the optimum estimator in the *post-data* setting. In other words, it is the optimum estimator for μ given our sample data and using our prior.

We will compare its performance to that of $\hat{\mu}_f = \bar{y}$ under the frequentist assumption that the true mean μ is a fixed but unknown constant. The probabilities will be calculated from the sampling distribution of $\bar{y}$. In other words, we are comparing the two estimators for μ in the *pre-data* setting.

The posterior mean is a linear function of the random variable $\bar{y}$, so its expected value is

$$E(\hat{\mu}_B) = \frac{1/s^2}{n/\sigma^2 + 1/s^2} \times m + \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} \times \mu\,.$$

The bias of the posterior mean is its expected value minus the true parameter value, which simplifies to

$$\frac{\sigma^2}{ns^2 + \sigma^2}(m - \mu)\,.$$

The posterior mean is a biased estimator of μ. The bias could only be 0 if our prior mean coincides with the unknown true value. The probability of that happening is 0. The bias increases linearly with the distance the prior mean m is from the true unknown mean μ. The variance of the posterior mean is

$$\left[\frac{n/\sigma^2}{n/\sigma^2 + 1/s^2}\right]^2 \times \frac{\sigma^2}{n} = \frac{s^2}{ns^2 + \sigma^2}\sigma^2$$

[1] The maximum likelihood estimator is the value of the parameter that maximizes the likelihood function. It turns out that $\bar{y}$ is the maximum likelihood estimator of μ for a normal random sample.

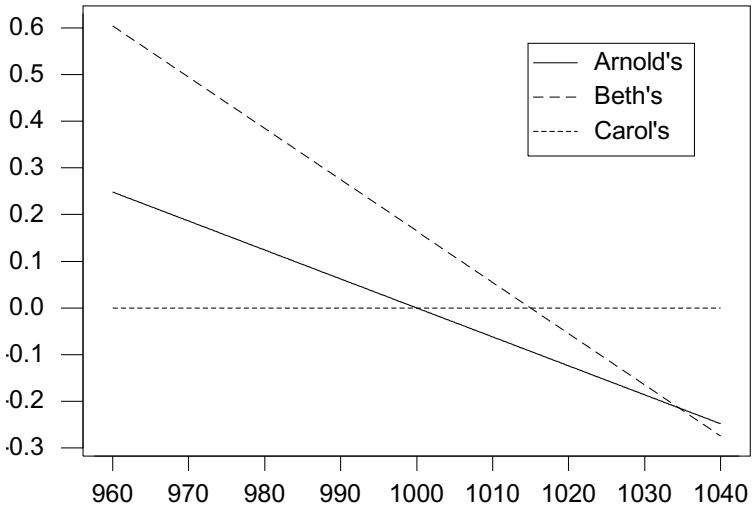


Figure 11.1 Biases of Arnold's, Beth's, and Carol's estimators.

and is seen to be clearly smaller than $\frac{\sigma^2}{n}$, which is the variance of the frequentist estimator $\hat{\mu}_f = \bar{y}$. The mean squared error of an estimator combines both the bias and the variance into a single measure:

$$MS(\hat{\mu}_B) = bias^2 + Var(\hat{\mu}) .$$

The frequentist estimator $\hat{\mu}_f = \bar{y}$ is an unbiased estimator of μ, so its mean squared error equals its variance:

$$MS(\hat{\mu}_f) = \frac{\sigma^2}{n} .$$

When there is prior information, we will see that the Bayesian estimator has smaller mean squared error over the range of μ values that are realistic.

Example 19 *Arnold, Beth, and Carol want to estimate the mean weight of "1 kg" packages of milk powder produced at a dairy company. The weight in individual packages is subject to random variation. They know that when the machine is adjusted properly, the weights are normally distributed with mean 1015 grams, and standard deviation 5 gm. They are going to base their estimate on a sample of size 10. Arnold decides to use a normal prior with mean 1000 gm and standard deviation 20 gm. Beth decides she will use a normal prior with mean 1015 and standard deviation 15. Carol decides she will use a "flat" prior. They calculate the bias, variance, and mean squared error of their estimators for various values of μ to see how well they perform.*

Figure 11.1 shows that only Carol's prior will give an unbiased Bayesian estimator. Her posterior Bayesian estimator corresponds exactly to the frequentist estimator

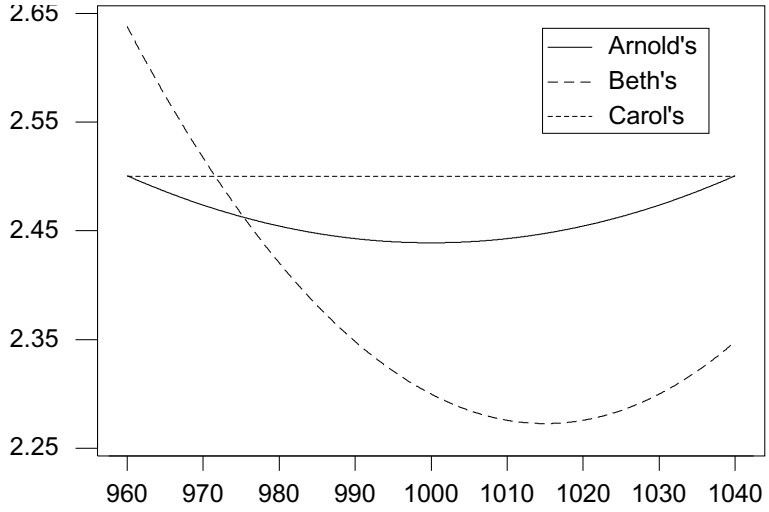


Figure 11.2 Mean-squared errors of Arnold's, Beth's, and Carol's estimators.

$\hat{\mu}_f = \bar{y}$, since she used the "flat" prior. In Figure 11.2 we see the ranges over which the Bayesian estimators have smaller MS than the frequentist estimator. In that range they will be closer to the true value, on average, than the frequentist estimator. The realistic range is the target mean (1015) plus or minus 3 standard deviations (5) which is from 1000 to 1030.

Although both Arnold and Beth's estimators are biased since they are using the Bayesian approach, they have smaller mean squared error over the feasible range than Carol's estimator (which equals the ordinary frequentist estimator). Since they have smaller mean squared error, on average, they will be closer to the true value in the feasible range. In particular, Beth's estimator seems to offer substantially better performance over this range.

11.2 COMPARING CONFIDENCE AND CREDIBLE INTERVALS FOR MEAN

Frequentist statisticians compute confidence intervals for the parameter μ to determine an interval that "has a high probability of containing the true value." Since they are done from the frequentist perspective, the parameter μ is considered a fixed but unknown constant. The coverage probability is found from the sampling distribution of an estimator, in this case $\bar{y}$, the sample mean. The sampling distribution of $\bar{y}$ is *normal* with mean μ and variance σ^2. We know before we take the sample that $\bar{y}$ is a random variable, so we can make the probability statement about $\bar{y}$:

$$P\left(\mu - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \bar{y} < \mu + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha\,,$$

where $z_{\frac{\alpha}{2}}$ is the value from the standard normal table having tail area $\frac{\alpha}{2}$. We rearrange this probability statement to have μ in the middle. The upper inequality in the first statement becomes the lower inequality in the second statement, and vice versa:

$$P\left(\bar{y} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha\,.$$

The endpoints of the interval are random because they depend on $\bar{y}$, which is the random variable in this interpretation. The parameter μ is considered a fixed but unknown constant. So the correct interpretation is that $(1-\alpha) \times 100\%$ of the intervals calculated this way will contain the true value. When we take our random sample and calculate $\bar{y}$, there is nothing random left to attach a probability to. The actual interval we calculate either contains the true value or it does not. Only we don't know which is true. So we say that we are $(1-\alpha) \times 100\%$ *confident* that the interval we calculated using the observed value of $\bar{y}$,

$$\bar{y} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\,, \tag{11.1}$$

does contain the true value. Our confidence comes from the sampling distribution of the statistic. It does not come from the actual sample values we used to calculate the endpoints of the confidence interval. Sometimes we write the confidence interval as

$$\left(\bar{y} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\,,\;\; \bar{y} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right)\,.$$

This contrasts with the Bayesian credible interval for μ that we calculated in the previous chapter. The probability statement we make is from the posterior distribution of the parameter μ given the sample data $y_1, \cdots, y_n$. It is conditional on the actual sample data we obtained. The probability given in the statement is our probability given the actual sample. It is a legitimate probability statement, since μ is considered random. But it is *subjective* because we constructed it using our *subjective* prior. Someone else who started with a different prior would end up with a (slightly) different credible interval.

Relationship between Frequentist Confidence Interval and Bayesian Credible Interval from "Flat" Prior

With a flat prior for μ, the posterior mean equals $m' = \bar{y}$, and the posterior variance equals $(s')^2 = \sigma^2/n$. So for this case the Bayesian credible interval and the frequentist confidence interval will have the form

$$\left(\bar{y} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right)\,.$$

However, they have different interpretations.

The frequentist interpretation is that μ is fixed. The endpoints of the random interval are calculated using a probability statement on the sampling distribution of

the statistic $\bar{y}$. There is no randomness left after the actual sample data have been used to calculate the endpoints. No probability statements can be made about the actual calculated interval. The confidence level $(1-\alpha)\times 100\%$ associated with the interval means that $(1-\alpha)\times 100\%$ of the random intervals calculated this way will contain the true unknown parameter, so we are that $(1-\alpha)\times 100\%$ *confident* that the one we calculate does.

The Bayesian interpretation lets μ be a random variable, so probability statements are allowed. The credible interval is calculated from the posterior distribution given the actual sample data that occurred. The credible interval has the stated conditional probability of containing μ, given the data.

Scientists are not interested in what would happen with hypothetical repetitions of the experiment giving all possible data sets. The only data set that matters is the one that occurred. They find direct probability statements about the parameter, conditional on their actual data set to be the most useful. Scientists often take the confidence interval given by the frequentist statistician and misinterpret it as a probability interval for the parameter given the data. The statistician knows this is interpretation is not the correct one but lets the scientist make the misinterpretation. The correct interpretation is scientifically useless.

Fortunately for frequentist statisticians, when they allow their clients to make the probability interpretation from the confidence interval for the mean of a normal distribution, μ, they can get away with it. Their interval is equivalent to the Bayesian credible interval from a "flat" prior, which allows the probability interpretation in this case

Example 18 (continued) *Previous studies have determined that the length of yearling trout have a normal $(\mu, \sigma^2 = 2^2)$ distribution. Arnie, Barb, and Chuck obtained a random sample of 12 yearling trout. The sample mean $\bar{y} = 32$ cm. The 95% confidence interval for μ is given by*

$$\bar{y} \pm z_{.025} \times \frac{\sigma}{\sqrt{n}} = 32 \pm 1.96 \times \frac{2}{\sqrt{12}} = (30.87, 33.13)\,.$$

Compare this with the 95% credible intervals they found in Table 10.5. We see that it is the same as the credible interval Barb found because she used the "flat" prior.

11.3 TESTING A ONE-SIDED HYPOTHESIS ABOUT A NORMAL MEAN

Often we get data from a new population similar to a population we already know about. For instance, the new population may be the set of all possible outcomes of an experiment, where we have changed one of the experimental factors from its standard value to a new value. We know the mean value of the standard population is μ_0. We assume each observation from the new population is *normal* (μ, σ^2) where σ^2 is known, and that the observations are independent of each other. The question we want to answer is, Is the mean μ for the new population greater than the mean of the standard population? A one-sided hypothesis test attempts to answer that question.

We consider there are two possible explanations to any discrepancy between the observed data and μ_0.

1. The mean of the new population is less than or equal to the mean of the standard population, and any discrepancy is due to chance alone.

2. The mean of the new population is greater than the mean of the standard population and at least part of the discrepancy is due to this fact.

Hypothesis testing is a way to protect our credibility by making sure that we don't reject the first explanation unless it has probability less than our chosen level of significance α. Note that we set up the positive answer to the question we are asking as the alternative hypothesis. The null hypothesis will be the negative answer to the question. We will compare the frequentist and Bayesian approaches.

Frequentist One-Sided Hypothesis Test about μ

As we saw in Chapter 9, frequentist tests are based on the sampling distribution of a statistic. This makes the probabilities *pre-data* in that they arise from all possible random samples that could have occurred. The steps are:

1. Set up the null and alternative hypothesis

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_0 : \mu > \mu_0 \, .$$

   Note the alternative hypothesis is the change in the direction we are interested in detecting. Any change in the other direction gets lumped into the null hypothesis. (We are trying to detect $\mu > \mu_0$. If $\mu < \mu_0$, it is not of any interest to us, so those values get included in the null hypothesis.)

2. The null distribution of $\bar{y}$ is *normal* $(\mu_0, \frac{\sigma^2}{n})$. This is the sampling distribution of $\bar{y}$ when the null hypothesis is true. Hence the null distribution of the standardized variable

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

   will be *normal* $(0, 1)$.

3. Choose a level of significance α. Commonly this is .10, .05, or .01.

4. Determine the rejection region. This is a region that has probability α when the null hypothesis is true ($\mu = \mu_0$). When $\alpha = .05$, the rejection region is $z > 1.645$. This is shown in Figure 11.3.

5. Take the sample data and calculate $\bar{y}$. If the value falls in the rejection region, we reject the hypothesis at level of significance $\alpha = .05$, else we can't reject the null hypothesis.

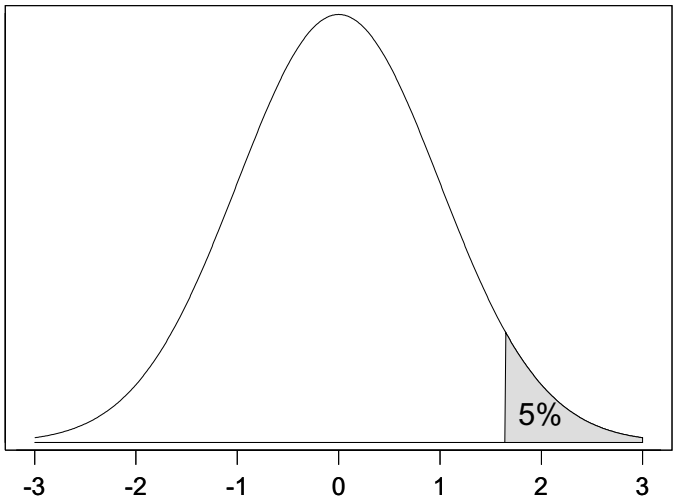


Figure 11.3 Null distribution of $z = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$ with rejection region for one-sided frequentist hypothesis test at 5% level of significance.

6. Another way to perform the test is to calculate the *p-value* which is the probability of observing what we observed, or something even more extreme, given the null hypothesis $H_0 : \mu = \mu_0$ is true:

$$\textit{p-value} = P\left(Z \geq \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}\right) . \tag{11.2}$$

If *p-value*$\leq \alpha$, then we reject the null hypothesis, else we can't reject it.

Bayesian One-Sided Hypothesis Test about μ

The posterior distribution $g(\mu|y_1, \cdots, y_n)$ summarizes our entire belief about the parameter, after viewing the data. Sometimes we want to answer a specific question about the parameter. This could be, Given the data, can we conclude the parameter μ is greater than μ_0? The value μ_0 ordinarily comes from previous experience. If the parameter is still equal to that value, then the experiment has not demonstrated anything new that requires explaining. We would lose our scientific credibility if we go around concocting explanations for effects that may not exist. The answer to the question can be resolved by testing

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0 .$$

This is an example of a one-sided hypothesis test. We decide on a level of significance α that we wish to use. It is the probability below which we will reject the null

hypothesis. Usually α is small, for instance, .10, .05, .01, .005, or .001. Testing a one-sided hypothesis in Bayesian statistics is done by calculating the posterior probability of the null hypothesis:

$$P(H_0 : \mu \leq \mu_0 | y_1, \ldots, y_n) = \int_{-\infty}^{\mu_0} g(\mu | y_1, \ldots, y_n)\, d\mu\,. \tag{11.3}$$

When the posterior distribution $g(\mu|y_1, \ldots, y_n)$ is *normal*(m', s'^2) this can easily be found from standard normal tables.

$$\begin{aligned} P(H_0 : \mu \leq \mu_0 | y_1, \ldots, y_n) &= P\left(\frac{\mu - m'}{s'} \leq \frac{\mu_0 - m'}{s'}\right) \\ &= P\left(Z \leq \frac{\mu_0 - m'}{s'}\right), \end{aligned} \tag{11.4}$$

where Z is a standard normal random variable. If the probability is less than our chosen α, we reject the null hypothesis and can conclude that $\mu > \mu_0$. Only then can we search for an explanation of why μ is now larger than μ_0.

Example 18 (continued from Chapter 10.) *Arne, Barb, and Chuck read in a journal that the mean length of yearling rainbow trout in a typical stream habitat is 31 cm. The each decide to determine if the mean length of trout in the stream they are researching is greater than that by testing*

$$H_0 : \mu \leq 31 \quad \text{versus} \quad H_1 : \mu > 31$$

at the $\alpha = 5\%$ level. For one-sided Bayesian hypothesis tests, they calculate the posterior probability of the null hypothesis. Arnie and Barb have normal posteriors, so they use Equation 11.4. Chuck has a nonnormal posterior that he calculated numerically. He calculates the posterior probability of the null hypothesis using Equation 11.3, and evaluates it numerically using the Minitab macro tintegral.mac. The results of the Bayesian hypothesis tests are shown in Table 11.1.

They also decide that they will perform the corresponding frequentist hypothesis test of

$$H_0 : \mu \leq 31 \quad \text{versus} \quad H_1 : \mu > 31$$

and compare the results. The null distribution of $z = \frac{\bar{y}-31}{\sigma/\sqrt{n}}$ and the correct rejection region are given in Figure 11.3. For this data, $z = \frac{32-31}{2/\sqrt{12}} = 1.732$. This lies in the rejection region, hence the null hypothesis is rejected at the 5% level. The other way we could perform this frequentist hypothesis test is to calculate the p-value using Equation 11.3. For these data,

$$\begin{aligned} p\text{-}value &= P\left(Z > \frac{32 - 31}{2/\sqrt{12}}\right) \\ &= P(Z > 1.732) \end{aligned}$$

Table 11.1 Results of Bayesian one-sided hypothesis tests

Person	Posterior	$P(\mu \leq 31 \mid y_1, \ldots, y_n)$		
Arnie	*normal*$(31.96, .5714^2)$	$P(Z \leq \frac{31-31.96}{.5714})$	=.0465	reject
Barb	*normal*$(32.00, .5774^2)$	$P(Z \leq \frac{31-32}{.5774})$	=.0416	reject
Chuck	*numerical*	$\int_{\infty}^{31} g(\mu \mid y_1, \ldots, y_n) d\mu$	=.0489	reject

which equals .0416 from the standard normal table in Appendix B (Table B.2). This is less than the level of significance α, so the null hypothesis is rejected, same as before[2].

11.4 TESTING A TWO-SIDED HYPOTHESIS ABOUT A NORMAL MEAN

Sometimes the question we want to have answered is, Is the mean for the new population μ, the same as the mean for the standard population which we know equals μ_0? A two-sided hypothesis test attempts to answer this question. We are interested in detecting a change in the mean, in either direction. We set this up as

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0 \,. \tag{11.5}$$

The null hypothesis is known as a *point hypothesis*. This means that, it is true only for the exact value μ_0. This is only a single point along the number line. At all the other values in the parameter space the null hypothesis is false. When we think of the infinite number of possible parameter values in an interval of the real line, we see that the it is impossible for the null hypothesis to be literally true. There are an infinite number of values that are extremely close to μ_0 but eventually differ from μ_0 when we look at enough decimal places. So rather than testing whether we believe the null hypothesis to actually be true, we are testing whether the null hypothesis is in the range that could be true.

Frequentist Two-Sided Hypothesis Test About μ

1. The null and alternative hypothesis are set up as in Equation 11.5. Note that we are trying to detect a change in either direction.

2. The null distribution of the standardized variable

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

will be *normal* $(0, 1)$.

[2]We note that in this case, the *p-value* equals Barb's probability of the null hypothesis because she used the "flat" prior. For the *normal* case, the *p-value* can be interpreted as the posterior probability of the null hypothesis when the noninformative "flat" prior was used. However, it is not generally true that *p-value* has any meaning in the Bayesian perspective.

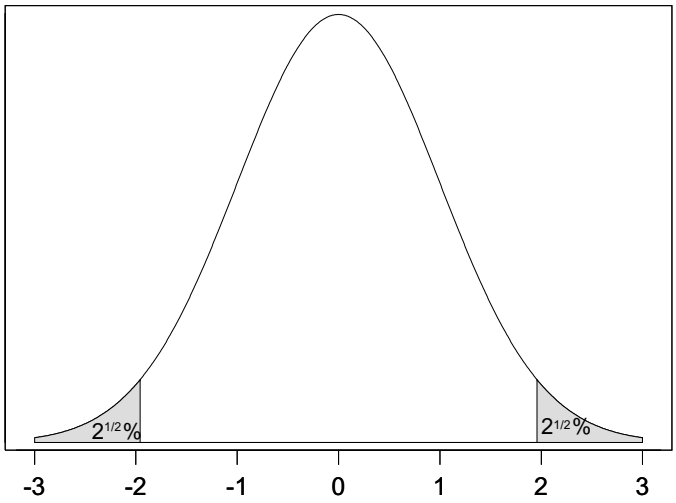


Figure 11.4 Null distribution of $z = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$ with rejection region for two-sided frequentist hypothesis test at 5% level of significance.

3. Choose α, the level of significance. This is usually a low value such as .10, .05, .01, or .001.

4. Determine the rejection region. This is a region that has probability = α when the null hypothesis is true. For a two-sided hypothesis test, we have a two-sided rejection region. When $\alpha = .05$, the rejection region is $|z| > 1.96$. This is shown in Figure 11.4.

5. Take the sample and calculate $z = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$. If it falls in the rejection region, reject the null hypothesis at level of significance α, else we can't reject the null hypothesis.

6. Another way to do the test is to calculate the *p-value* which is the probability of observing what we observed, or something even more extreme than what we observed, given the null hypothesis is true. Note that the *p-value* includes probability of two tails:

$$p\text{-}value = P\left(Z < -\left|\frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}\right|\right) \quad + \quad P\left(Z > \left|\frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}\right|\right).$$

If *p-value*$\leq \alpha$, then we can reject the null hypothesis, otherwise we can't reject it.

Relationship between two-sided hypothesis test and confidence interval. We note that the rejection region for the two-sided test at level α is

$$z = \left| \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\frac{\alpha}{2}} ,$$

and this can be manipulated to give either

$$\mu_0 < \bar{y} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \mu_0 < \bar{y} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} .$$

We see that if we reject $H_0 : \mu = \mu_0$ at the level α, then μ_0 lies outside the $(1-\alpha) \times 100\%$ confidence interval for μ. Similarly we can show that if we accept $H_0 : \mu = \mu_0$ at level α, then μ_0 lies inside $(1-\alpha) \times 100\%$ confidence interval for μ. So the confidence interval contains all those values of μ_0 that would be accepted if tested for.

Bayesian Two-Sided Hypothesis Test about μ

If we wish to test the two-sided hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

in a Bayesian manner, and we have a continuous prior, we can't calculate the posterior probability of the null hypothesis as we did for the one-sided hypothesis. Since we have a continuous prior, we have a continuous posterior. We know that the probability of any specific value of a continuous random variable always equals 0. The posterior probability of the null hypothesis $H_0 : \mu = \mu_0$ will equal zero. This means we can't test this hypothesis by calculating the posterior probability of the null hypothesis and comparing it to α.

Instead, we calculate a $(1-\alpha) \times 100\%$ credible interval for μ using our posterior distribution. If μ_0 lies inside the credible interval, we conclude that μ_0 still has credibility as a possible value. In that case we will not reject the null hypothesis $H_0 : \mu = \mu_0$, so we consider that it is credible that there is no effect. (However, we realize it has zero probability of being exactly true if we look at enough decimal places.) There is no need to search for an explanation of a nonexistent effect. However, if μ_0 lies outside the credible interval we conclude that μ_0 does not have credibility as a possible value, and we will reject the null hypothesis. Then it is reasonable to attempt to explain why the mean has shifted from μ_0 for this experiment.

Main Points

- When we have prior information on the values of the parameter that are realistic, we can find a prior distribution so that the mean of the posterior distribution of μ (the Bayesian estimator) has a smaller mean squared error than the sample mean (the frequentist estimator) over the range of realistic values. This means that on the average, it will be closer to the true value of the parameter.

- A confidence interval for μ is found by inverting a probability statement for $\bar{y}$, and plugging in the sample value to compute the endpoints. It is called a confidence interval because there is nothing left to be random, so no probability statement can be made after the sample value is plugged in.

- The interpretation of a $(1-\alpha)\times 100\%$ frequentist confidence interval for μ is that $1-\alpha)\times 100\%$ of the random intervals calculated this way would cover the true parameter, so we are $(1-\alpha)\times 100\%$ *confident* that the interval we calculated does.

- A $(1-\alpha)\times 100\%$ Bayesian credible interval is an interval such that the posterior probability it contains the random parameter is $(1-\alpha)\times 100\%$.

- This is more useful to the scientist because he/she is only interested in his/her particular interval.

- The $(1-\alpha)\times 100\%$ frequentist confidence interval for μ corresponds to the $(1-\alpha)\times 100\%$ Bayesian credible interval for μ when we used the "flat prior." So, in this case, frequentist statisticians can get away with misinterpreting their confidence interval for μ as a probability interval.

- In the general, misinterpreting a frequentist confidence interval as a probability interval for the parameter will be wrong.

- Hypothesis testing is how we protect our credibility, by not attributing an effect to a cause if that effect could be due to chance alone.

- If we are trying to detect an effect in one direction, say $\mu > \mu_0$, we set this up as the one-sided hypothesis test

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0 \, .$$

  Note that the alternative hypothesis contains the effect we wish to detect. The null hypothesis is that the mean is still at the old value (or is changed in the direction we aren't interested in detecting).

- If we are trying to detect an effect in either direction, we set this up as the two-sided hypothesis test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0 \, .$$

  The null hypothesis contains only a single value μ_0 and is called a point hypothesis.

- Frequentist hypothesis tests are based on the sample space.

- The level of significance α is the low probability we allow for rejecting the null hypothesis when it is true. We choose α.

- A frequentist hypothesis test divides the sample space into a rejection region, and an acceptance region such that the probability the test statistic lies in the rejection region if the null hypothesis is true is less than the level of significance α. If the test statistic falls into the rejection region we reject the null hypothesis at level of significance α.
- Or we could calculate the *p-value*. If the *p-value*$< \alpha$, we reject the null hypothesis at level α.
- The *p-value* is not the probability the null hypothesis is true. Rather it is the probability of observing what we observed, or even something more extreme, given the null hypothesis is true.
- We can test a one-sided hypothesis in a Bayesian manner by computing the posterior probability of the null hypothesis by integrating the posterior density over the null region. If this probability is less than the level of significance α, then we reject the null hypothesis.
- We cannot test a two-sided hypothesis by integrating the posterior probability over the null region because with a continuous prior, the prior probability of a point null hypothesis is zero, so the posterior probability will also be zero. Instead, we test the credibility of the null value by observing whether or not it lies within the Bayesian credible interval. If it does, the null value remains credible and we can't reject it.

Exercises

11.1 A statistician buys a pack of 10 new golf balls, and drops each golf ball from a height of one meter, and measures the height in centimeters it returns on the first bounce. The ten values are:

79.9	80.0	78.9	78.5	75.6	80.5	82.5	80.1	81.6	76.7

Assume y, the height (in cm) a golf ball bounces when dropped from a one meter height is *normal* (μ, σ^2), where the standard deviation $\sigma = 2$.

(a) Assume a *normal* $(75, 10^2)$ prior for μ. Find the posterior distribution of μ.

(b) Calculate a 95% Bayesian credible interval for μ.

(c) Perform a Bayesian test of the hypothesis

$$H_0 : \mu \geq 80 \quad \text{versus} \quad H_1 : \mu < 80$$

at the 5% level of significance.

11.2 The statistician buys ten used balls that have been recovered from a water hazard. He drops each from a height of one meter and measures the height in centimeters it returns on the first bounce. The values are:

73.1	71.2	69.8	76.7	75.3	68.0	69.2	73.4	74.0	78.2

Assume y, the height (in cm) a golf ball bounces when dropped from a one meter height is *normal* (μ, σ^2), where the standard deviation $\sigma = 2$.

(a) Assume a *normal* $(75, 10^2)$ prior for μ. Find the posterior distribution of μ.

(b) Calculate a 95% Bayesian credible interval for μ.

(c) Perform a Bayesian test of the hypothesis

$$H_0 : \mu \geq 80 \quad \text{versus} \quad H_1 : \mu < 80$$

at the 5% level of significance.

11.3 The local consumer watchdog group was concerned about the cost of electricity to residential customers over the New Zealand winter months (Southern Hemisphere). They took a random sample of 25 residential electricity accounts and looked at the total cost of electricity used over the three months of June, July, and August. The costs were:

514	536	345	440	427
443	386	418	364	483
506	385	410	561	275
306	294	402	350	343
480	334	324	414	296

Assume that the amount of electricity used over the three months by a residential account is *normal* (μ, σ^2), where the known standard deviation $\sigma = 80$.

(a) Use a *normal* $(325, 80^2)$ prior for μ. Find the posterior distribution for μ.

(b) Find a 95% Bayesian credible interval for μ.

(c) Perform a Bayesian test of the hypothesis

$$H_0 : \mu = 350 \quad \text{versus} \quad H_1 : \mu \neq 350$$

at the 5% level.

(d) Perform a Bayesian test of the hypothesis

$$H_0 : \mu \leq 350 \quad \text{versus} \quad H_1 : \mu > 350$$

at the 5% level.

11.4 A medical researcher collected the systolic blood pressure reading for a random sample of $n = 30$ female students under the age of 21 who visited the Student's Health Service. The blood pressures are:

120	122	121	108	133	119	136	108	106	105
122	139	133	115	104	94	118	93	102	114
123	125	124	108	111	134	107	112	109	125

Assume that systolic blood pressure comes from a *normal* (μ, σ^2) distribution where the standard deviation $\sigma = 12$ is known.

(a) Use a *normal* $(120, 15^2)$ prior for μ. Calculate the posterior distribution of μ.

(b) Find a 95% Bayesian credible interval for μ.

(c) Suppose we had not actually known the standard deviation σ. Instead, the value $\hat{\sigma} = 12$ was calculated from the sample and used in place of the unknown true value. Recalculate the 95% Bayesian credible interval.

12

Bayesian Inference for Difference between Means

Comparisons are the main tool of experimental science. When there is uncertainty present due to observation errors or experimental unit variation, comparing observed values can't establish the existence of a difference because of the uncertainty within each of the observations. Instead, we must compare the means of the two distributions the observations came from. In many cases the distributions are normal, so we are comparing the means of two normal distributions. There are two experimental situations that the data could arise from.

The most common experimental situation is where there are independent random samples from each distribution. The treatments have been applied to different random samples of experimental units. The second experimental situation is where the random samples are paired. It could be that the two treatments have been applied to the same set of experimental units (at separate times). The two measurements on the same experimental unit cannot be considered independent. Or it could be that the experimental units were formed into pairs of similar units, and one of each pair randomly assigned to each treatment group. Again, the two measurements in the same pair cannot be considered independent. We say the observations are paired. The random samples from the two populations are dependent.

In Section 12.1 we look at how to analyze data from independent random samples. If the treatment effect is an additive constant, we get equal variances for the two distributions. If the treatment effect is random, not constant, we get unequal variances for the two distributions. In Section 12.2 we investigate the case where we have independent random samples from two normal distributions with equal variances. In

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

Section 12.3, we investigate the case where we have independent random samples from two normal distributions with unequal variances. In Section 12.4 we investigate how to find the difference between proportions using the normal approximation, when we have independent random samples. In Section 12.5 we investigate the case where we have paired samples.

12.1 INDEPENDENT RANDOM SAMPLES FROM TWO NORMAL DISTRIBUTIONS

We may want to determine whether or not a treatment is effective in increasing growth rate in lambs. We know that lambs vary in their growth rate. Each lamb in a flock is randomly assigned to either the treatment group or the control group that will not receive the treatment. The assignments are done independently. This is called a completely randomized design, and we discussed it in Chapter 2. The reason the assignments are done this way is that any differences among lambs enters the treatment group and control group randomly. There will be no bias in the experiment. On average, both groups have been assigned similar groups of lambs over the whole range of the flock. The distribution of underlying growth rates for lambs in each group is assumed to be normal with the same means and variances σ^2. The means and variances for the two groups are equal because the assignment is done randomly.

The mean growth rate for a lamb in the treatment group, μ_1, equals the mean underlying growth rate plus the treatment effect for that lamb. The mean growth rate for a lamb in the control group, μ_2, equals the mean underlying growth rate plus zero, since the control group doesn't receive the treatment. Adding a constant to a random variable doesn't change the variance, so if the treatment effect is constant for all lambs, the variances of the two groups will be equal. We call that an *additive* model. If the treatment effect is different for different lambs, the variances of the two groups will be unequal. This is called a *nonadditive* model.

If the treatment is effective, μ_1 will be greater than μ_2. In this chapter we will develop Bayesian methods for inference about the difference between means $\mu_1 - \mu_2$ for both additive and nonadditive models.

12.2 CASE 1: EQUAL VARIANCES

We often assume the treatment effect is the same for all units. The observed value for a unit given the treatment is the mean for that unit plus the constant treatment effect. Adding a constant doesn't change the variance, so the variance of the treatment group is equal to the variance of the control group. That sets up an *additive* model.

When the Variance Is Known

Suppose we know the variance σ^2. Since we know the two samples are independent of each other, we will use independent priors for both means. They can either be

normal (m_1, s_1^2) and *normal*(m_2, s_2^2) priors, or we can use *flat* priors for one or both of the means.

Because the priors are independent, and the samples are independent, the posteriors are also independent. The posterior distributions are

$$\mu_1|y_{11}, \cdots, y_{n_1 1} \sim Normal(m_1', (s_1')^2)$$

and

$$\mu_2|y_{12}, \cdots, y_{n_2 2} \sim Normal(m_2', (s_2')^2)\,,$$

where the $m_1', (s_1')^2, m_2'$, and $(s_2')^2$ are found using the simple updating formulas given by Equations 10.5 and 10.6.

Since $\mu_1|y_{11}, \cdots, y_{n_1 1}$ and $\mu_2|y_{12}, \cdots, y_{n_2 2}$ are independent of each other, we can use the rules for mean and variance of a difference between independent random variables. This gives the posterior distribution of $\mu_d = \mu_1 - \mu_2$. It is

$$\mu_d|y_{11}, \cdots, y_{n_1 1}, y_{12}, \cdots, y_{n_2 2} \sim Normal(m_d', (s_d')^2)\,,$$

where $m_d' = m_1' - m_2'$, and $(s_d')^2 = (s_1')^2 + (s_2')^2)$. We can use this posterior distribution to make further inferences about the difference between means $\mu_1 - \mu_2$.

Credible interval for difference between means, known equal variance case. The general rule for finding a $(1-\alpha)\times 100\%$ Bayesian credible interval when the posterior distribution is *normal* $(m', (s')^2)$ is to take the *posterior mean* $\pm$ *critical value* $\times$ *posterior standard deviation*. When the observation variance (or standard deviation) is assumed known, the critical value comes from the standard normal table. In that case the $(1-\alpha)\times 100\%$ Bayesian credible interval for $\mu_d = \mu_1 - \mu_2$ is

$$m_d' \pm z_{\frac{\alpha}{2}} \times s_d'\,. \tag{12.1}$$

This can be written as

$$m_1' - m_2' \ \pm z_{\frac{\alpha}{2}} \times \sqrt{(s_1')^2 + (s_2')^2}\,. \tag{12.2}$$

Thus, given the data, the probability that $\mu_1 - \mu_2$ lies between the endpoints of the credible interval equals $(1-\alpha)\times 100\%$.

Confidence interval for difference between means, known equal variance case. The frequentist confidence interval for $\mu_d = \mu_1 - \mu_2$ when the two distributions have equal known variance is given by

$$\bar{y}_1 - \bar{y}_2 \ \pm z_{\frac{\alpha}{2}} \times \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,. \tag{12.3}$$

This is the same formula as the Bayesian credible interval would be if we had used independent "flat" priors for μ_1 and μ_2, but the interpretations are different. The endpoints of the confidence interval are what is random under the frequentist viewpoint. $(1-\alpha)\times 100\%$ of the intervals calculated using this formula would

contain the fixed, but unknown value $\mu_1 - \mu_2$. We would have that confidence that the particular interval we calculated using our data contains the true value.

Example 20 *In Example 3 (Chapter 3), we looked at two series of measurements Michelson made on the speed of light, in 1879 and 1882, respectively. The data are shown in Table 3.1. (The measurements are figures given plus 299,000.) Suppose we assume each speed of light measurement is normally distributed with known standard deviation 100. Let us use independent normal (m, s^2) priors for the 1879 and 1882 measurements, where $m = 300,000$ and $s^2 = 500^2$.*

The posterior distributions of μ_{1879} and μ_{1882} can be found using the updating rules. For μ_{1879} they give

$$\frac{1}{(s'_{1879})^2} = \frac{1}{500^2} + \frac{20}{100^2} = .002004\,,$$

so $(s'_{1879})^2 = 499$, and

$$m'_{1879} = \frac{\frac{1}{500^2}}{.002004} \times 300000 + \frac{\frac{20}{100^2}}{.002004} \times (299000 + 909) = 299909\,.$$

Similarly, for μ_{1882} they give

$$\frac{1}{(s'_{1882})^2} = \frac{1}{500^2} + \frac{23}{100^2} = .002304\,,$$

so $(s'_{1882})^2 = 434$, and

$$m'_{1882} = \frac{\frac{1}{500^2}}{.002304} \times 300000 + \frac{\frac{23}{100^2}}{.002304} \times (299000 + 756) = 299757\,.$$

The posterior distribution of $\mu_d = \mu_{1879} - \mu_{1882}$ will be normal $(m'_d, (s'_d)^2)$ where

$$m'_d = 299909 - 299757 = 152$$

and

$$(s'_d)^2 = 499 + 434 = 30.5^2\,.$$

The 95% Bayesian credible interval for $\mu_d = \mu_{1879} - \mu_{1882}$ is

$$152 \pm 1.96 \times 30.5 = (92.1, 211.9)\,.$$

One-sided Bayesian hypothesis test. If we wish to determine whether or not the treatment mean μ_1 is greater than the control mean μ_2, we will use hypothesis testing. We test the null hypothesis

$$H_0 : \mu_d \leq 0 \quad \text{versus} \quad H_1 : \mu_d > 0$$

where $\mu_d = \mu_1 - \mu_2$ is the difference between the two means. To do this test in a Bayesian manner, we calculate the posterior probability of the null hypothesis $P(\mu_d \leq 0|data)$ where $data$ includes the observations from both samples

$y_{11}, \ldots, y_{n_1 1}$ and $y_{12}, \ldots, y_{n_2 2}$. Standardizing by subtracting the mean and dividing by the standard deviation gives

$$\begin{aligned} P(\mu_d \leq 0|data) &= P\left(\frac{\mu_d - m'_d}{s'_d} \leq \frac{0 - m'_d}{s'_d}\right) \\ &= P\left(Z \leq \frac{0 - m'_d}{s'_d}\right), \end{aligned} \tag{12.4}$$

where Z has the standard normal distribution. We find this probability in Table B.2 in Appendix B. If it is less than α, we can reject the null hypothesis at that level. Then we can conclude that μ_1 is indeed greater than μ_2 at that level of significance.

Two-sided Bayesian hypothesis test. We can't test the two-sided hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

in a Bayesian manner by calculating the posterior probability of the null hypothesis. It is a point null hypothesis since it is only true for a single value $\mu_d = \mu_1 - \mu_2 = 0$. When we used the continuous prior, we got a continuous posterior, and the probability that any continuous random variable takes on any particular value always equals 0.

Instead, we use the credible interval for μ_d. If 0 lies in the interval, we cannot reject the null hypothesis and 0 remains a credible value for the difference between the means. However, if 0 lies outside the interval, then 0 is no longer a credible value at the significance level α.

Example 20 (continued) *The* 95% *Bayesian credible interval for* $\mu_d = \mu_{1879} - \mu_{1882}$ *is* $(92.1, 211.9)$. *0 lies outside the interval; hence we reject the null hypothesis that the means for the two measurement groups were equal, and conclude that they are different. This shows that there was a bias in Michelson's first group of measurements, which was very much reduced in the second group of measurements.*

When the Variance Is Unknown and Flat Priors Are Used

Suppose we use independent "flat" priors for μ_1 and μ_2. Then $(s'_1)^2 = \frac{\sigma^2}{n_1}$, $(s'_2)^2 = \frac{\sigma^2}{n_2}$, $m'_1 = \bar{y}_1$ and $m'_2 = \bar{y}_2$.

Credible interval for difference between means, unknown equal variance case. If we knew the variance σ^2 the credible interval could be written as

$$\bar{y}_1 - \bar{y}_2 \ \pm z_{\frac{\alpha}{2}} \times \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,.$$

However, we don't know σ^2. We will have to estimate it from the data. We can get an estimate from each of the samples. The best thing to do is to combine these estimates to get the pooled variance estimate

$$\hat{\sigma}_p^2 = \frac{\sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{j2} - \bar{y}_2)^2}{n_1 + n_2 - 2}\,. \tag{12.5}$$

Since we used the estimated $\hat{\sigma}_p^2$ instead of the unknown true variance σ^2, the credible interval should be widened to allow for the additional uncertainty. We will get the critical value from the *Student's t* table with $n_1 + n_2 - 2$ degrees of freedom. The approximate $(1-\alpha) \times 100\%$ Bayesian credible interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \ \pm t_{\frac{\alpha}{2}} \times \hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,, \tag{12.6}$$

where the critical value comes from the *Student's t* table with $n_1 + n_2 - 2$ degrees of freedom[1].

Confidence interval for difference between means, unknown equal variance case. The frequentist confidence interval for $\mu_d = \mu_1 - \mu_2$ when the two distributions have equal unknown variance is

$$\bar{y}_1 - \bar{y}_2 \ \pm t_{\frac{\alpha}{2}} \times \hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,, \tag{12.7}$$

where the critical value again comes from the *Student's t* table with $n_1 + n_2 - 2$ degrees of freedom. The confidence interval has exactly the same form as the Bayesian credible interval when we use independent "flat" priors for μ_1 and μ_2. Of course, the interpretations are different.

The frequentist has $(1-\alpha) \times 100\%$ confidence that the interval contains the true value of the difference because $(1-\alpha) \times 100\%$ of the random intervals calculated this way do contain the true value. The Bayesian interpretation is that given the data from the two samples, the posterior probability the random parameter $\mu_1 - \mu_2$ lies in the interval is $(1-\alpha)$.

In this case the scientist who misinterprets the confidence interval for a probability statement about the parameter gets away with it, because it actually is a probability statement using independent *flat* priors. It is fortunate for frequentist statisticians that their most commonly used techniques (confidence intervals for means and proportions) are equivalent to Bayesian credible intervals for some specific prior[2]. Thus a scientist who misinterpret his/her confidence interval as a probability statement, can do so in this case, but he/she is implicitly assuming independent flat priors. The

[1]Actually, we are treating the unknown σ^2 as a nuisance parameter, and using an independent prior $g(\sigma^2) \propto \frac{1}{\sigma^2}$ for it. We find the marginal posterior distribution of $\mu_1 - \mu_2$ from the joint posterior of $\mu_1 - \mu_2$ and σ^2 by integrating out the nuisance parameter. The marginal posterior will be *Student's t* with $n_1 + n_2 - 2$ degrees of freedom instead of normal. This gives us the credible interval with the z critical value replaced by the t critical value. We see that our approximation gives us the correct credible interval for these assumptions.

[2]In the case of a single random sample from a *normal* distribution, frequentist confidence intervals are equivalent to Bayesian credible intervals with *flat* prior for μ. In the case of independent random samples from *normal* distributions having equal unknown variance σ^2, confidence intervals for the difference between means are equivalent to Bayesian credible intervals with independent flat priors for μ_1 and μ_2, and the improper prior $g(\sigma) \propto \sigma^{-1}$ for the nuisance parameter.

only loss that the scientist will have incurred is he/she didn't get to use any prior information he/she may have had[3].

One-sided Bayesian hypothesis test. If we want to test

$$H_0 : \mu_d \leq 0 \quad \text{versus} \quad H_1 : \mu_d > 0$$

when we assume that the two random samples come from *normal* distributions having the same unknown variance σ^2, and we use the pooled estimate of the variance $\hat{\sigma}_p^2$ in place of the unknown σ^2 and assume independent "flat" priors for the means μ_1 and μ_2, we calculate the posterior probability of the null hypothesis using Equation 12.4, but instead of finding the probability in the standard normal table, we find it from the *Student's t* distribution with $n_1 + n_2 - 2$ degrees of freedom. We could calculate it using Minitab or R, or alternatively, we could find values that bound this probability in the *Student's t* table.

Two-sided Bayesian hypothesis test. When we assume both samples come from *normal* distributions with equal unknown variance σ^2, and we use the pooled estimate of the variance $\hat{\sigma}_p^2$ in place of the unknown variance σ^2 and assume independent "flat" priors, we can test the two-sided hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

using the credible interval for $\mu_1 - \mu_2$ given in Equation 12.6. There are $n_1 + n_2 - 2$ degrees of freedom. If 0 lies in the credible interval, we cannot reject the null hypothesis, and 0 remains a credible value for the difference between the means. However, if 0 lies outside the interval, then 0 is no longer a credible value at the significance level α.

12.3 CASE 2: UNEQUAL VARIANCES

When the Variances Are Known

In this section we will look at a nonadditive model, but with known variances. Let $y_{11}, \ldots, y_{n_1 1}$ be a random sample from normal distribution having mean μ_1 and known variance σ_1^2. Let $y_{12}, \ldots y_{n_2 2}$ be a random sample from normal distribution having mean μ_2 and known variance σ_2^2. The two random samples are independent of each other.

We use independent priors for μ_1 and μ_2. They can be either normal priors or "flat" priors. Since the samples are independent, and the priors are independent, we can find each posterior independently of the other. We find these using the simple updating formulas given in Equations 10.5 and 10.6. The posterior of $\mu_1 | y_{11}, \cdots, y_{n_1 1}$ is

[3] Frequentist techniques such as the confidence intervals used in many other situations do not have Bayesian interpretations. Interpreting the confidence interval as the basis for a probability statement about the parameter would be completely wrong in those situations.

normal$[m_1', (s_1')^2]$. The posterior of $\mu_2|y_{12}, \cdots, y_{n_2 2}$ is *normal*$[m_2', (s_2')^2]$. The posteriors are independent since the priors are independent and the samples are independent. The posterior distribution of $\mu_d = \mu_1 - \mu_2$ is normal with mean equal to the *difference* of the posterior means, and variance equal to the *sum* of the posterior variances.

$$(\mu_d|y_{11}, \cdots, y_{n_1 1}, y_{12}, \cdots, y_{n_2 2}) \sim Normal[m_d', (s_d')^2]\,,$$

where $m_d' = m_1' - m_2'$ and $(s_d')^2 = (s_1')^2 + (s_2')^2$

Credible interval for difference between means, known unequal variance case. A $(1-\alpha)\times 100\%$ Bayesian credible interval for $\mu_d = \mu_1 - \mu_2$, the difference between means is

$$m_d' \pm z_{\frac{\alpha}{2}} \times (s_d')\,, \tag{12.8}$$

which can be written as

$$m_1' - m_2' \pm z_{\frac{\alpha}{2}} \times \sqrt{(s_1')^2 + (s_2')^2}\,. \tag{12.9}$$

Note these are identical to Equations 12.1 and 12.2.

Confidence interval for difference between means, known unequal variance case. The frequentist confidence interval for $\mu_d = \mu_1 - \mu_2$ in this case would be

$$\bar{y}_1 - \bar{y}_2 \pm z_{\frac{\alpha}{2}} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,. \tag{12.10}$$

Note that this is identical to the Bayesian credible interval we would get if we had used flat priors for both μ_1 and μ_2. However, the interpretations are different.

When the Variances Are Unknown

When the variances are unequal and unknown, each of them will have to be estimated from the sample data

$$\hat{\sigma}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 \quad \text{and} \quad \hat{\sigma}_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{i2} - \bar{y}_2)^2\,.$$

These estimates will be used in place of the unknown true values in the simple updating formulas. This adds extra uncertainty. To allow for this, we should use the *Student's t* table to find the critical values. However, it is no longer straightforward what degrees of freedom should be used. Satterthwaite suggested that the adjusted degrees of freedom be

$$\frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{(\hat{\sigma}_1^2/n_1)^2}{n_1+1} + \frac{(\hat{\sigma}_2^2/n_2)^2}{n_2+1}}$$

rounded down to the nearest integer.

Credible interval for difference between means, unequal unknown variances. When we use the sample estimates of the variances in place of the true unknown variances in Equations 10.5 and 10.6, an approximate $(1-\alpha)\times 100\%$ credible interval for $\mu_d = \mu_1 - \mu_2$ is given by

$$m_1' - m_2' \pm t_{\frac{\alpha}{2}} \times \sqrt{(s_1')^2 + (s_2')^2}\,,$$

where we find the degrees of freedom using Satterthwaites adjustment. In the case where we use independent "flat" priors for μ_1 and μ_2, this can be written as

$$m_1' - m_2' \pm t_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}\,. \tag{12.11}$$

Confidence interval for difference between means, unequal unknown variances. An approximate $(1-\alpha)\times 100\%$ confidence interval for $\mu_d = \mu_1 - \mu_2$ is given by

$$m_1' - m_2' \pm t_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}\,. \tag{12.12}$$

We see this is the same form as the $(1-\alpha)\times 100\%$ credible interval found when we used independent flat priors[4]. However, the interpretations are different.

Bayesian hypothesis test of $H_0 : \mu_1 - \mu_2 \leq 0$ versus $H_1 : \mu_1 - \mu_2 > 0$. To test

$$H_0 : \mu_1 - \mu_2 \leq 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 > 0$$

at the level α in a Bayesian manner, we calculate the posterior probability of the null hypothesis. We would use Equation 12.4. If the variances σ_1^2 and σ_2^2 are

[4]Finding the posterior distribution of $\mu_1 - \mu_2 - (\bar{y}_1 - \bar{y}_2)|y_{11}, \cdots, y_{n_1 1}, y_{12}, \cdots, y_{n_2 2}$ in the Bayesian paradigm, or equivalently finding the sampling distribution of $\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)$ in the frequentist paradigm when the variances are both unknown and not assumed equal has a long and controversial history. In the one sample case, the sampling distribution of $\bar{y} - \mu$ is the same as the posterior distribution of $\mu - \bar{y}|y_1, \ldots, y_n$ when we use the flat prior for $g(\mu) = 1$ and the noninformative prior $g(\sigma^2) \propto \frac{1}{\sigma^2}$, and marginalize σ^2 out of the joint posterior. This leads to the equivalence between the confidence interval and the credible interval for that case. Similarly, in the two-sample case with equal variances, the sampling distribution of $\bar{y}_1 - \bar{y}_2$ equals the posterior distribution of $\mu_1 - \mu_2|y_{11}, \ldots, y_{n_1 1}, y_{12}, \ldots, y_{n_2 2}$ where we use flat priors for μ_1 and μ_2 and the noninformative prior $g(\sigma^2) \propto \frac{1}{\sigma^2}$, and marginalized σ^2 out of the joint posterior. Again, that led to the equivalence between the confidence interval and the credible interval for that case. One might be led to believe this pattern would hold in general. However, it doesn't hold in the two sample case with unknown unequal variances. The Bayesian posterior distribution in this case is known as the *Behrens-Fisher* distribution. The frequentist distribution depends on the ratio of the unknown variances. Both of the distributions can be approximated by *Student's t* with an adjustment made to the degrees of freedom. Satterthwaite suggested that the adjusted degrees of freedom be

$$\frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{(\hat{\sigma}_1^2/n_1)^2}{n_1+1} + \frac{(\hat{\sigma}_2^2/n_2)^2}{n_2+1}}$$

rounded down to the nearest integer.

known, we get the critical value from the standard normal table. However, when we use estimated variances instead of the true unknown variances, we will find the probabilities using the *Student's t* distribution with degrees of freedom given by Satterthwaites approximation. If this probability is less than α, then we reject the null hypothesis and conclude that $\mu_1 > \mu_2$. In other words, that the treatment is effective. Otherwise, we can't reject the null hypothesis.

12.4 BAYESIAN INFERENCE FOR DIFFERENCE BETWEEN TWO PROPORTIONS USING NORMAL APPROXIMATION

Often we want to compare the proportions of a certain attribute in two populations. The true proportions in population 1 and population 2 are π_1 and π_2, respectively. We take a random sample from each of the populations and observe the number of each sample having the attribute. The distribution of $y_1|\pi_1$ is *binomial*(n_1, π_1) and the distribution of $y_2|\pi_2$ is *binomial*$(\ n_2, \pi_2)$, and they are independent of each other

We know that if we use independent prior distributions for π_1 and π_2, we will get independent posterior distributions. Let the prior for π_1 be *beta*(a_1, b_1) and for π_2 be *beta*(a_2, b_2). The posteriors are independent beta distributions. The posterior for π_1 is *beta*(a_1', b_1'), where $a_1' = a_1 + y_1$ and $b_1' = b_1 + n_1 - y_1$. Similarly the posterior for π_2 is *beta*(a_2, b_2), where $a_2' = a_2 + y_2$ and $b_2' = b_2 + n_2 - y_2$

Approximate each posterior distribution with the normal distribution having same mean and variance as the beta. The posterior distribution of $\pi_d = \pi_1 - \pi_2$ is approximately *normal*$(m_d', (s_d')^2)$ where the posterior mean is given by

$$m_d' = \frac{a_1'}{a_1' + b_1'} - \frac{a_2'}{a_2' + b_2'}$$

and the posterior variance is given by

$$(s_d')^2 = \frac{a_1' b_1'}{(a_1' + b_1')^2(a_1' + b_1' + 1)} + \frac{a_2' b_2'}{(a_2' + b_2')^2(a_2' + b_2' + 1)}\,.$$

Credible interval for difference between proportions. We find the $(1 - \alpha) \times 100\%$ Bayesian credible interval for $\pi_d = \pi_1 - \pi_2$ using the general rule for the (approximately) normal posterior distribution. It is

$$m_d' \pm z_{\frac{\alpha}{2}} \times s_d'\,. \tag{12.13}$$

One-sided Bayesian hypothesis test for difference between proportions. Suppose we are trying to detect whether $\pi_d = \pi_1 - \pi_2 > 0$. We set this up as a test of

$$H_0 : \pi_d \leq 0 \quad \text{versus} \quad H_1 : \pi_d > 0\,.$$

Note, the alternative hypothesis is what we are trying to detect. We calculate the approximate posterior probability of the null distribution by

$$P(Z = \pi_d \leq 0) \quad = \quad P\left(\frac{\pi_d - m_d'}{s_d'} \leq \frac{0 - m_d'}{s_d'}\right) \tag{12.14}$$

$$= \quad P\left(Z \leq \frac{0 - m'_d}{s'_d}\right) .$$

If this probability is less than the level of significance α that we chose, we would reject the null hypothesis at that level, and conclude $\pi_1 > \pi_2$. Otherwise, we can't reject the null hypothesis.

Two-sided Bayesian hypothesis test for difference between proportions. To test the hypothesis

$$H_0 : \pi_1 - \pi_2 = 0 \quad \text{versus} \quad H_1 : \pi_1 - \pi_2 \neq 0$$

in a Bayesian manner check whether the null hypothesis value (0) lies inside the credible interval for π_d given in Equation 12.13. If it lies inside the interval, we cannot reject the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$ at the level α. If it lies outside the interval, we can reject the null hypothesis at the level α and accept the alternative $H_1 : \pi_1 - \pi_2 \neq 0$.

Example 21 *The student newspaper wanted to write an article on the smoking habits of students A random sample of 200 students (100 males and 100 females) between ages of 16 and 21 were asked about whether they smoked cigarettes. Out of the 100 males, 22 said they were regular smokers, and out of the 100 females, 31 said they were regular smokers. The editor of the paper asked Donna, a statistics student, to analyze the data.*

Donna considered the male and female samples would be independent. Her prior knowledge was that a minority of students smoked cigarettes, so she decided to use independent beta(1,2) priors for π_m and π_f, the male and female proportions respectively. Her posterior distribution of π_m will be beta(23,80), and her posterior distribution of π_f will be beta(32,71). Hence, her posterior distribution of the difference between proportions, $\pi_d = \pi_m - \pi_f$, will be approximately normal$(m'_d, (s'_d)^2)$ where

$$\begin{aligned} m'_d &= \frac{23}{23+80} - \frac{32}{32+71} \\ &= -.087 \end{aligned}$$

and

$$\begin{aligned} (s'_d)^2 &= \frac{23 * 80}{(23+80)^2 * (23+80+1)} + \frac{32 * 71}{(32+71)^2 * (32+71+1)} \\ &= .061^2 . \end{aligned}$$

Her 95 % credible interval for π_d will be (-.207, .032) which contains 0. She can't reject the null hypothesis $H_0 : \pi_m - \pi_f = 0$ at the 5% level, so she tells the editor that the data does not conclusively show that there is any difference between the proportions of male and female students who smoke.

12.5 NORMAL RANDOM SAMPLES FROM PAIRED EXPERIMENTS

Variation between experimental units often is a major contributor to the variation in the data. When the two treatments are administered to two independent random samples of the experimental units, this variation makes it harder to detect any difference between the treatment effects, if one exists.

Often designing a paired experiment makes it much easier to detect the difference between treatment effects. For a paired experiment, the experimental units are matched into pairs of similar units. Then one of the units from each pair is assigned to the first treatment, and the other in that pair is assigned the second treatment. This is a *randomized block* experimental design, where the pairs are blocks. We discussed this design in Chapter 2. For example, in the dairy industry, identical twin calves are often used for experiments. They are exact genetic copies. One of each pair is randomly assigned to the first treatment, and the other to the second treatment.

Paired data can arise other ways. For instance, if the two treatments are applied to the same experimental units (at different times) giving the first treatment effect time to dissipate before the second treatment is applied. Or, we can be looking at "before treatment" and "after treatment" measurements on the same experimental units.

Because of the variation between experimental units, the two observations from units in the same pair will be more similar than two observations from units in different pairs. In the same pair, the only difference between the observation given treatment A and the observation given treatment B is the treatment effect plus the measurement error. In different pairs, the difference between the observation given treatment A and the observation given treatment B is the treatment effect plus the experimental unit effect plus the measurement error. Because of this we cannot treat the paired random samples as independent of each other. The two random samples come from *normal* populations with means μ_A and μ_B, respectively. The populations will have equal variances σ^2 when we have an additive model. We consider that the variance comes from two sources, measurement error plus random variation between experimental units.

Take Differences within Each Pair

Let y_{i1} be the observation from pair i given treatment A, and y_{i2} be the observation from pair i given treatment B. If we take the difference between the observations within each pair, $d_i = y_{i1} - y_{i2}$, then these d_i will be a random sample from a *normal* population with mean $\mu_d = \mu_A - \mu_B$, and variance σ_d^2. We can treat this (differenced) data as a sample from a single *normal* distribution and do inference using techniques found in Chapters 10 and 11.

Example 22 *An experiment was designed to determine whether a mineral supplement was effective in increasing annual yield in milk. Fifteen pairs of identical twin dairy cows were used as the experimental units. One cow from each pair was randomly assigned to the treatment group that received the supplement. The other cow from the pair was assigned to the control group that did not receive the supplement.*

Table 12.1 Milk annual yield

Twin Set	Milk Yield: Control (liters)	Milk Yield: Treatment (liters)
1	3525	3340
2	4321	4279
3	4763	4910
4	4899	4866
5	3234	3125
6	3469	3680
7	3439	3965
8	3658	3849
9	3385	3297
10	3226	3124
11	3671	3218
12	3501	3246
13	3842	4245
14	3998	4186
15	4004	3711

The annual yields are given in Table 12.1. Assume that the annual yields from cows receiving the treatment are normal (μ_t, σ_t^2)*, and that the annual yields from the cows in the control group are normal* (μ_c, σ_c^2)*. Aleece, Brad, and Curtis decided that since the two cows in the same pair share identical genetic background, their responses will be more similar than two cows that were from different pairs. There is natural pairing. As the samples drawn from the two populations cannot be considered independent of each other, they decided to take differences* $d_i = y_{i1} - y_{i2}$*. The differences will be normal* (μ_d, σ_d^2)*, where* $\mu_d = \mu_t - \mu_c$ *and we will assume that* $\sigma_d^2 = 270^2$ *is known.*

Aleece decided she would use a "flat" prior for μ_d*. Brad decided he would use a normal* (m, s^2) *prior for* μ_d *where he let* $m = 0$ *and* $s = 200$*. Curtis decided that his prior for* μ_d *matched a triangular shape. He set up a numerical prior that interpolated between the heights given in Table 12.2 The shapes of the priors are shown in Figure 12.1.*

Aleece used a "flat" prior, so her posterior will be $normal[m', (s')^2]$ *where* $m' = \bar{y} = 7.07$ *and* $(s')^2 = 270^2/15 = 4860$*. Her posterior standard deviation* $s' = \sqrt{4860} = 69.71$*. Brad used a normal* $(0, 200^2)$ *prior, so his posterior will be normal* $[m', (s')^2]$ *where* m' *and* s' *are found by using Equations 10.5 and 10.6.*

$$\frac{1}{(s')^2} = \frac{1}{200^2} + \frac{15}{270^2} = 0.000230761\,,$$

Table 12.2 Curtis' prior weights. His continuous prior is found by linearly interpolating between them.

value	weight
-300	0
0	3
300	0

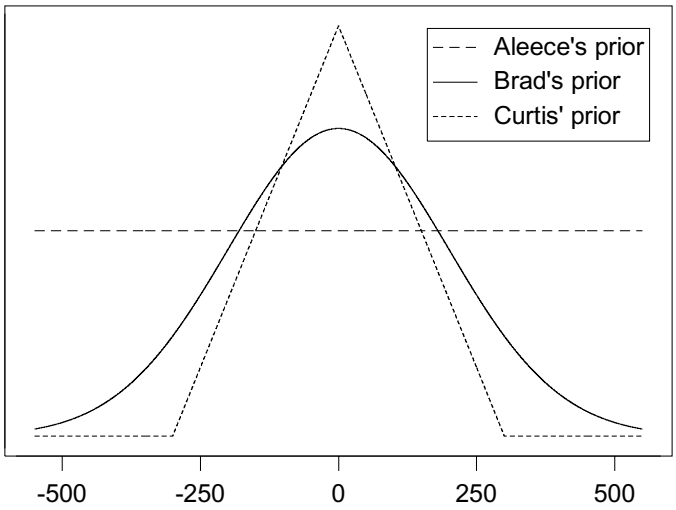


Figure 12.1 Aleece's, Brad's and Curtis' prior distributions.

so his $s' = 65.83$*, and*

$$m' = \frac{\frac{1}{200^2}}{.000230761} \times 0 + \frac{\frac{15}{270^2}}{.000230761} \times 7.07 = 6.33\,.$$

Curtis has to find his posterior numerically using Equation 10.3. He uses the Minitab macro NormGCP.mac to do the numerical integration. The three posteriors are shown in Figure 12.2.

They decided that to determine whether or not the treatment was effective in increasing the yield of milk protein, they would perform the one-sided hypothesis test

$$H_0 : \mu_d \leq 0 \quad \text{vs} \quad H_1 : \mu_d > 0$$

at the 95% level of significance. Aleece and Brad had normal posteriors, so they used Equation 11.4 to calculate the posterior probability of the null hypothesis. Curtis had a numerical posterior, so he used Equation 11.3 and performed the integration using the Minitab macro tintegral.mac. The results are shown in Table 12.3.

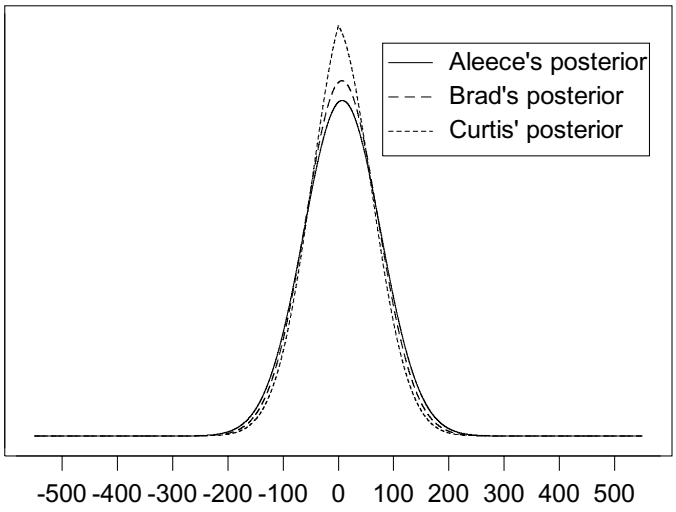


Figure 12.2 Aleece's, Brad's and Curtis's posterior distributions.

Table 12.3 Results of Bayesian one-sided hypothesis tests

person	posterior	$P(\mu_d \leq 0 \mid d_1, \ldots, d_n)$		
Aleece	*normal*$(7.07, 69.71^2)$	$P(Z \leq \frac{0-7.07}{69.71})$	=.4596	don't reject
Brad	*normal*$(6.33, 65.83^2)$	$P(Z \leq \frac{0-6.33}{65.83})$	=.4619	don't reject
Curtis	*numerical*	$\int_{\infty}^{0} g(\mu_d \mid d_1, \ldots, d_n) d\mu$	=.4684	don't reject

Main Points

- The difference between normal means are used to make inferences about the size of a treatment effect.
- Each experimental unit is randomly assigned to the treatment group or control group. The unbiased random assignment method ensures that both groups have similar experimental units assigned to them. On average, the means are equal.
- The treatment group mean is the mean of the experimental units assigned to the treatment group, plus the treatment effect.
- If the treatment effect is constant, we call it an additive model, and both sets of observations have the same underlying variance, assumed to be known.
- If the data in the two samples are independent of each other, we use independent priors for the two means. The posterior distributions $\mu_1 | y_{11}, \ldots, y_{n_1 1}$ and $\mu_2 | y_{12}, \ldots, y_{n_2 2}$ are also independent of each other, and can be found using methods from Chapter 10.

- Let $\mu_d = \mu_1 - \mu_2$. The posterior distribution of $\mu_d | y_{11}, \ldots, y_{n_1 1}, y_{12}, \ldots, y_{n_2 2}$ is *normal* with mean $m'_d = m'_1 - m'_2$ and variance $(s'_d)^2 = (s'_1)^2 + (s'_2)^2$

- The $(1-\alpha) \times 100\%$ credible interval for $\mu_d = \mu_1 - \mu_2$ is given by

$$m'_d \pm z_{\alpha/2} \times s'_d \,.$$

- If the variance is unknown, use the pooled estimate from the two samples. The credible interval will have to be widened to account for the extra uncertainty. This is accomplished by taking the critical values from the *Student's t* table (with $n_1 + n_2 - 2$ degrees of freedom) instead of the *standard normal* table.

- The confidence interval for $\mu_d | y_{11}, \ldots, y_{n_1 1}, y_{12}, \ldots, y_{n_2 2}$ is the same as the Bayesian credible interval where flat priors are used.

- If the variances are unknown, and not equal, use the sample estimates as if they were the correct values. Use the *Student's t* for critical values, with the degrees given by Satterthwaite's approximation. This is true for both credible intervals and confidence intervals.

- The posterior distribution for a difference between proportions can be found using the normal approximation. The posterior variances are known, so the critical values for credible interval come from *standard normal* table.

- When the observations are paired, the samples are dependent. Calculate the differences $d_i = y_{i1} - y_{i2}$ and treat them as a single sample from a *normal* (μ_d, σ_d^2), where $\mu_d = \mu_1 - \mu_2$. Inferences about μ_d are made using the single sample methods found in Chapters 10 and 11.

Exercises

12.1 The Human Resources Department of a large company wishes to compare two methods of training industrial workers to perform a skilled task. Twenty workers are selected, and 10 of them are randomly assigned to be trained using method A, and the other 10 are assigned to be trained using method B. After the training is complete, all the workers are tested on the speed of performance at the task. The times taken to complete the task are:

Method A	Method B
115	123
120	131
111	113
123	119
116	123
121	113
118	128
116	126
127	125
129	128

(a) We will assume that the observations come from *normal*(μ_A, σ^2) and *normal*(μ_B, σ^2), where $\sigma^2 = 6^2$. Use independent *normal* (m, s^2) prior distributions for μ_A and μ_B, respectively, where $m = 100$ and $s^2 = 20^2$. Find the posterior distributions of μ_A and μ_B, respectively.

(b) Find the posterior distribution of $\mu_A - \mu_B$.

(c) Find a 95% Bayesian credible interval for $\mu_A - \mu_B$.

(d) Perform a Bayesian test of the hypothesis

$$H_0 : \mu_A - \mu_B = 0 \quad \text{versus} \quad H_1 : \mu_A - \mu_B \neq 0$$

at the 5% level of significance. What conclusion can we draw?

12.2 A consumer testing organization obtained samples of size 12 from two brands of emergency flares, and measured the burning times. They are:

Brand A	Brand B
17.5	13.4
21.2	9.9
20.3	13.5
14.4	11.3
15.2	22.5
19.3	14.3
21.2	13.6
19.1	15.2
18.1	13.7
14.6	8.0
17.2	13.6
18.8	11.8

(a) We will assume that the observations come from *normal*(μ_A, σ^2) and *normal*(μ_B, σ^2), where $\sigma^2 = 3^2$. Use independent *normal* (m, s^2) prior distributions for μ_A and μ_B, respectively, where $m = 20$ and $s^2 = 8^2$. Find the posterior distributions of μ_A and μ_B, respectively.

(b) Find the posterior distribution of $\mu_A - \mu_B$.

(c) Find a 95% Bayesian credible interval for $\mu_A - \mu_B$.

(d) Perform a Bayesian test of the hypothesis

$$H_0 : \mu_A - \mu_B = 0 \quad \text{versus} \quad H_1 : \mu_A - \mu_B \neq 0$$

at the 5% level of significance. What conclusion can we draw?

12.3 The quality manager of a dairy company is concerned whether the levels of butterfat in a product are equal at two dairy factories which produce the product. He obtains random samples of size 10 from each of the factories output, and measures the butterfat. The results are:

Factory 1	Factory 2
16.2	16.1
12.7	16.3
14.8	14.0
15.6	16.2
14.7	15.2
13.8	16.5
16.7	14.4
13.7	16.3
16.8	16.9
14.7	13.7

(a) We will assume that the observations come from *normal*(μ_1, σ^2) and *normal*(μ_2, σ^2), where $\sigma^2 = 1.2^2$. Use independent *normal* (m, s^2) prior distributions for μ_1 and μ_2, respectively, where $m = 15$ and $s^2 = 4^2$. Find the posterior distributions of μ_1 and μ_2, respectively.

(b) Find the posterior distribution of $\mu_1 - \mu_1$.

(c) Find a 95% Bayesian credible interval for $\mu_1 - \mu_2$.

(d) Perform a Bayesian test of the hypothesis

$$H_0 : \mu_A - \mu_B = 0 \quad \text{versus} \quad H_1 : \mu_A - \mu_B \neq 0$$

at the 5% level of significance. What conclusion can we draw?

12.4 Independent random samples of ceramic produced by two different processes were tested for hardness. The results were:

Process 1	Process 2
8.8	9.2
9.6	9.5
8.9	10.2
9.2	9.5
9.9	9.8
9.4	9.5
9.2	9.3
10.1	9.2

(a) We will assume that the observations come from *normal*(μ_1, σ^2) and *normal*(μ_2, σ^2), where $\sigma^2 = .4^2$. Use independent *normal* (m, s^2) prior

distributions for μ_1 and μ_2, respectively, where $m = 10$ and $s^2 = 1^2$. Find the posterior distributions of μ_1 and μ_2, respectively.

(b) Find the posterior distribution of $\mu_1 - \mu_1$.

(c) Find a 95% Bayesian credible interval for $\mu_1 - \mu_2$.

(d) Perform a Bayesian test of the hypothesis

$$H_0 : \mu_1 - \mu_2 \geq 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 < 0$$

at the 5% level of significance. What conclusion can we draw?

12.5 A thermal power station discharges its cooling water into river. An environmental scientist wants to determine if this has adversely affected the dissolved oxygen level. She takes samples of water one kilometer upstream from the power station, and one kilometer downstream from the power station, and measures the dissolved oxygen level. The data are:

Upstream	Downstream
10.1	9.7
10.2	10.3
13.4	6.4
8.2	7.3
9.8	11.7
	8.9

(a) We will assume that the observations come from *normal*(μ_1, σ^2) and *normal*(μ_2, σ^2), where $\sigma^2 = 2^2$. Use independent *normal* (m, s^2) prior distributions for μ_1 and μ_2, respectively, where $m = 10$ and $s^2 = 2^2$. Find the posterior distributions of μ_1 and μ_2, respectively.

(b) Find the posterior distribution of $\mu_1 - \mu_1$.

(c) Find a 95% Bayesian credible interval for $\mu_1 - \mu_2$.

(d) Perform a Bayesian test of the hypothesis

$$H_0 : \mu_1 - \mu_2 \leq 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 > 0$$

at the 5% level of significance. What conclusion can we draw?

12.6 Cattle being ruminants have multiple chambers in their stomaches. Stimulating specific receptors causes reflex contraction of the reticular groove and swallowed fluid then bypasses the reticulo-rumen and moves directly to the abomasum. Scientists wanted to develop a simple nonradioactive, noninvasive test to determine when this occurs. In a study to determine the fate of swallowed fluids in cattle, McLeay, Carruthers, and Neil (1997) investigate a carbon

13 (^{13}C) octanoic acid breath test as a means of detecting a reticular groove contraction in cattle. Twelve adult cows were randomly assigned to two groups of 6 cows. The first group had 200 mg of ^{13}C octanoic acid administered into the reticulum, and the second group had the same dose of ^{13}C octanoic acid administered into the reticulo-osmasal orifice. Change in the enrichment of ^{13}C in breath was measured for each cow 10 minutes later. The results are:

^{13}C *Administered into Reticulum*		^{13}C *Administered into Reticulo-omasal Orifice*	
Cow ID	x	Cow ID	y
8	1.5	14	3.5
9	1.9	15	4.7
10	0.4	16	4.8
11	-1.2	17	4.1
12	1.7	18	4.1
13	0.7	19	5.3

(a) Explain why the observations of variables x and y can be considered independent in this experiment.

(b) Suppose the change in the enrichment of ^{13}C for cows administered in the *reticulum* is *normal* (μ_1, σ_1^2), where $\sigma_1^2 = 1.00^2$. Use a *normal* $(2, 2^2)$ prior for μ_1. Calculate the posterior distribution of $\mu_1|x_8 \ldots, x_{13}$.

(c) Suppose the change in the enrichment of ^{13}C for cows administered in the *reticulo-omasal orifice* is *normal* (μ_2, σ_2^2), where $\sigma_2^2 = 1.40^2$. Use a *normal* $(2, 2^2)$ prior for μ_2. Calculate the posterior distribution of $\mu_1|y_{14} \ldots, y_{19}$.

(d) Calculate the posterior distribution of $\mu_d = \mu_1 - \mu_2$, the difference between the means.

(e) Calculate a 95% Bayesian credible interval for μ_d.

(f) Test the hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

at the 5% level of significance. What conclusion can be drawn.

12.7 Glass fragments found on a suspect's shoes or clothes are often used to connect the suspect to a crime scene. The index of refraction is the fragments are compared to the refractive index of the glass from the crime scene. To make this comparison rigorous, we need to know the variability the index of refraction is over a pane of glass. Bennet et al. (2002) analyzed the refractive index in a pane of float glass, searching for any spatial pattern. Here are samples of the refractive index from the edge and from the middle of the pane.

Edge of Pane		Middle of Pane	
1.51996	1.51997	1.52001	1.51999
1.51998	1.52000	1.52004	1.51997
1.51998	1.52004	1.52005	1.52000
1.52000	1.52001	1.52004	1.52002
1.52000	1.51997	1.52004	1.51996

For these data, $\bar{y}_1 = 1.51999$, $\bar{y}_2 = 1.52001$, $\sigma_1 = .00002257$, and $\sigma_2 = .00003075$.

(a) Suppose glass at the edge of the pane is *normal* (μ_1, σ_1^2), where $\sigma_1 = .00003$. Calculate the posterior distribution of μ_1 when you use a *normal* $(1.52000, .0001^2)$ prior for μ_1.

(b) Suppose glass in the middle of the pane is *normal* (μ_2, σ_2^2), where $\sigma_2 = .00003$. Calculate the posterior distribution of μ_2 when you use a *normal* $(1.52000, .0001^2)$ prior for μ_2.

(c) Find the posterior distribution of $\mu_d = \mu_1 - \mu_2$.

(d) Find a 95% credible interval for μ_d.

(e) Perform a Bayesian test of the hypothesis

$$H_0 : \mu_d = 0 \quad \text{versus} \quad \mu_d \neq 0$$

at the 5% level of significance.

The last half of the twentieth century saw great change in the role of women in New Zealand society. These changes included education, employment, family formation, and fertility, where women took control of these aspects of their lives. During those years phrases such as "women's liberation movement" and "the sexual revolution" were used to describe the changing role of women in society. In 1995 the Population Studies Centre at the University of Waikato sponsored the New Zealand Women Family, Employment, and Education Survey (NZFEE) to investigate these changes. A random sample of New Zealand women of all ages between 20 and 59 was taken, and the women were interviewed about their educational, employment, and personal history. The details of this survey are summarized in Marsault et al. (1997). Detailed analysis of the data from this survey is in Johnstone et al. (2001).

12.8 Have the educational qualifications of younger New Zealand women changed from those of previous generations of New Zealand women? To shed light on this question, we will compare the educational qualifications of two generations of New Zealand women 25 years apart. The women in the age group 25-29 at the time of the survey were born between 1966 and 1970. The women in the age group 50-54 at the time of the survey were born between 1941 and 1945.

(a) Out of 314 women in the age group 25-29, 234 had completed a secondary school qualification. Find the posterior distribution of π_1, the proportion of New Zealand women in that age group who have a completed a secondary school qualification. (Use a *uniform* prior for π_1.)

(b) Out of 219 women in the age group 50-54, 120 had completed a secondary school qualification. Find the posterior distribution of π_2, the proportion of New Zealand women in that age group who have a completed a secondary school qualification. (Use a *uniform* prior for π_2.)

(c) Find the approximate posterior distribution of $\pi_1 - \pi_2$.

(d) Find a 99% Bayesian credible interval for $\pi_1 - \pi_2$.

(e) What would be the conclusion if you tested the hypothesis

$$H_0 : \pi_1 - \pi_2 = 0 \quad \text{versus} \quad H_1 : \pi_1 - \pi_2 \neq 0$$

at the 1% level of significance?

12.9 Are younger New Zealand women more likely to be in paid employment than previous generations of New Zealand women? To shed light on this question, we will look at the current employment status of two generations of New Zealand women 25 years apart.

(a) Out of 314 women in the age group 25-29, 171 were currently in paid employment. Find the posterior distribution of π_1, the proportion of New Zealand women in that age group who are currently in paid employment. (Use a *uniform* prior for π_1.)

(b) Out of 219 women in the age group 50-54, 137 were currently in paid employment. Find the posterior distribution of π_2, the proportion of New Zealand women in that age group who are currently in paid employment. (Use a *uniform* prior for π_2.)

(c) Find the approximate posterior distribution of $\pi_1 - \pi_2$.

(d) Find a 99% Bayesian credible interval for $\pi_1 - \pi_2$.

(e) What would be the conclusion if you tested the hypothesis

$$H_0 : \pi_1 - \pi_2 = 0 \quad \text{versus} \quad H_1 : \pi_1 - \pi_2 \neq 0$$

at the 1% level of significance?

12.10 Are younger New Zealand women becoming sexually active at an earlier age than previous generations of New Zealand women? To shed light on this question, we look at the proportions of Zealand women who report having experienced sexual intercourse before age 18 for the two generations of New Zealand women.

(a) Out of the 298 women in the age group 25-29 who responded to this question, 180 report having experienced sexual intercourse before reaching the age of 18. Find the posterior distribution of π_1, the proportion

of New Zealand women in that age group who had experienced sexual intercourse before age 18. (Use a *uniform* prior for π_1.)

(b) Out of the 218 women in the age group 50-54 who responded to this question, 52 report having experienced sexual intercourse before reaching the age of 18. Find the posterior distribution of π_2, the proportion of New Zealand women in that age group who had experienced sexual intercourse before age 18. (Use a *uniform* prior for π_2.)

(c) Find the approximate posterior distribution of $\pi_1 - \pi_2$.

(d) Test the hypothesis

$$H : 0 : \pi_1 - \pi_2 \leq 0 \quad \text{versus} \quad H_1 : \pi_1 - \pi_2 > 0$$

in a Bayesian manner at the 1% level of significance. Can we conclude that New Zealand women in the generation aged 25-29 have experienced sexual intercourse at an earlier age than New Zealand women in the generation aged 50-54?

12.11 Are younger New Zealand women marrying at a later age than previous generations of New Zealand women? To shed light on this question, we look at the proportions of Zealand women who report having been married before age 22 for the two generations of New Zealand women.

(a) Out of the 314 women in the age group 25-29, 69 report having been married before the age 22. Find the posterior distribution of π_1, the proportion of New Zealand women in that age group who have married before age 22. (Use a *uniform* prior for π_1.)

(b) Out of the 219 women in the age group 50-54, 114 report having been married before age 22. Find the posterior distribution of π_2, the proportion of New Zealand women in that age group who have been married before age 22. (Use a *uniform* prior for π_2.)

(c) Find the approximate posterior distribution of $\pi_1 - \pi_2$.

(d) Test the hypothesis

$$H : 0 : \pi_1 - \pi_2 \geq 0 \quad \text{versus} \quad H_1 : \pi_1 - \pi_2 < 0$$

in a Bayesian manner at the 1% level of significance. Can we conclude that New Zealand women in the generation aged 25-29 have married at an earlier age than New Zealand women in the generation aged 50-54?

12.12 Family formation patterns in New Zealand have changed over the time frame covered by the survey. New Zealand society has become more accepting of couples co-habiting (living together before or instead of legally marrying). When we take this into account, are younger New Zealand women forming family like units at a similar age to previous generations?

(a) Out of the 314 women in the age group 25-29, 199 report having formed a domestic partnership (either co-habiting or legal marriage) before age 22. Find the posterior distribution of π_1, the proportion of New Zealand women in that age group who have formed a domestic partnership before age 22. (Use a *uniform* prior for π_1.)

(b) Out of the 219 women in the age group 50-54, 116 report having formed a domestic partnership before age 22. Find the posterior distribution of π_2, the proportion of New Zealand women in that age group who have formed a domestic partnership before age 22. (Use a *uniform* prior for π_2.)

(c) Find the approximate posterior distribution of $\pi_1 - \pi_2$.

(d) Find a 99% Bayesian credible interval for $\pi_1 - \pi_2$.

(e) What would be the conclusion if you tested the hypothesis

$$H_0 : \pi_1 - \pi_2 = 0 \quad \text{versus} \quad H_1 : \pi_1 - \pi_2 \neq 0$$

at the 1% level of significance.

12.13 Are young New Zealand women having their children at a later age than previous generations?

(a) Out of the 314 women in the age group 25-29, 136 report having given birth to their first child before the age of 25. Find the posterior distribution of π_1, the proportion of New Zealand women in that age group who have given birth before age 25. (Use a *uniform* prior for π_1.)

(b) Out of the 219 women in the age group 50-54, 135 report having given birth to their first child before age 25. Find the posterior distribution of π_2, the proportion of New Zealand women in that age group who have given birth before age 25. (Use a *uniform* prior for π_2.)

(c) Find the approximate posterior distribution of $\pi_1 - \pi_2$.

(d) Test the hypothesis

$$H : 0 : \pi_1 - \pi_2 \geq 0 \quad \text{versus} \quad H_1 : \pi_1 - \pi_2 < 0$$

in a Bayesian manner at the 1% level of significance. Can we conclude that New Zealand women in the generation aged 25-29 have had their first child at a later age than New Zealand women in the generation aged 50-54?

12.14 The experiment described in Exercise 6 was repeated on another set of 7 cows, McLeay, Carruthers, and Neil (1997). However, in this case, the second treatment was given to the same set of 7 cows that were given the first treatment, at a later time when the first dose of ^{13}C had been eliminated from the cow. The data are given below:

Cow ID	^{13}C *Administered into Reticulum* x	^{13}C *Administered into Reticulo-omasal Orifice* y
1	1.1	3.5
2	0.8	3.6
3	1.7	5.1
4	1.1	5.6
5	2.0	6.2
6	1.6	6.5
7	3.1	8.3

(a) Explain why the variables x and y cannot be considered independent in this experiment.

(b) Calculate the differences $d_i = x_i - y_i$ for $i = 1, \ldots, 7$.

(c) Assume that the differences come from a *normal* (μ_d, σ_d^2) distribution, where $\sigma_d^2 = 1$. Use a *normal* $(0, 3^2)$ prior for μ_d. Calculate the posterior for $\mu_d | d_1, \ldots, d_7$.

(d) Calculate a 95% Bayesian credible interval for μ_d.

(e) Test the hypothesis

$$H_0 : \mu_d = 0 \quad \text{versus} \quad H_1 : \mu_d \neq 0$$

at the 5% level of significance. What conclusion can be drawn?

12.15 One of the advantages of Bayesian statistics is that evidence from different sources can be combined. In Exercise 6 and Exercise 14, we found posterior distributions of μ_d using data sets from two different experiments. In the first experiment, the two treatments were given to two sets of cows, and the measurements were independent. In the second experiment, the two treatments were given to a third set of cows at different times and the measurements were paired. When we want to find the posterior distribution given data sets from two independent experiments, we should use the posterior distribution after the first experiment as the prior distribution for the second.

(a) Explain why the two data sets can be considered independent.

(b) Find the posterior distribution of $\mu_d | data$ where the *data* include all of the measurements $x_8 \ldots, x_{13}, y_{14} \ldots, y_{19}, d_1, \ldots, d_7$.

(c) Find a 95 % credible interval for μ_d based on all the data.

(d) Test the hypothesis

$$H_0 : \mu_d = 0 \quad \text{versus} \quad H_1 : \mu_d \neq 0$$

at the 5 % level of significance. Can we conclude that ^{13}C octanic acid breath test is effective in detecting reticular groove contraction in cattle?

13

Bayesian Inference for Simple Linear Regression

Sometimes we want to model a relationship between two variables, x and y. We might want to find an equation that describes the relationship. Often we plan to use the value of x to help predict y using that relationship.

The data consist of n ordered pairs of points (x_i, y_i) for $i = 1, \ldots, n$. We think of x as the predictor variable (independent variable) and consider that we know it without error. We think y is a response variable that depends on x in some unknown way, but that each observed y contains an error term as well. We plot the points on a two-dimensional *scatterplot*; the predictor variable is measured along the horizontal axis, and the response variable is measured along the vertical axis.

We examine the scatterplot for clues about the nature of the relationship. To construct a regression model, we first decide on the type of equation that appears to fit the data. A *linear* relationship is the simplest equation relating two variables. This would give a straight line relationship between the predictor x and the response y. We leave the parameters of the line, the slope β, and the y-intercept α_0 unknown, so all lines are possible.

Then we determine the best estimates of the unknown parameters by some criterion. The criterion that is most frequently used is *least squares*. This is where we find the parameter values that minimize the sum of squares of the *residuals*, which are the vertical distances of the observed points to the fitted equation. We do this for the simple linear regression in Section 13.1. In Section 13.2 we look at how an exponential growth model can be fitted using least squares regression on the logarithm of the response variable.

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

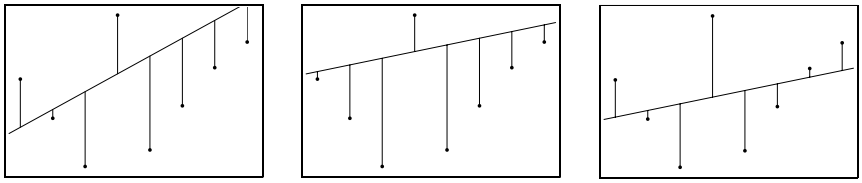


Figure 13.1 Scatterplot with three possible lines, and the residuals from each of the lines. The third line is the least squares line. It minimizes the sum of squares of the residuals.

At this stage no inferences are possible because there is no probability model for the data. In Section 13.3 we construct a regression model that makes assumptions on how the response variable depends on the predictor variable, and how randomness enters the data. Inferences can be done on the parameters of this model. The most important one is determining the predictive distribution of new observations, given the data. In Section 13.4 we fit a *linear* relationship between the two variables using Bayesian methods, and perform Bayesian inferences on the parameters of the model.

13.1 LEAST SQUARES REGRESSION

We could draw any number of lines on the scatterplot. Some of them would fit fairly well, others would be extremely far from the points. A *residual* is the *vertical* distance from an observed point on the scatterplot to the line. We can put in any line that we like, and calculate the residuals from that line. Least squares is a method for finding the line that *best* fits the points in terms of minimizing *sum of squares of the residuals*. Figure13.1 shows a scatterplot, three possible lines, and the residuals from each line.

The equation of a line is determined by two things: its slope β and its y-intercept α_0. Actually its slope and any other point on the line will do, for instance, $\alpha_{\bar{x}}$, the intercept of the vertical line at $\bar{x}$. Finding the least squares line is equivalent to finding its slope and the y-intercept (or another intercept).

The Normal Equations

The sum of squares of the residuals from line $y = \alpha_0 + \beta x$ is

$$SS_{res} = \sum_{i=1}^{n} [y_i - (\alpha_0 + \beta x_i)]^2 .$$

To find values of α_0 and β that minimize SS_{res} using calculus, take derivatives with respect to each α_0 and β and set equal to 0, and solve the resulting set of simultaneous equations. First, take the derivative with respect to intercept α_0. This

gives the equation

$$\frac{\partial SS}{\partial \alpha_0} = \sum_{i=1}^{n} 2 \times [y_i - (\alpha_0 + \beta x_i)]^1 \times (-1) = 0$$

which simplifies to

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \alpha_0 - \sum_{i=1}^{n} \beta x_i = 0$$

and further to

$$\bar{y} - \alpha_0 - \beta \bar{x} = 0\,. \tag{13.1}$$

Second, taking the derivative with respect to the slope β gives the equation

$$\frac{\partial SS}{\partial \beta} = \sum_{i=1}^{n} 2 \times [y_i - (\alpha_0 + \beta x_i)]^1 \times (-x_i) = 0\,,$$

which simplifies to

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \alpha_0 x_i - \sum_{i=1}^{n} \beta x_i^2 = 0$$

and further to

$$\overline{xy} - \alpha_0 \bar{x} - \beta \overline{x^2} = 0\,. \tag{13.2}$$

Equation 13.1 and Equation 13.2 are known as the *normal* equations. Here *normal* refers to right angles[1] and has nothing to do with the normal distribution.

Solve Equation 13.1 for α_0 in terms of β and substitute into Equation 13.2 and solve for β

$$\overline{xy} - (\bar{y} - \beta \bar{x})\bar{x} - \beta \overline{x^2} = 0\,.$$

The solution is the least squares slope[2]

$$B = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}\,. \tag{13.3}$$

Substitute this back into Equation 13.1 and solve for the least squares y-intercept,

$$A_0 = \bar{y} - B\bar{x}\,. \tag{13.4}$$

The equation of the least squares line is

$$y = A_0 + Bx\,. \tag{13.5}$$

[1] Least squares finds the projection of the (n-dimensional) observation vector onto the plane containing all possible values of (α_0, β).

[2] There are many different formulas for the least squares slope. This can be a source of confusion as many books give formulas that look quite dissimilar. However, all can be shown to be equivalent. I use this one because it is easy to remember. The average of $x \times y$ minus the average of $x \times$ the average of y all divided by the average of x^2 minus the square of the average of x.

The slope and any other point besides y-intercept also determines the line. Say the point is $A_{\bar{x}}$, where the least squares line intercepts the vertical line at $\bar{x}$:

$$A_{\bar{x}} = A_0 + B\bar{x} = \bar{y}\,.$$

Thus the least squares line goes through the point $(\bar{x}, \bar{y})$. An alternative equation for the least squares line is

$$y = A_{\bar{x}} + B(x - \bar{x}) = \bar{y} + B(x - \bar{x})\,, \tag{13.6}$$

which is particularly useful.

Estimating the Variance around the Least Squares Line

The estimate of the variance around the least squares line is

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}[y_i - (A_{\bar{x}} + B(x_i - \bar{x}))]^2}{n-2}$$

which is the sum of squares of the residuals divided by $n-2$. The reason we use $n-2$ is that we have used two estimates, $A_{\bar{x}}$ and B in calculating the sum of squares[3].

Example 23 *A company is manufacturing a food product, and must control the moisture level in the final product. It is cheaper (and hence preferable) to measure the level at an in-process stage rather than in the final product. Michael, the company statistician, recommends to the engineers running the process that a measurement of the moisture level at an in-process stage may give a good prediction of what the final moisture level will be. He organizes the collection of data from 25 batches, giving the moisture level at the in-process stage and the final moisture level for each batch. These are shown in the first three columns of Table 13.1.*

Summary statistics for these data are: $\bar{x} = 14.389$, $\bar{y} = 14.221$, $\overline{x^2} = 207.0703$, $\overline{y^2} = 202.3186$, *and* $\overline{xy} = 204.6628$. *Note that he needs to keep all the significant figures in the squared terms. The formula for* B *uses subtraction, and if he rounds off too early, the differences will have too few significant figures and accuracy will be lost.*

He then calculates the least squares line relating the final moisture level to the in-process moisture level:

$$B = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{204.6628 - 14.389 \times 14.221}{207.0703 - (14.389)^2} = \frac{.042569}{.032755} = 1.30\,.$$

The equation of the least squares line is

$$y = 14.221 + 1.29963 \times (x - 14.389).$$

[3]The general rule for finding an unbiased estimate of the variance is that the sum of squares is divided by the degrees of freedom, and we lose a degree of freedom for every estimated parameter in the sum of squares formula.

Table 13.1 In-process and final moisture levels

Batch	In-Process Level x	Final Level y	LS Fits $\hat{y} = A_0 + Bx_i$	Residual $y - \hat{y}$	Residual^2 $(y - \hat{y})^2$
1	14.36	13.84	14.1833	-0.343256	0.117825
2	14.48	14.41	14.3392	0.070792	0.005012
3	14.53	14.22	14.4042	-0.184188	0.033925
4	14.52	14.63	14.3912	0.238808	0.057029
5	14.35	13.95	14.1703	-0.220260	0.048514
6	14.31	14.37	14.1183	0.251724	0.063365
7	14.44	14.41	14.2872	0.122776	0.015074
8	14.23	13.99	14.0143	-0.024308	0.000591
9	14.32	13.89	14.1313	-0.241272	0.058212
10	14.57	14.59	14.4562	0.133828	0.017910
11	14.28	14.32	14.0793	0.240712	0.057942
12	14.36	14.31	14.1833	0.126744	0.016064
13	14.50	14.43	14.3652	0.064800	0.004199
14	14.52	14.44	14.3912	0.048808	0.002382
15	14.28	14.14	14.0793	0.060712	0.003686
16	14.13	13.90	13.8843	0.015652	0.000245
17	14.54	14.37	14.4172	-0.047184	0.002226
18	14.60	14.34	14.4952	-0.155160	0.024075
19	14.86	14.78	14.8331	-0.053056	0.002815
20	14.28	13.76	14.0793	-0.319288	0.101945
21	14.09	13.85	13.8324	0.017636	0.000311
22	14.20	13.89	13.9753	-0.085320	0.007280
23	14.50	14.22	14.3652	-0.145200	0.021083
24	14.02	13.80	13.7414	0.058608	0.003435
25	14.45	14.67	14.3002	0.369780	0.136737
Mean	14.389	14.221			

The scatterplot of final moisture level and in-process moisture level together with the least squares line is given in Figure 13.2.

He calculates the least squares fitted values $\bar{y} + B(x_i - \bar{x})$, the residuals, and the squared residuals. They are in the last three columns of Table 13.1. The estimated

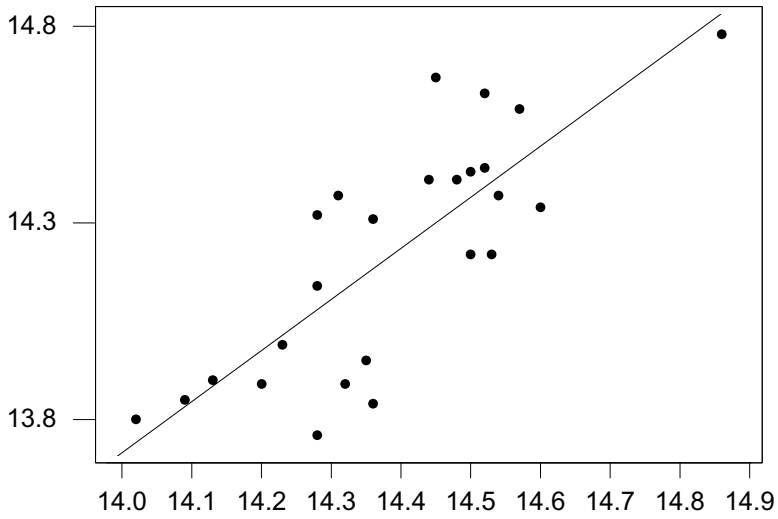


Figure 13.2 Scatterplot and least squares line for the moisture data.

variance about the least squares line is

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(y-\hat{y})^2}{n-2} = \frac{0.80188}{23} = 0.0320753\,.$$

To find the estimated standard deviation about the least squares line, he takes the square root:

$$\hat{\sigma} = \sqrt{(0.0320753)} = 0.179096\,.$$

13.2 EXPONENTIAL GROWTH MODEL

When we look at economic time series, the predictor variable is time t, and we want to see how some response variable u depends on t. Often, when we graph the response variable versus time on a scatterplot, we notice two things. First, the plotted points seem to go up not at a linear rate but at a rate that increases with time. Second, the variability of the plotted points seems to be increasing at about the same rate as the response variable. This will be shown more clearly if we graph the residuals versus time. In this case the exponential growth model will usually give a better fit:

$$u = e^{\alpha_0+\beta\times t}\,.$$

We note that if we let $y = \log_e(u)$, then

$$y = \alpha_0 + \beta \times t$$

Table 13.2 Annual poultry production in New Zealand

Year	Poultry Production	Linear			Exponential
t	u	Fitted Value	$\log_e(u)$	Fitted $\log_e u$	Fitted Value
1987	44,085	47,757	10.7739	10.7776	47,934
1988	51,646	48,725	10.8522	10.8393	50,986
1989	57,241	53,364	10.9550	10.9010	54,232
1990	56,261	58,004	10.9378	10.9628	57,686
1991	58,257	62,643	10.9726	11.0245	61,359
1992	60,944	67,283	11.0177	11.0862	65,266
1993	68,214	71,922	11.1304	11.1479	69,421
1994	74,037	76,562	11.2123	11.2097	73,842
1995	88,646	81,201	11.3924	11.2714	78,543
1996	86,869	85,841	11.3722	11.3331	83,545
1997	86,534	90,480	11.3683	11.3949	88,864
1998	95,682	95,120	11.4688	11.4566	94,522
1999	97,400	99,759	11.4866	11.5183	100,541
2000	10,4927	104,398	11.5610	11.5801	106,943
2001	11,4010	109,038	11.6440	11.6418	113,752

is a linear relationship. We can estimate the parameters of the relationship using least squares using response variable y. The fitted exponential growth model is

$$u = e^{A_0 + B \times t},$$

where B and A_0 are the least squares slope and intercept for the logged data.

Example 24 *The annual New Zealand poultry production (in tonnes) for the years 1987-2001 is given in Table 13.2.*

The scatterplot showing the residuals and least squares line is shown in Figure 13.3. We see that the residuals are mostly positive at the ends of the data, and mostly negative in the center. This indicates that an exponential growth model would give a better fit. The scatterplot, and the exponential growth model found by exponentiating the least squares line to the logged data are shown in Figure 13.4.

13.3 SIMPLE LINEAR REGRESSION ASSUMPTIONS

The method of least squares is *nonparametric* or *distribution free*, since it makes no use of the probability distribution of the data. It is really a data analysis tool, and can

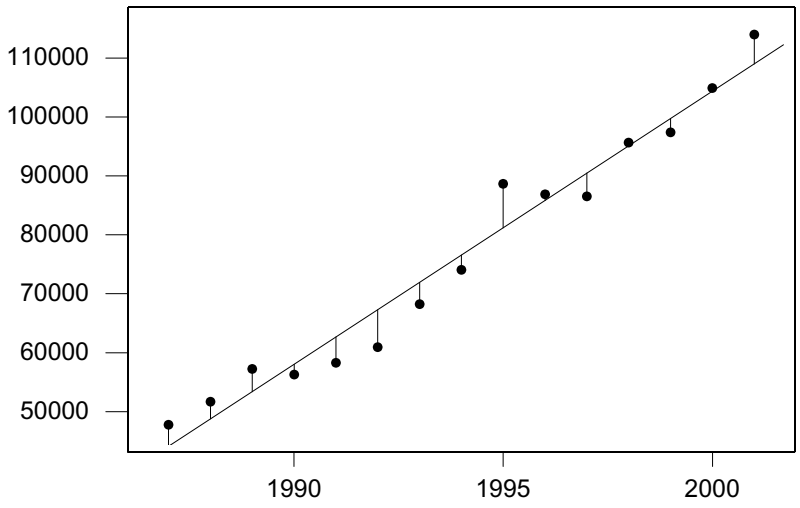


Figure 13.3 Scatterplot and least squares line for the poultry production data.

be applied to any bivariate data. We can't make any inferences about the slope and intercept nor about any predictions from the least squares model, unless we make some assumptions about the probability model underlying the data. The simple linear regression assumptions are:

1. *Mean assumption*. The conditional mean of y given x is an unknown linear function of x.

   $$\mu_{y|x} = \alpha_0 + \beta x\,,$$

   where β is the unknown slope and α_0 is the unknown *y intercept*, the intercept of the vertical line $x = 0$. In the alternate parameterization

   $$\mu_{y|x} = \alpha_{\bar{x}} + \beta(x - \bar{x})\,,$$

   where $\alpha_{\bar{x}}$ is the unknown intercept of the vertical line $x = \bar{x}$. In this parameterization the least squares estimates $A_{\bar{x}} = \bar{y}$ and B will be independent under our assumptions, so the likelihood will factor into a part depending on $\alpha_{\bar{x}}$ and a part depending on β. This greatly simplifies things, so we will use this parameterization. The mean assumption is shown in the first graph of Figure 13.5.

2. *Error assumption*. Observation equals mean plus error, which is *normally* distributed with mean 0 and *known* variance σ^2. All errors have equal variance. The equal variance assumption is shown in the second graph of Figure 13.5.

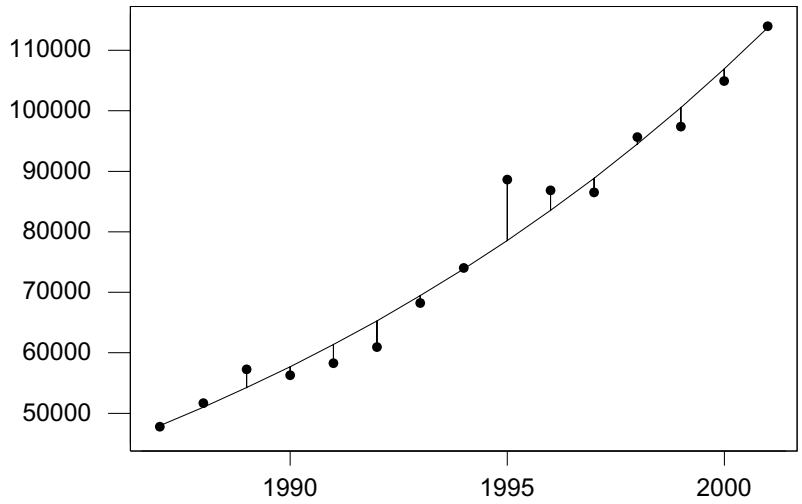


Figure 13.4 Scatterplot and fitted exponential growth model for the poultry production data.

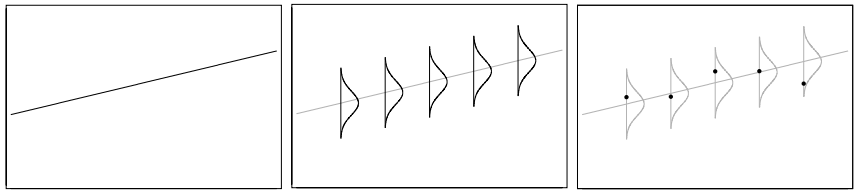


Figure 13.5 Assumptions of linear regression model. The mean of Y given X is a linear function. The observation errors are normally distributed with mean 0 and equal variances. The observations are independent of each other.

3. *Independence assumption*. The errors for all of the observations are independent of each other. The independent draw assumption is shown in the third graph of Figure 13.5.

Using the alternate parameterization

$$y_i = \alpha_{\bar{x}} + \beta \times (x_i - \bar{x}) + e_i \,,$$

where $\alpha_{\bar{x}}$ is the mean value for y given $x = \bar{x}$, and β is the slope. Each e_i is normally distributed with mean 0 and known variance σ^2. The e_i are all independent of each other. Therefore $y_i|x_i$ is normally distributed with mean $a_{\bar{x}} + \beta(x_i - \bar{x})$ and variance σ^2 and all the $y_i|x_i$ are all independent of each other.

13.4 BAYES' THEOREM FOR THE REGRESSION MODEL

Bayes' theorem is always summarized by

$$posterior \propto prior \times likelihood\,,$$

so we need to determine the likelihood and decide on our prior for this model.

The Joint Likelihood for β and $\alpha_{\bar{x}}$

The joint likelihood of the i^{th} observation is its probability density function as a function of the two parameters $\alpha_{\bar{x}}$ and β, where (x_i, y_i) are fixed at the observed values. It gives relative weights to all possible values of both parameters $a_{\bar{x}}$ and β from the observation. The likelihood of observation i is

$$likelihood_i(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}[y_i-(\alpha_{\bar{x}}+\beta(x_i-\bar{x}))]^2}\,,$$

since we can ignore the part not containing the parameters. The observations are all independent, so the likelihood of the whole sample of all the observations is the product of the individual likelihoods:

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}[y_i-(\alpha_{\bar{x}}+\beta(x_i-\bar{x}))]^2}\,.$$

The product of exponentials is found by summing the exponents, so

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}[y_i-(a_{\bar{x}}+\beta(x_i-\bar{x}))]^2\right]}\,.$$

The term in brackets in the exponent equals

$$\left[\sum_{i=1}^{n}[y_i - \bar{y} + \bar{y} - (\alpha_{\bar{x}} + \beta(x_i - \bar{x}))]^2\right]\,.$$

Breaking this into three sums and multiplying it out gives us

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \bar{y})(\bar{y} - (\alpha_{\bar{x}} + \beta(x_i - \bar{x})))$$

$$+\sum_{i=1}^{n}(\bar{y} - (\alpha_{\bar{x}} + \beta(x_i - \bar{x})))^2\,.$$

This simplifies into

$$SS_y - 2\beta SS_{xy} + \beta^2 SS_x + n(\alpha_{\bar{x}} - \bar{y})^2\,,$$

where $SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$, and $SS_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}))$, and $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Thus the joint likelihood can be written as

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}[SS_y-2\beta SS_{xy}+\beta^2 SS_x+n(\alpha_{\bar{x}}-\bar{y})^2]}\,.$$

Writing this as a product of two exponentials gives

$$\propto e^{-\frac{1}{2\sigma^2}[SS_y - 2\beta SS_{xy} + \beta^2 SS_x]} \times e^{-\frac{1}{2\sigma^2}[n(\alpha_{\bar{x}} - \bar{y})^2]} .$$

We factor out SS_x in the first exponential, complete the square, and Absorb the part that doesn't depend on any parameter into the proportionality constant. This gives us

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2/SS_x}[\beta - \frac{SS_{xy}}{SS_x}]^2} \times e^{-\frac{1}{2\sigma^2/n}[(\alpha_{\bar{x}} - \bar{y})^2]} .$$

Note that $\frac{SS_{xy}}{SS_x} = B$, the least squares slope, and $\bar{y} = A_{\bar{x}}$, the least squares estimate of the intercept of the vertical line $x = \bar{x}$. We have factored the joint likelihood into the product of two individual likelihoods

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto likelihood_{sample}(\alpha_{\bar{x}}) \times likelihood_{sample}(\beta) ,$$

where

$$likelihood_{sample}(\beta) \propto e^{-\frac{1}{2\sigma^2/SSx}(\beta - B)^2}$$

and

$$likelihood_{sample}(\alpha_{\bar{x}}) \propto e^{-\frac{1}{\sigma^2/n}(\alpha_{\bar{x}} - A_{\bar{x}})^2} .$$

Since the joint likelihood has been factored into the product of the individual likelihoods we know the individual likelihoods are independent. We recognize that the likelihood of the slope β has the *normal* shape with mean B, the least squares slope, and variance $\frac{\sigma^2}{SS_x}$. Similarly the likelihood of $\alpha_{\bar{x}}$ has the *normal* shape with mean $A_{\bar{x}}$ and variance $\frac{\sigma^2}{n}$.

The Joint Prior for β and $\alpha_{\bar{x}}$

If we multiply the joint likelihood by a joint prior, it is proportional to the joint posterior. We will use independent priors for each parameter. The joint prior of the two parameters is the product of the two individual priors:

$$g(\alpha_{\bar{x}}, \beta) = g(\alpha_{\bar{x}}) \times g(\beta) .$$

We can either use *normal* priors, or *flat* priors.

The Joint Posterior for β and $\alpha_{\bar{x}}$

The joint posterior then is proportional to the joint prior times the joint likelihood.

$$g(\alpha_{\bar{x}}, \beta | data) \propto g(\alpha_{\bar{x}}, \beta) \times likelihood_{sample}(\alpha_{\bar{x}}, \beta) ,$$

where the $data$ is the set of ordered pair $(x_1, y_1), \ldots, (x_n, y_n)$. The joint prior and the joint likelihood both factor into a part depending on $\alpha_{\bar{x}}$ and a part depending on β. Rearranging them gives the joint posterior factored into the marginal posteriors

$$g(\alpha_{\bar{x}}, \beta | data) \propto g(\alpha_{\bar{x}} | data) \times g(\beta | data) .$$

Since the joint posterior is the product of the marginal posteriors, they are independent. Each of these marginal posteriors can be found by using the simple updating rules for normal distributions, which works for *normal* and *flat* priors. For instance, if we use a *normal*(m_β, s_β^2) prior for β, we get a *normal*$(m'_\beta, (s'_\beta)^2)$, where

$$\frac{1}{(s'_\beta)^2} = \frac{1}{s_\beta^2} + \frac{SS_x}{\sigma^2} \tag{13.7}$$

and

$$m'_\beta = \frac{\frac{1}{s_\beta^2}}{\frac{1}{(s'_\beta)^2}} \times m_\beta + \frac{\frac{SS_x}{\sigma^2}}{\frac{1}{(s'_\beta)^2}} \times B\,. \tag{13.8}$$

The posterior precision equals the prior precision plus the precision of the likelihood. The posterior mean equals the weighted average of the prior mean and the likelihood mean where the weights are the proportions of the precisions to the posterior precision. And the posterior distribution is normal.

Similarly if we use a *normal*$(m_{\alpha_{\bar{x}}}, s^2_{\alpha_{\bar{x}}})$ prior for $\alpha_{\bar{x}}$, we get a *normal*$(m'_{\alpha_{\bar{x}}}, (s'_{\alpha_{\bar{x}}})^2)$ where

$$\frac{1}{(s'_{\alpha_{\bar{x}}})^2} = \frac{1}{s^2_{\alpha_{\bar{x}}}} + \frac{n}{\sigma^2}\,,$$

and

$$m'_{\alpha_{\bar{x}}} = \frac{\frac{1}{s^2_{\alpha_{\bar{x}}}}}{\frac{1}{(s'_{\alpha_{\bar{x}}})^2}} \times m_{\alpha_{\bar{x}}} + \frac{\frac{n}{\sigma^2}}{\frac{1}{(s'_{\alpha_{\bar{x}}})^2}} \times A_{\bar{x}}\,.$$

Example 23 (continued) *The statistician decides that he will use a normal* $(1, (.3)^2)$ *prior for* β *and a normal* $(15, 1^2)$ *prior for* $\alpha_{\bar{x}}$. *Since he doesn't know the true variance, he will use the estimated variance about the least squares regression line* $\hat{\sigma^2} = 0.0320753$. *Note that* $SS_x = \sum_{i=1}^n (x_i - \bar{x})^2 = n(\bar{x^2} - \bar{x}^2) = 25 * (207.0703 - 14.389^2) = .674475$.

The posterior precision of β *is*

$$\frac{1}{(s'_\beta)^2} = \frac{1}{.3^2} + \frac{25}{.674475} = 48.177\,,$$

so the posterior standard deviation of β *is*

$$s'_\beta = 48.177^{-\frac{1}{2}} = .144\,.$$

The posterior mean of β *is*

$$m'_\beta = \frac{\frac{1}{.3^2}}{48.177} \times 1 + \frac{\frac{25}{.64775}}{48.177} \times 1.29963 = 1.231\,.$$

Similarly the posterior precision of $\alpha_{\bar{x}}$ *is*

$$\frac{1}{(s'_{\alpha_{\bar{x}}})^2} = \frac{1}{1^2} + \frac{25}{.674475} = 38.066\,,$$

so the posterior standard deviation is

$$s'_{\alpha_{\bar{x}}} = 38.066^{-\frac{1}{2}} = .162\,.$$

The posterior mean of $\alpha_{\bar{x}}$ *is*

$$m'_{\alpha_{\bar{x}}} = \frac{\frac{1}{1^2}}{38.066} \times 15 + \frac{\frac{25}{.64775}}{38.066} \times 14.221 = 14.242\,.$$

Bayesian Credible Interval for Slope

The posterior distribution of β summarizes our entire belief about it after examining the data. We may want to summarize it by a $(1-\alpha)\times 100\%$ Bayesian credible interval for slope β. This will be

$$m'_\beta \pm z_{\frac{\alpha}{2}} \times \sqrt{(s'_\beta)^2}\,. \tag{13.9}$$

More realistically, we don't know σ^2. A sensible approach in that instance is to use the estimate calculated from the residuals

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - (A_{\bar{x}} + B(x_i - \bar{x})))^2}{n-2}\,.$$

We have to widen the confidence interval to account for the increased uncertainty due to not knowing σ^2. We do this by using a *Student's t* critical value with $n-2$ degrees of freedom[4]. The credible interval becomes

$$m'_\beta \pm t_{\frac{\alpha}{2}} \times \sqrt{(s'_\beta)^2}\,. \tag{13.10}$$

Frequentist Confidence Interval for Slope

When the variance σ^2 is unknown, the $(1-\alpha)\times 100\%$ confidence interval for the slope β is

$$B \pm t_{\frac{\alpha}{2}} \times \frac{\hat{\sigma}}{\sqrt{SS_x}}\,,$$

where $\hat{\sigma}^2$ is the estimate of the variance calculated from the residuals from the least squares line. The confidence interval is the same form as the Bayesian credible interval when we used *flat* priors for β and $\alpha_{\bar{x}}$. Of course the interpretation is different. Under the frequentist assumptions we are $(1-\alpha)\times 100\%$ confident that the interval contains the true, unknown parameter value. Once again, the frequentist confidence interval is equivalent to a Bayesian credible interval, so if the scientist misinterprets it as a probability interval, he/she will get away with it. The only loss experienced will be that the scientist did not get to put in any of his/her prior knowledge.

[4]Actually we are treating the unknown parameter σ^2 as a nuisance parameter and using the prior $g(\sigma^2) \propto (\sigma^2)^{-1}$. The marginal posterior of β is found by integrating σ^2 out of the joint posterior.

Testing One-Sided Hypothesis about Slope

Often we want to determine whether or not the amount of increase in y associated with one unit increase in x is greater than some value, β_0. We can do this by testing

$$H_0 : \beta \leq \beta_0 \quad \text{versus.} \quad H_1 : \beta > \beta_0$$

at the α level of significance in a Bayesian manner. To do the test in a Bayesian manner, we calculate the posterior probability of the null hypothesis. This is

$$\begin{aligned} P(\beta \leq \ \beta_0 | data) &= \int_{-\infty}^{\beta_0} g(\beta|data)d\beta \\ &= P\left(Z \leq \frac{\beta_0 - m'_\beta}{\sqrt{\frac{(s'_\beta)^2}{SS_x}}}\right). \end{aligned} \tag{13.11}$$

If this probability is greater than α, then we reject H_0 and conclude that indeed, the slope β is greater than β_0. (If we used the estimate of the variance, then we would use a *Student's t* with $n-2$ degrees of freedom instead of the standard normal Z.)

Example 23 (continued) *Since he used the estimated variance in place of the unknown true variance, he used Equation 13.10 to find the Bayesian credible interval where there are 23 degrees of freedom. The interval is (.933,1.529).*

Testing Two-Sided Hypothesis about Slope

If $\beta = 0$, then the mean of y does not depend on x at all. We really would like to test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ at the α level of significance in a Bayesian manner, before we use the regression model to make predictions.

To do the test in a Bayesian manner, look where 0 lies in relation to the credible interval. If it lies outside interval, reject H_0. Otherwise, we can't reject the null hypothesis, and we should not use the regression model to help with predictions.

13.5 PREDICTIVE DISTRIBUTION FOR FUTURE OBSERVATION

Making predictions of future observations for specified x values is one of the main purposes of linear regression modelling. When we have established that there is a linear relationship between the explanatory variable x and the response variable y, we often want to use that relationship to make predictions of future value y_{n+1} given the value of the explanatory variable x_{n+1}. We can make better predictions using the value of the explanatory variable than without it. The best prediction is

$$\tilde{y}_{n+1} = \hat{\alpha}_{\bar{x}} + \hat{\beta} \times (x_{n+1} - \bar{x}),$$

where $\hat{\beta}$ is the slope estimate, and $\hat{\alpha}_{\bar{x}}$ is the estimate of the intercept of the line $x = \bar{x}$.

How good is the prediction? There are two sources of uncertainty. First, we are using the estimated values of the parameters in the prediction, not the true values, which are unknown. We are considering the parameters to be random variables and have found their posterior distribution in the previous section. Second, the new observation y_{n+1} contains its own observation error e_{n+1}, which will be independent of all previous observation errors. The *predictive distribution* of the next observation y_{n+1} given the value x_{n+1} and the $data$ accounts for both sources of uncertainty. It is denoted $f(y_{n+1}|x_{n+1}, data)$ and is found by Bayes' theorem.

Finding the Predictive Distribution

The predictive distribution is found by integrating the parameters $\alpha_{\bar{x}}$ and β out of the joint posterior distribution of the next observation y_{n+1} and the parameters given the value x_{n+1} and the $data$ that equals the previous observations $(x_1, y_1), \ldots, (x_n, y_n)$:

$$f(y_{n+1}|x_{n+1}, data)$$

$$= \int\int f(y_{n+1}, \alpha_{\bar{x}}, \beta|x_{n+1}, data) d\alpha_{\bar{x}}\, d\beta\,.$$

Integrating out nuisance parameters from the joint posterior like this is known as *marginalization*. This is one of the clear advantages of Bayesian statistics. It has a single method of dealing with nuisance parameters that always works. When we find the predictive distribution, we consider all the parameters to be nuisance parameters.

First, we have to determine the joint posterior distribution of the parameters and next observation, given the value x_{n+1} and the $data$:

$$f(y_{n+1}, \alpha_{\bar{x}}, \beta|x_{n+1}, data)$$

$$= f(y_{n+1}|\alpha_{\bar{x}}, \beta, x_{n+1}, data) \times g(\alpha_{\bar{x}}, \beta|x_{n+1}, data)\,.$$

But the next observation y_{n+1} at the known value x_{n+1} is another random observation from the regression model. Given the parameters $\alpha_{\bar{x}}$ and β, the observations are all independent of each other. This means that given the parameters, the new observation y_{n+1} does not depend on the $data$, which are the previous observations from the regression. Also the posterior of $\alpha_{\bar{x}}, \beta$, given the *data* and the value x_{n+1} does not depend on x_{n+1}. The posterior was calculated from the data alone. So the joint distribution of new observation and parameters simplifies to

$$f(y_{n+1}, \alpha_{\bar{x}}, \beta|x_{n+1}, data)$$

$$= f(y_{n+1}|\alpha_{\bar{x}}, \beta, x_{n+1}) \times g(\alpha_{\bar{x}}, \beta|data)\,.$$

This is the distribution of the next observation given the parameters, times the posterior distribution of the parameters given the previous sample. The next observation $y_{n+1}|\alpha_{\bar{x}}, \beta, x_{n+1}$ is another random observation from the regression model. By our assumptions it is normally distributed with mean given by the linear function of the parameters $\mu_{n+1} = \alpha_{\bar{x}} + \beta(x_{n+1} - \bar{x})$ and known variance σ^2.

The posterior distributions of the parameters given the previous data are independently *normal* $(m'_\alpha, (s'_\alpha)^2)$ and *normal*$(m'_\beta, (s'_\beta)^2)$, which we found using the updating rules in the previous section. The two components of the linear function are independent. Thus posterior distribution of μ_{n+1} will be *normal* with mean $m'_{n+1} = m'_\alpha + (x_{n+1} - \bar{x}) \times m'_\beta$ and variance $(s'_{n+1})^2 = (s'_\alpha)^2 + (x_{n+1} - \bar{x})^2 \times (s'_\beta)^2)$.

Since the next observation only depends on the parameters through the linear function, $\mu_{n+1} = \alpha_{\bar{x}} + \beta(x_{n+1} - \bar{x})$, we will let μ_{n+1} be the parameter. We will find the predictive distribution by marginalizing the μ_{n+1} out of the joint posterior of y_{n+1} and μ_{n+1}.

$$
\begin{aligned}
f(y_{n+1}|x_{n+1}, data) &= \int f(y_{n+1}|\mu_{n+1}, x_{n+1}, data) \times \\
&\quad g(\mu_{n+1}|x_{n+1}, data)\, d\mu_{n+1} \\
&= \int f(y_{n+1}|\mu_{n+1}) \times g(\mu_{n+1}|x_{n+1}, data)\, d\mu_{n+1} \\
&\propto \int e^{-\frac{1}{2\sigma^2}(y_{n+1}-\mu_{n+1})^2} \times e^{-\frac{1}{2(s'_{n+1})^2}(\mu_{n+1}-m'_{n+1})^2}\, d\mu_{n+1} \\
&\propto \int e^{-\frac{1}{2\sigma^2 (s'_{n+1})^2/(\sigma^2+(s'_{n+1})^2)}\left(\mu - \frac{y(s'_{n+1})^2 + m'_{n+1}\sigma^2}{((s'_{n+1})^2+\sigma^2}\right)^2} \\
&\quad \times e^{-\frac{1}{2((s'_{n+1})^2+\sigma^2)}\left(y_{n+1}-m'_{n+1}\right)^2}\, d\mu_{n+1}\,.
\end{aligned}
$$

The second factor doesn't depend on μ_{n+1}, so it can be brought in front of the integral. We recognize that the first term integrates out, so we are left with

$$
f(y_{n+1}|x_{n+1}, data) \propto e^{-\frac{1}{2((s'_{n+1})^2+\sigma^2)}\left(y_{n+1}-m'_{n+1}\right)^2}\,. \tag{13.12}
$$

We recognize that this as a *normal* $(m'_\mu, \sigma^2 + (s'_{n+1})^2$. The predictive distribution of the next observation y_{n+1} at x_{n+1} is normal with mean equal to the posterior mean of $\mu_{n+1} = \alpha_{\bar{x}} + \beta(x_i - \bar{x})$ and variance equal to the posterior variance of $\mu_{n+1} = \alpha_{\bar{x}} + \beta(x_i - \bar{x})$ plus σ^2. The predictive distribution allows for both sources of uncertainty.

Main Points

- Our goal is to use one variable x, called the predictor variable to help us predict another variable y, called the response variable.
- We think the two variables are related by a linear relationship, $y = a_0 + b \times x$. b is the slope and a_0 is the y-intercept (where the line intersects the y-axis.)
- The scatterplot of the points (x, y) would indicate a perfect linear relationship if the points lie along a straight line.

- However, the points usually do not lie perfectly along a line but are scattered around, yet still show a linear pattern.
- We could draw any line on the scatterplot. The residuals from that line would be the vertical distance from the plotted points to the line.
- Least squares is a method for finding a line that best fits a plotted points by minimizing the sum of squares of residuals from a fitted line.
- The slope and intercept of the least squares line are found by solving the *normal equations*.
- The linear regression model has three assumptions:
    1. The mean of y is an unknown linear function of x. Each observation y_i is made at a known value x_i.
    2. Each observation y_i is subject to a random error that is normally distributed with mean 0 and variance σ^2. We will assume that σ^2 is known.
    3. The observation errors are independent of each other.
- Bayesian regression is much easier if we reparameterize the model to be $y = \alpha_{\bar{x}} + \beta \times (x - \bar{x})$.
- The joint likelihood of the sample factors into a part dependent on the slope β and a part dependant on $\alpha_{\bar{x}}$.
- We use independent priors for the slope β and intercept $\alpha_{\bar{x}}$. They can either be normal priors or "flat" priors. The joint prior is the product of the two priors.
- The joint posterior is proportional to the joint prior times the joint likelihood. Since both the joint prior and joint likelihood factor, the joint posterior is the product of two individual posteriors. Each of them is normal where the constants can be found from the simple updating rules.
- Ordinarily we are more interested in the posterior distribution of the slope β, which is *normal* $(m', (s')^2)$. In particular, we are interested in knowing whether the belief $\beta = 0$ is credible given the data. If so, we should not be using x to help predict y.
- The Bayesian credible interval for β is the *posterior mean* $\pm$ the *critical value* $\times$ the *posterior standard deviation*.
- The critical value is taken from the normal table if we assume the variance σ^2 is known. If we don't know it and use the sample estimate calculated from the residuals then we take the critical value from the *Student's t* table.
- The credible interval can be used to test the two-sided hypothesis $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

- We can test a one-sided hypothesis $H_0 : \beta \leq 0$ versus $H_1 : \beta > 0$ by calculating the probability of the null hypothesis, and comparing it to the level of significance.
- We can compute the predictive probability distribution for the next observation y_{n+1} taken when x_{n+1}. It is the *normal* distribution with mean equal to the mean of the linear function $\mu_{n+1} = \alpha_{\bar{x}} + (x_{n+1} - \bar{x})$, and its variance equal to the variance of the linear function plus the observation variance.

Exercises

13.1 A researcher measured heart rate (x) and oxygen uptake (y) for one person under varying exercise conditions. He wishes to determine if heart rate which is easier to measure can be used to predict oxygen uptake. If so, then the estimated oxygen uptake based on the measured heart rate can be used in place of the measured oxygen uptake for later experiments on the individual:

Heart Rate	Oxygen Uptake
x	y
94	.47
96	.75
94	.83
95	.98
104	1.18
106	1.29
108	1.40
113	1.60
115	1.75
121	1.90
131	2.23

(a) Plot a scatterplot of oxygen uptake y versus heart rate x.

(b) Calculate the parameters of the least squares line.

(c) Graph the least squares line on your scatterplot.

(d) Calculate the estimated variance about the least squares line.

(e) Suppose that we know that oxygen uptake given the heart rate is *normal* $(\alpha_0 + \beta \times x, \sigma^2)$, where $\sigma^2 = .13^2$ is known. Use a *normal* $(0, 1^2)$ prior for β. What is the posterior distribution of β?

(f) Find a 95% credible interval for β.

(g) Perform a Bayesian test of

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0$$

at the 95 % level of significance.

13.2 A researcher is investigating the relationship between yield of potatoes (y) and level of fertilizer (x.) She divides a field into eight plots of equal size and applied fertilizer at a different level to each plot. The level of fertilizer and yield for each plot is recorded below:

Fertilizer Level	Yield
x	y
1	25
1.5	31
2	27
2.5	28
3	36
3.5	35
4	32
4.5	34

(a) Plot a scatterplot of yield versus fertilizer level.

(b) Calculate the parameters of the least squares line.

(c) Graph the least squares line on your scatterplot.

(d) Calculate the estimated variance about the least squares line.

(e) Suppose that we know that yield given the fertilizer level is *normal* $(\alpha_0 + \beta \times x, \sigma^2)$, where $\sigma^2 = 3.0^2$ is known. Use a *normal* $(2, 2^2)$ prior for β. What is the posterior distribution of β?

(f) Find a 95% credible interval for β.

(g) Perform a Bayesian test of

$$H_0 : \beta \leq 0 \quad \text{versus} \quad H_1 : \beta > 0$$

at the 95 % level of significance.

13.3 A researcher is investigating the relationship between fuel economy and driving speed. He makes six runs on a test track, each at a different speed, and measures the kilometers travelled on one liter of fuel. The speeds (in kilometers per hour) and distances (in kilometers) are recorded below:

Speed x	Distance y
80	55.7
90	55.4
100	52.5
110	52.1
120	50.5
130	49.2

(a) Plot a scatterplot of distance travelled versus speed.

(b) Calculate the parameters of the least squares line.

(c) Graph the least squares line on your scatterplot.

(d) Calculate the estimated variance about the least squares line.

(e) Suppose that we know distance travelled given the speed is *normal* $(\alpha_0 + \beta \times x, \sigma^2)$ where $\sigma^2 = .57^2$ is known. Use a *normal* $(0, 1^2)$ prior for β. What is the posterior distribution of β?

(f) Perform a Bayesian test of

$$H_0 : \beta \geq 0 \quad \text{versus} \quad H_1 : \beta < 0$$

at the 95 % level of significance.

13.4 The Police Department is interested in determining the effect of alcohol consumption on driving performance. Twelve male drivers of similar weight, age, and driving experience were randomly assigned to three groups of four. The first group consumed two cans of beer within 30 minutes, the second group consumed four cans of beer within 30 minutes, and the third group was the control, and did not consume any beer. Twenty minutes later, each of the twelve took a driving test under the same conditions, and their individual scores were recorded. (The higher score, the better the driving performance.) The results were:

Cans x	Score y
0	78
0	82
0	75
0	58
2	75
2	42
2	50
2	55
4	27
4	48
4	49
4	39

(a) Plot a scatterplot of score versus cans.

(b) Calculate the parameters of the least squares line.

(c) Graph the least squares line on your scatterplot.

(d) Calculate the estimated variance about the least squares line.

(e) Suppose we know that the driving score given the number of cans of beer drunk is *normal* $(\alpha_0 + \beta \times x, \sigma^2)$, where $\sigma^2 = 12^2$ is known. Use a *normal* $(0, 10^2)$ prior for β. What is the posterior distribution of β?

(f) Perform a Bayesian test of

$$H_0 : \beta \geq 0 \quad \text{versus} \quad H_1 : \beta < 0$$

at the 95 % level of significance.

13.5 A textile manufacturer is concerned about the strength of cotton yarn. In order to find out whether fiber length is an important factor in determining the strength of yarn, the quality control manager checked the fiber length (x) and strength (y) for a sample of 10 segments of yarn. The results are:

Fiber Length x	Strength y
85	99
82	93
75	103
73	97
76	91
73	94
96	135
92	120
70	88
74	92

(a) Plot a scatterplot of strength versus fiber length.

(b) Calculate the parameters of the least squares line.

(c) Graph the least squares line on your scatterplot.

(d) Calculate the estimated variance about the least squares line.

(e) Suppose we know that the strength given the fiber length is *normal* $(\alpha_0 + \beta \times x, \sigma^2)$, where $\sigma^2 = 7.7^2$ is known. Use a *normal* $(0, 10^2)$ prior for β. What is the posterior distribution of β.

(f) Find a 95% credible interval for β.

(g) Perform a Bayesian test of

$$H_0 : \beta \leq 0 \quad \text{versus} \quad H_1 : \beta > 0$$

at the 95 % level of significance.

13.6 In Chapter 3, Exercise 7, we were looking at the relationship between log(mass) and log(length) for a sample of 100 New Zealand slugs of the species *Limax maximus* from a study conducted by Barker and McGhie (1984.) These data are in the Minitab worksheet *slug.mtw*. We identified observation 90 that did not appear to fit the pattern It is likely that this observation is an outlier that was recorded incorrectly, so remove it from the data set. The summary statistics for the 99 remaining observations are. Note: x is log(length), and y is log(weight)

$$\sum x = 352.399 \qquad \sum y = -33.6547 \qquad \sum x^2 = 1292.94$$

$$\sum xy = -18.0147 \qquad \sum y^2 = 289.598$$

(a) Calculate the least squares line for the regression of y on x from the formulas.

(b) Using Minitab, calculate the least squares line. Plot a scatterplot of log weight on log length. Include the least squares line on your scatterplot.

(c) Using Minitab, calculate the residuals from the least squares line, and plot the residuals versus x. From this plot, does it appear the linear regression assumptions are satisfied?

(d) Using Minitab, calculate the estimate of the standard deviation of the residuals.

(e) Suppose we use a *normal* $(3, .5^2)$ prior for β, the regression slope coefficient. Calculate the posterior distribution of $\beta|data$. (Use the standard deviation you calculated from the residuals as if it is the true observation standard deviation.)

(f) Find a 95% credible interval for the true regression slope β.

(g) If the slug stay the same shape as they grow (allotropic growth) the height and width would both be proportional to the length, so the weight would be proportional to the cube of the length. In that case the coefficient of log(weight) on log(length) would equal 3. Test the hypothesis

$$H_0 : \beta = 3 \quad \text{versus} \quad H_1 : \beta \neq 3$$

at the 5% level of significance. Can you conclude this slug species shows allotropic growth?

13.7 Endophyte is a fungus *Neotyphodium lolli* that lives inside ryegrass plants. It does not spread between plants, but plants grown from endophyte-infected seed will be infected. One of its effects is that it produces a range of compounds that are toxic to Argentine stem weevil *Listronotus bonariensis*, which feeds on ryegrass. AgResearch New Zealand did a study on the persistence of perennial ryegrass at four rates of Argentine stem weevil infestation. For ryegrass that was infected with endophyte the following data were observed:

Infestation Rate x	Number of Ryegrass Plants (n)	$\log_e(n+1)$ y
0	19	2.99573
0	23	3.17805
0	2	1.09861
0	0	0.00000
0	24	3.21888
5	20	3.04452
5	18	2.94444
5	10	2.39790
5	6	1.94591
5	6	1.94591
10	12	2.56495
10	2	1.09861
10	11	2.48491
10	7	2.07944
10	6	1.94591
20	3	1.38629
20	16	2.83321
20	14	2.70805
20	9	2.30259
20	12	2.56495

(a) Plot a scatterplot of number of ryegrass plants versus the infestation rate.

(b) The relationship between infestation rate and number of ryegrass plants is clearly nonlinear. Look at the transformed variable $y = \log_e(n+1)$. Plot y versus x on a scatterplot. Does this appear to be more linear?

(c) Find the least squares line relating y to x. Include the least squares line on your scatterplot.

(d) Find the estimated variance about the least squares line.

(e) Assume that the observed y_i are normally distributed with mean $\alpha_{\bar{x}} + \beta \times (x_i - \bar{x})$ and known variance σ^2 equal to that calculated in part (b.) Find the posterior distribution of $\beta|(x_1, y_1), \ldots, (x_{20}, y_{20})$. Use a *normal* $(0, 1^2)$ prior for β.

13.8 For ryegrass that was not infected with endophyte the following data were observed:

Infestation Rate x	Number of Ryegrass Plants (n)	$\log_e(n+1)$ y
0	16	2.83321
0	23	3.17805
0	2	1.09861
0	16	2.83321
0	6	1.94591
5	8	2.19722
5	6	1.94591
5	1	0.69315
5	2	1.09861
5	5	1.79176
10	5	1.79176
10	0	0.00000
10	6	1.94591
10	2	1.09861
10	2	1.09861
20	1	0.69315
20	0	0.00000
20	0	0.00000
20	1	0.69315
20	0	0.00000

(a) Plot a scatterplot of number of ryegrass plants versus the infestation rate.

(b) The relationship between infestation rate and number of ryegrass plants is clearly nonlinear. Look at the transformed variable $y = \log_e(n+1)$. Plot y versus x on a scatterplot. Does this appear to be more linear?

(c) Find the least squares line relating y to x.

(d) Find the estimated variance about the least squares line.

(e) Assume that the observed y_i are normally distributed with mean $\alpha_{\bar{x}} + (x_i - \bar{x}) \times \beta$ and variance equal to that calculated in part (b.) Find the posterior distribution of $\beta|(x_1, y_1), \ldots, (x_{20}, y_{20})$. Use a *normal* $(0, 1^2)$ prior for β.

13.9 In the previous two problems we found the posterior distribution of the slope of y on x, the rate of weevil infestation for endophyte infected and noninfected ryegrass. Let β_1 be the slope for noninfected ryegrass, and let β_2 be the slope for infected ryegrass

(a) Find the posterior distribution of $\beta_1 - \beta_2$.

(b) Calculate a 95 % credible interval for $\beta_1 - \beta_2$.

(c) Test the hypothesis

$$H_0 : \beta_1 - \beta_2 \leq 0 \quad \text{versus} \quad H_1 : \beta_1 - \beta_2 > 0$$

at the 10 % level of significance.

14

Robust Bayesian Methods

Many statisticians hesitate to use Bayesian methods because they are reluctant to let their prior belief into their inferences. In almost all cases they have some prior knowledge, but they may not wish to formalize it into a prior distribution. They know some values are more likely than others, and some are not realistically possible. Scientists are studying and measuring something they have observed. They know the scale of possible measurements. We saw in previous chapters that all priors that have reasonable probability over the range of possible values will give similar, although not identical posteriors. And we saw that Bayes' theorem using the prior information will give better inferences than frequentist ones that ignore prior information, even when judged by frequentist criteria. The scientist would be better off if he formed a prior from his prior knowledge and used Bayesian methods.

However, it is possible that a scientist could have a strong prior belief, yet that belief could be incorrect. When the data are taken, the likelihood is found to be very different from that expected from the prior. The posterior would be strongly influenced by the prior. Most scientists would be very reluctant to use that posterior. If there is a strong disagreement between the prior and the likelihood, the scientist would want to go with the likelihood, since it came from the data.

In this chapter we look at how we can make Bayesian inference more robust against a poorly specified prior. We find that using a *mixture* of conjugate priors enables us to do this. We allow a small prior probability that our prior is misspecified. If the likelihood is very different than what would be expected under the prior, the

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

posterior probability of misspecification is large, and our posterior distribution will depend mostly on the likelihood.

14.1 EFFECT OF MISSPECIFIED PRIOR

One of the main advantages of Bayesian methods is that it uses your prior knowledge, along with the information from the sample. Bayes' theorem combines both prior and sample information into the posterior. Frequentist methods only use sample information. Thus Bayesian methods usually perform better than frequentist ones because they are using more information. The prior should have relatively high values over the whole range where the likelihood is substantial.

However, sometimes this does not happen. A scientist could have a strong prior belief, yet it could be wrong. Perhaps he (wrongly) bases his prior on some past data that arose from different conditions than the present data set. If a strongly specified prior is incorrect, it has a substantial effect on the posterior. This is shown in the following two examples.

Example 25 *Archie is going to conduct a survey about how many Hamilton voters say they will attend a casino if it is built in town. He decides to base his prior on the opinions of his friends. Out of the 25 friends he asks, 15 say they will attend the casino. So he decides on a beta*(a, b) *prior that matches those opinions. The prior mean is .6, and the equivalent samples size is 25. Thus* $a + b + 1 = 25$ *and* $\frac{a}{a+b} = .6$. *Thus* $a = 14.4$ *and* $b = 9.6$. *Then he takes a random sample of 100 Hamilton voters and finds that 25 say they will attend the casino. His posterior distribution is beta*$(39.4, 84.60)$. *Archie's prior, the likelihood, and his posterior are shown in Figure 14.1. We see that the prior and the likelihood do not overlap very much. The posterior is in between. It gives high posterior probability to values that aren't supported strongly by the data (likelihood) and aren't strongly supported by prior either. This is not satisfactory.*

Example 26 *Andrea is going to take a sample of measurements of dissolved oxygen level from a lake during the summer. Assume that the dissolved oxygen level is approximately normal with mean* μ *and known variance* $\sigma^2 = 1$. *She had previously done a similar experiment from the river that flowed into the lake. She considered that she had a pretty good idea of what to expect. She decided to use a normal*$(8.5, .7^2)$ *prior for* μ, *which was similar to her river survey results. She takes a random sample of size 5 and the sample mean is 5.45. The parameters of the posterior distribution are found using the simple updating rules for normal. The posterior is normal*$(6.334, .3769^2)$. *The prior, likelihood, and posterior are shown in 2. The posterior density is between the prior and likelihood, and gives high probability to values that aren't supported strongly either by the data or by the prior, which is a very unsatisfactory result. Figure 14.2 shows Andrea's prior, likelihood, and posterior.*

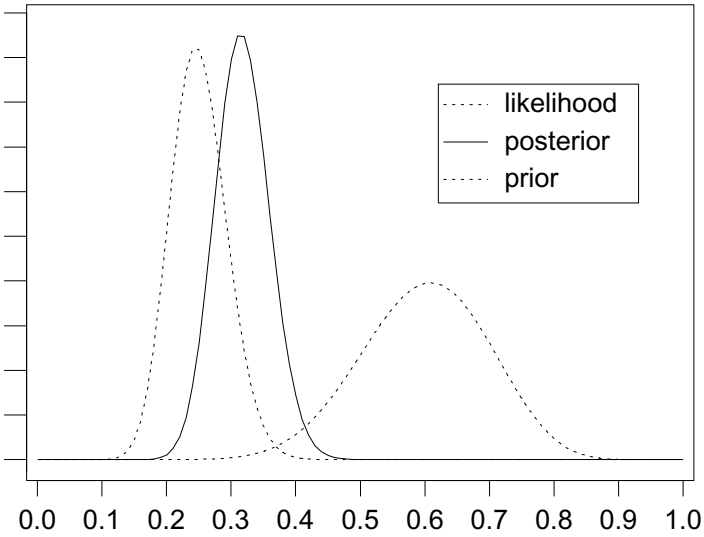


Figure 14.1 Archie's prior, likelihood, and posterior.

These two examples show how an incorrect prior can arise. Both Archie and Andrea based their priors on past data, each judged to arise from a situation similar the one to be analyzed. They were both wrong. In Archie's case he considered his friends to be representative of the population. However, they were all similar in age and outlook to him. They do not constitute a good data set to base a prior on. Andrea considered that her previous data from the river survey would be similar to data from the lake. She neglected the effect of water movement on dissolved oxygen. She is basing her prior on data obtained from an experiment under different conditions than the one she is now undertaking.

14.2 BAYES' THEOREM WITH MIXTURE PRIORS

Suppose our prior density is $g_0(\theta)$ and it is quite precise, because we have substantial prior knowledge. However, we want to protect ourselves from the possibility that we misspecified the prior by using prior knowledge that is incorrect. We don't consider it likely, but concede that it is possible that we failed to see the reason why our prior knowledge will not applicable to the new data. If our prior is misspecified, we don't really have much of an idea what values θ should take. In that case the prior for θ is $g_1(\theta)$, which is either a very vague conjugate prior or a flat prior. Let $g_0(\theta|y_1, \cdots, y_n)$ be the posterior distribution of θ given the observations when we start with $g_0(\theta)$ as the prior. Similarly we let $g_1(\theta|y_1, \cdots, y_n)$ be the posterior distribution of θ given

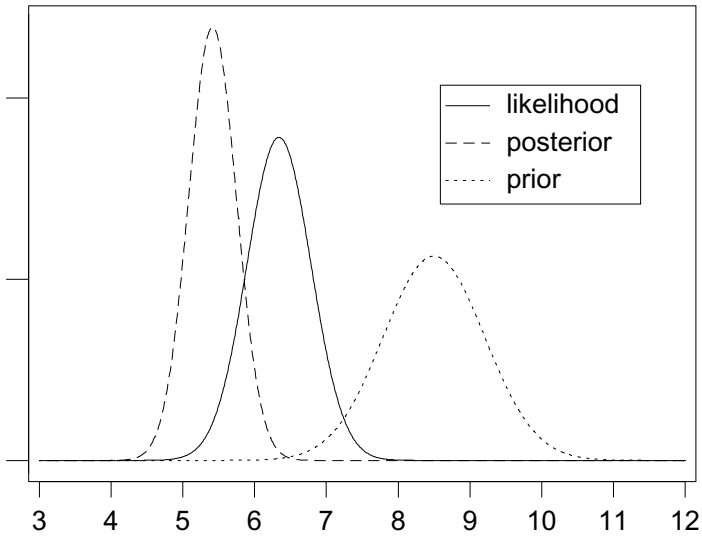


Figure 14.2 Andrea's prior, likelihood, and posterior.

the observations when we start with $g_1(\theta)$ as the prior:

$$g_i(\theta|y_1, \cdots, y_n) \propto g_i(\theta) f(y_1, \cdots, y_n|\theta)\,.$$

These are found using the simple updating rules, since we are using priors that are either from the conjugate family or are flat.

The Mixture Prior

We introduce a new parameter, I that takes two possible values. If $i = 0$, then θ comes from $g_0(\theta)$. However, if $i = 1$, then θ comes from $g_1(\theta)$. The conditional prior probability of θ given i is

$$g(\theta|i) = \begin{cases} g_0(\theta) & \text{if} \quad i = 0 \\ g_1(\theta) & \text{if} \quad i = 1 \end{cases} .$$

We let the prior probability distribution of I be $P(I = 0) = p_0$, where p_0 is some high value like .9, .95, or .99, because we think our prior $g_0(\theta)$ is correct. The prior probability that our prior is misspecified is $p_1 = 1 - p_0$. The joint prior distribution of θ and I is

$$g(\theta, i) = p_i \times g_i(\theta) \quad \text{for} \quad i = 0, 1\,.$$

We note this joint distribution is continuous in the parameter θ and discrete in the parameter I. The marginal prior density of the random variable θ is found by

marginalizing (summing I over all possible values) the joint density. It has a *mixture* prior distribution since its density

$$g(\theta) = \sum_0^1 p_i g_i(\theta) \tag{14.1}$$

is a mixture of the two prior densities.

The Joint Posterior

The joint posterior distribution of θ, I given the observations $y_1, \ldots, y_n$ is proportional to the joint prior times the joint likelihood. This gives

$$g(\theta, i|y_1, \cdots, y_n) = c \times g(\theta, i) \times f(y_1, \cdots, y_n|\theta, i) \quad \text{for} \quad i = 0, 1$$

for some constant c. But the sample only depends on θ, not on i, so the joint posterior

$$\begin{aligned} g(\theta, i|y_1, \cdots, y_n) &= c \times p_i g_i(\theta) f(y_1, \cdots, y_n|\theta) \quad \text{for} \quad i = 0, 1 \\ &= c \times p_i h_i(\theta, y_1, \cdots, y_n) \quad \text{for} \quad i = 0, 1\,, \end{aligned}$$

where $h_i(\theta, y_1, \cdots, y_n) = g_i(\theta) f(y_1, \ldots, y_n|\theta)$ is the joint distribution of the parameter and the data, when $g_i(\theta)$ is the correct prior. The marginal posterior probability $P(I = i|y_1, \cdots, y_n)$ is found by integrating θ out of the joint posterior:

$$\begin{aligned} P(I = i|y_1, \cdots, y_n) &= \int g(\theta, i|y_1, \cdots, y_n) d\theta \\ &= c \times p_i \int h_i(\theta, y_1, \cdots, y_n) d\theta \\ &= c \times p_i f_i(y_1, \ldots, y_n) \end{aligned}$$

for $i = 0, 1$, where $f_i(y_1, \ldots, y_n)$ is the marginal probability (or probability density) of the data when $g_i(\theta)$ is the correct prior. The posterior probabilities sum to 1, and the constant c cancels, so

$$P(I = i|y_1, \ldots, y_n) = \frac{p_i f_i(y_1, \cdots, y_n)}{\sum_{i=0}^1 p_i f_i(y_1, \cdots, y_n)}\,.$$

These can be easily evaluated.

The Mixture Posterior

We find the marginal posterior of θ by summing all possible values of i out of the joint posterior:

$$g(\theta|y_1, \cdots, y_n) = \sum_{i=0}^1 g(\theta, i|y_1, \cdots, y_n)\,.$$

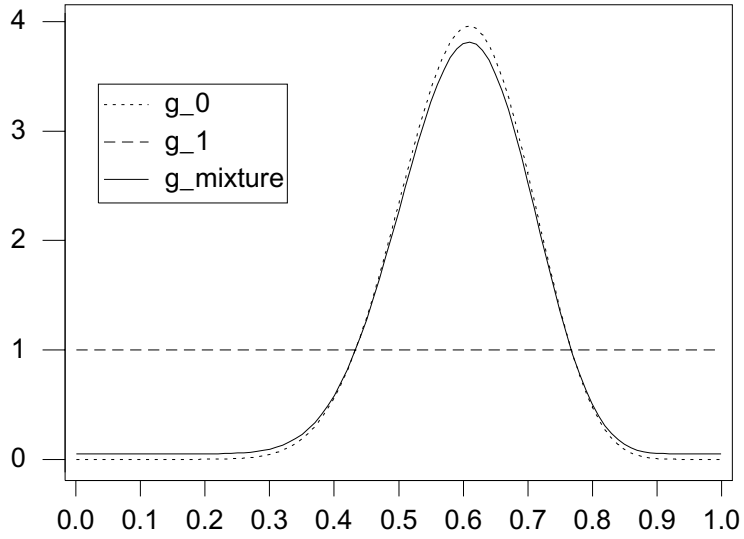


Figure 14.3 Ben's mixture prior and components.

But there is another way the joint posterior can be rearranged from conditional probabilities:

$$g(\theta, i|y_1, \cdots, y_n) = g(\theta|i, y_1, \cdots, y_n) \times P(I = i|y_1, \cdots, y_n)\,,$$

where $g(\theta|i, y_1, \ldots, y_n) = g_i(\theta|y_1, \ldots, y_n)$ is the posterior distribution when we started with $g_i(\theta)$ as the prior. Thus the marginal posterior of θ is

$$g(\theta|y_1, \cdots, y_n) = \sum_{i=0}^{1} g_i(\theta|y_1, \cdots, y_n) \times P(I = i|y_1, \cdots, y_n)\,. \tag{14.2}$$

This is the mixture of the two posteriors, where the weights are the posterior probabilities of the two values of i given the data.

Example 25 (continued) *One of Archie's friends, Ben, decided that he would reanalyze Archie's data with a mixture prior. He let g_0 be the same beta(14.4,9.6) prior that Archie used. He let g_1 be the (uniform) beta(1,1) prior. He let the prior probability $p_0 = .95$. Ben's mixture prior and its components are shown in Figure 14.3. His mixture prior is quite similar to Archie's. However, it has heavier weight in the tails. This gives makes his prior robust against prior misspecification. In this case, $h_i(\pi, y)$ is a product of a beta times a binomial. Of course, we are only*

interested in $y = 25$, the value that occurred:

$$\begin{aligned} h_0(\pi, y = 25) &= \frac{\Gamma(24)}{\Gamma(14.4)\Gamma(9.6)}\pi^{13.4}(1-\pi)^{8.6} \times \left(\frac{100!}{25!75!}\right)\pi^{25}(1-\pi)^{75} \\ &= \frac{\Gamma(24)}{\Gamma(14.4)\Gamma(9.6)} \times \left(\frac{100!}{25!75!}\right) \times \pi^{38.4}(1-\pi)^{83.6} \end{aligned}$$

and

$$\begin{aligned} h_1(\pi, y = 25) &= \pi^0(1-\pi)^0 \times \left(\frac{100!}{25!75!}\right)\pi^{25}(1-\pi)^{75} \\ &= \left(\frac{100!}{25!75!}\right)\pi^{25}(1-\pi)^{75}\,. \end{aligned}$$

We recognize each of these as a constant times a beta distribution. So integrating them with respect to π gives

$$\begin{aligned} \int_0^1 h_0(\pi, y = 25)d\pi &= \frac{\Gamma(24)}{\Gamma(14.4)\Gamma(9.6)} \times \left(\frac{100!}{25!75!}\right) \times \int_0^1 \pi^{38.4}(1-\pi)^{83.6}d\pi \\ &= \frac{\Gamma(24)}{\Gamma(14.4)\Gamma(9.6)} \times \left(\frac{100!}{25!75!}\right) \times \frac{\Gamma(39.4)\Gamma(84.6)}{\Gamma(124)} \end{aligned}$$

and

$$\begin{aligned} \int_0^1 h_1(\pi, y = 25)d\pi &= \left(\frac{100!}{25!75!}\right) \times \int_0^1 \pi^{25}(1-\pi)^{75}d\pi \\ &= \left(\frac{100!}{25!75!}\right) \times \frac{\Gamma(26)\Gamma(76)}{\Gamma(102)}\,. \end{aligned}$$

Remember that $\Gamma(a) = (a-1) \times \Gamma(a-1)$ and if a is an integer, $\Gamma(a) = (a-1)!$. The second integral is easily evaluated and gives

$$f_1(y = 25) = \int_0^1 h_1(\pi, y = 25)d\pi = \frac{1}{101} = 9.90099 \times 10^{-3}\,.$$

We can evaluate the second integral numerically

$$f_0(y = 25) = \int_0^1 h_0(\pi, y = 25)d\pi = 2.484 \times 10^{-4}\,.$$

So the posterior probabilities are $P(I = 0|25) = 0.323$ and $P(I = 1|25) = 0.677$. The posterior distribution is the mixture $g(\pi|25) = .323 \times g_0(\pi|25) + .677 \times g_1(\pi|25)$, where $g_0(\pi|y)$ and $g_1(\pi|y)$ are the conjugate posterior distributions found using g_0 and g_1 as the respective priors. Ben's mixture posterior distribution and its two components is shown in Figure 14.4. Ben's prior and posterior, together with the likelihood is shown in Figure 14.5. When the prior and likelihood disagree, we should go with the likelihood because it is from the data. Superficially, Ben's prior looks

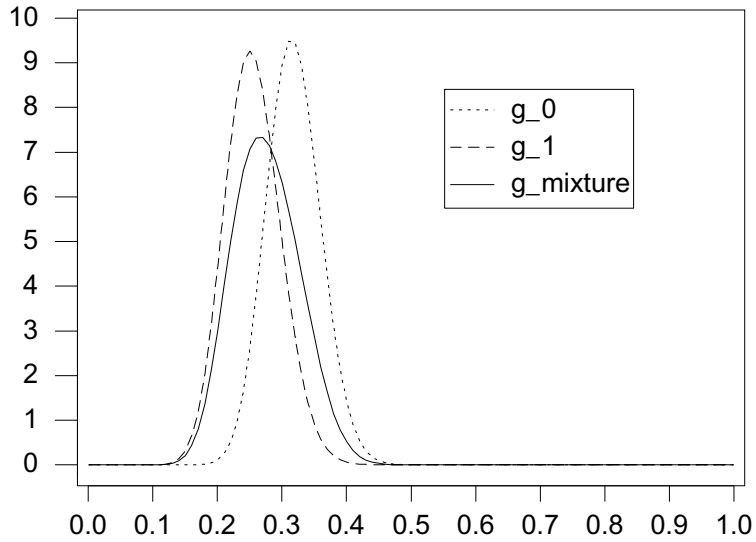


Figure 14.4 Ben's mixture posterior and its two components.

very similar to Archie's prior. However, it has a heavier tail allowed by the mixture, and this has allowed his posterior to be very close to the likelihood. We see that this is much more satisfactory than Archie's analysis shown in Figure 14.1.

Example 26 (continued) *Andrea's friend Caitlin looked at Figure 14.2 and told her it was not satisfactory. The values given high posterior probability were not supported strongly either by the data, or by the prior. She considered it likely that the prior was misspecified. She said to protect against that, she would do the analysis using a mixture of normal priors.* $g_0(\theta)$ *was the same as Andrea's, normal*$(8.5, .7^2)$*, and* $g_1(\theta)$ *would be normal* $(8.5, (4 \times .7)^2$*, which has the same mean as Andrea's prior, but with the standard deviation 4 times as large. She allows prior probability .05 that Andrea's prior was misspecified. Caitlin's mixture prior and its components are shown in Figure 14.6. We see that her mixture prior appears very similar to Andrea's except there is more weight in the tail regions. Caitlin's posterior* $g_0(\theta|\bar{y})$ *is normal*$(6.334, .3769^2)$*, the same as for Andrea. Caitlin's posterior when the original prior was misspecified* $g_1(\theta|\bar{y})$ *is normal*$(5.526, .4416^2)$ *where the parameters are found by the simple updating rules for the normal. In the normal case*

$$\begin{aligned} h_i(\mu, y_1, \cdots, y_n) &\propto g_i(\mu) \times f(\bar{y}|\mu) \\ &\propto e^{-\frac{1}{2s_i^2}(\mu - m_i)^2} \times e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2} \end{aligned}$$

where m_i *and* s_i^2 *are the mean and variance of the prior distribution* $g_i(\mu)$*. The integral* $\int h_i(\mu, y_1, \cdots, y_n)d\mu$ *gives the unconditional probability of the sample, when* g_i *is the correct prior. We multiply out the two terms, rearrange all the terms*

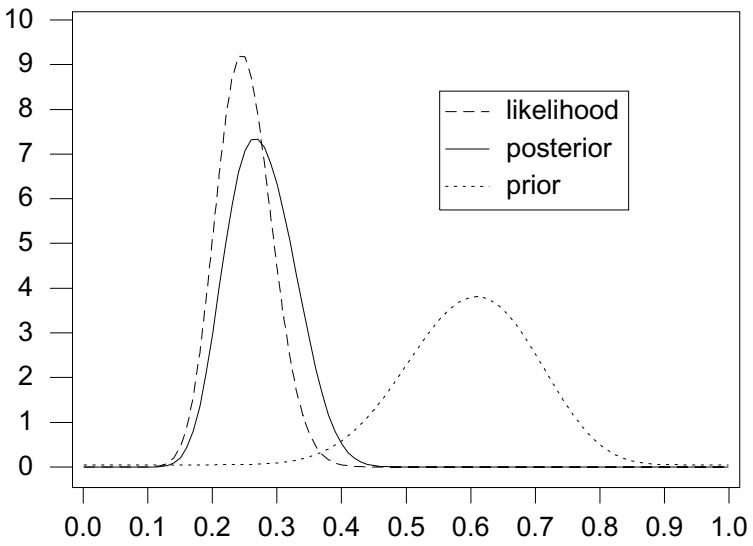


Figure 14.5 Ben's mixture prior, likelihood, and mixture posterior.

containing μ which is normal and integrates. The terms that are left simplify to

$$\begin{aligned} f_i(\bar{y}) &= \int h_i(\mu, \bar{y}) d\mu \\ &\propto e^{-\frac{1}{s_i^2+\sigma^2/n}(\bar{y}-m_i)^2}, \end{aligned}$$

which we recognize as a normal density with mean m_i and variance $\frac{\sigma^2}{n} + s_i^2$. In this example, $m_0 = 8.5, s_0^2 = .7^2, m_1 = 8.5, s_1^2 = (4 \times .7)^2), \sigma^2 = 1$, and $n = 5$. The data are summarized by the value $\bar{y} = 5.45$ that occurred in the sample. Plugging in these values we get $P(I = 0|\bar{y} = 5.45) = .12$ and $P(I = 1|\bar{y} = 5.45) = .88$. Thus Caitlin's posterior is the mixture $.12 \times g_0(\mu|\bar{y}) + .88 \times g_1(\mu|\bar{y})$. Caitlin's mixture posterior and its components are given in Figure 14.7. Caitlin's prior, likelihood, and posterior are shown in Figure 14.8. Comparing this with Andrea's analysis shown in Figure 14.2, we see that using mixtures has given her a posterior that is much closer to the likelihood than the one obtained with the original misspecified prior. This is a much more satisfactory result.

Summary

Our prior represents our prior belief about the parameter before looking at the data from this experiment. We should be getting our prior from past data from similar experiments. However, if we think an experiment is similar, but it is not, our prior can be quite misspecified. We may think we know a lot about the parameter, but

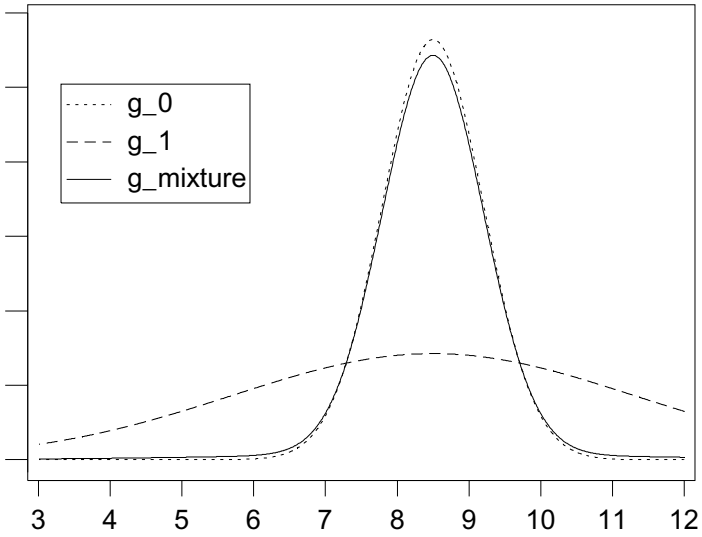


Figure 14.6 Caitlin's mixture prior and its components.

what we think is wrong. That makes the prior quite precise, but wrong. It will be quite a distance from the likelihood. The posterior will be in between, and will give high probability to values neither supported by the data or the prior. That is not satisfactory. If there is a conflict between the prior and the data, we should go with the data.

We introduce a indicator random variable that we give a small prior probability of indicating our original prior is misspecified. The mixture prior we use is the $P(I = 0) \times g_0(\theta) + P(I = 1) \times g_1(\theta)$, where g_0 and g_1 are the original prior, and a more widely spread prior respectively. We find the joint posterior of distribution of I and θ given the data. The marginal posterior distribution of θ given the data is found by marginalizing the indicator variable out. It will be a the mixture distribution

$$g_{mixture}(\theta|y_1, \cdots, y_n) = P(I = 0|y_1, \cdots, y_n)g_0(\theta|y_1, \cdots, y_n)$$

$$+P(I = 1|y_1, \cdots, y_n)g_1(\theta|y_1, \cdots, y_n)\,.$$

This posterior is very robust against a misspecified prior. If the original prior is correct, the mixture posterior will be very similar to the original posterior. However, if the original prior is very far from the likelihood, the posterior probability $p(i = 0|y_1, \cdots, y_n)$ will be very small, and the mixture posterior will be close to the likelihood. This has resolved the conflict between the original prior and the likelihood by giving much more weight to the likelihood.

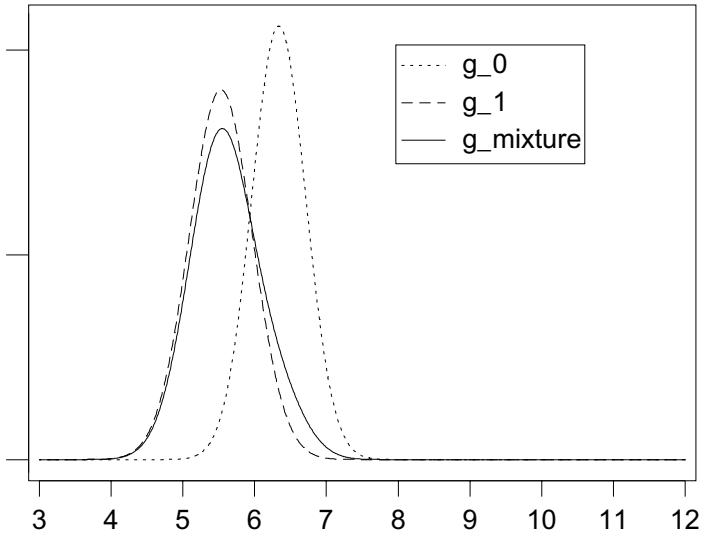


Figure 14.7 Caitlin's mixture posterior and its two components.

Main Points

- If the prior places high probability on values that have low likelihood, and low probability on values that have high likelihood, the posterior will place high probability on values that are not supported either by the prior or by the likelihood. This is not satisfactory.
- This could have been caused by a misspecified prior that arose when the scientist based his/her prior on past data, which had been generated by a process that differs from the process that will generate the new data in some important way that the scientist failed to take into consideration.
- Using mixture priors protects against this possible misspecification of the prior. We use mixtures of conjugate priors. We do this by introducing a mixture index random variable that takes on the values 0 or 1. The mixture prior is

$$g(\theta) = p_0 \times g_0(\theta) + p_1 \times g_1(\theta)\,,$$

  where $g_0(\theta)$ is the original prior we believe in, and g_1 is another prior that has heavier tails, and thus allows for our original prior being wrong. The respective posteriors that arise using each of the priors are $g_0(\theta|y_1, \ldots, y_n)$ and $g_1(\theta|y_1, \ldots, y_n)$.
- We give the original prior g_0 high prior probability by letting the prior probability $p_0 = P(I = 0)$ be high and the prior probability $p_1 = (1-p_0) = P(I = 1)$

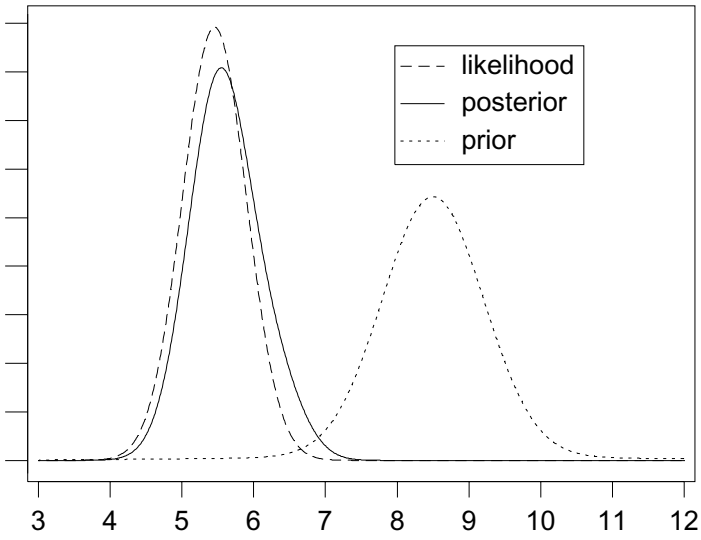


Figure 14.8 Caitlin's mixture prior, the likelihood, and her mixture posterior.

is low. We think the original prior is correct, but have allowed a small probability that we have it wrong.

- Bayes' theorem is used on the mixture prior to determine a mixture posterior. The mixture index variable is a nuisance parameter, and is marginalized out.

- If the likelihood has most of its value far from the original prior, the mixture posterior will be close to the likelihood. This is a much more satisfactory result. When the prior and likelihood are conflicting, we should base our posterior belief mostly on the likelihood, because it is based on the data. Our prior was based on faulty reasoning from past data that failed to note some important change in the process we are drawing the data from.

- The mixture posterior is a mixture of the two posteriors, where the mixing proportions $P(I = i)$ for $i = 0, 1$, are proportional to the prior probability times the the marginal probability (or probability density) evaluated at the data that occurred.

$$P(I = i) \quad \propto p_i \times f_i(y_1, \ldots y_n) \quad \text{for} \quad i = 0, 1\,.$$

- They sum to 1, so

$$P(I = i) \quad = \frac{p_i \times f_i(y_1, \ldots y_n)}{\sum_{i=0}^{1} p_i \times f_i(y_1, \ldots y_n)} \quad \text{for} \quad i = 0, 1\,.$$

Exercises

14.1 You are going to conduct a survey of the voters in the city you live in. They are being asked whether or not the city should build a new convention facility. You believe that most of the voters will disapprove the proposal because it may lead to increased property taxes for residents. As a resident of the city, you have been hearing discussion about this proposal, and most people have voiced disapproval. You think that only about 35% of the voters will support this proposal, so you decide that a *beta* $(7, 13)$ summarizes your prior belief. However, you have a nagging doubt that the group of people you have heard voicing their opinions is representative of the city voters. Because of this, you decide to use a mixture prior:

$$g(\pi|i) = \begin{cases} g_0(\pi) & \text{if} \quad i = 0 \\ g_1(\pi) & \text{if} \quad i = 1 \end{cases},$$

where $g_0(\pi)$ is the *beta* $(7, 13)$ density, and $g_1(\pi)$ is the *beta* $(1, 1)$ (uniform) density. The prior probability $P(I = 0) = .95$. You take a random sample of $n = 200$ registered voters who live in the city. Of these, $y = 10$ support the proposal.

(a) Calculate the posterior distribution of π when $g_0(\pi)$ is the prior.

(b) Calculate the posterior distribution of π when $g_1(\pi)$ is the prior.

(c) Calculate the posterior probability $P(I = 0|Y)$.

(d) Calculate the marginal posterior $g(\pi|Y)$.

14.2 You are going to conduct a survey of the students in your university to find out whether they read the student newspaper regularly. Based on your friends opinions, you think that a strong majority of the students do read the paper regularly. However, you are not sure your friends are representative sample of students. Because of this, you decide to use a mixture prior.

$$g(\pi|i) = \begin{cases} g_0(\pi) & \text{if} \quad i = 0 \\ g_1(\pi) & \text{if} \quad i = 1 \end{cases},$$

where $g_0(\pi)$ is the *beta* $(20, 5)$ density, and $g_1(\pi)$ is the *beta* $(1, 1)$ (uniform) density. The prior probability $P(I = 0) = .95$. You take a random sample of $n = 100$ students. Of these, $y = 41$ say they read the student newspaper regularly.

(a) Calculate the posterior distribution of π when $g_0(\pi)$ is the prior.

(b) Calculate the posterior distribution of π when $g_1(\pi)$ is the prior.

(c) Calculate the posterior probability $P(I = 0|Y)$.

(d) Calculate the marginal posterior $g(\pi|Y)$.

14.3 You are going take a sample of measurements of specific gravity of a chemical product being produced. You know the specific gravity measurements are approximately *normal* (μ, σ^2) where $\sigma^2 = .005^2$. You have precise *normal* $(1.10, .001^2)$ prior for μ because the manufacturing process is quite stable. However, you have a nagging doubt about whether the process is correctly adjusted, so you decide to use a mixture prior. You let $g_0(\mu)$ be your precise *normal* $(1.10, .001^2)$ prior, you let $g_1(\mu)$ be a *normal* $(1.10, .01^2)$, and you let $p_0 = .95$. You take a random sample of product and measure the specific gravity. The measurements are:

1.10352	1.10247	1.10305	1.10415	1.10382	1.10187

(a) Calculate the joint posterior distribution of I and μ given the data.

(b) Calculate the posterior probability $P(I = 0|y_1, \ldots, y_6)$.

(c) Calculate the marginal posterior $g(\mu|y_1, \ldots, y_6)$.

14.4 You are going take a sample of 500 gm blocks of cheese. You know they are approximately *normal* (μ, σ^2) where $\sigma^2 = 2^2$. You have a precise *normal* $(502, 1^2)$ prior for μ because this is what the the process is set for. However, you have a nagging doubt that maybe the machine needs adjustment, so you decide to use a mixture prior. You let $g_0(\mu)$ be your precise *normal* $(502, 1^2)$ prior, and you let $g_1(\mu)$ be a *normal* $(502, 2^2)$, and you let $p_0 = .95$. You take a random sample of ten blocks of cheese and weigh them. The measurements are:

501.5	499.1	498.5	499.9	500.4
498.9	498.4	497.9	498.8	498.6

(a) Calculate the joint posterior distribution of I and μ given the data.

(b) Calculate the posterior probability $P(I = 0|y_1, \ldots, y_{10})$.

(c) Calculate the marginal posterior $g(\mu|y_1, \ldots, y_{10})$.

A

Introduction to Calculus

FUNCTIONS

A function $f(x)$ defined on a set of real numbers, A, is a rule that associates each real number x in the set A with one and only one other real number y. The number x is associated with the number y by the rule $y = f(x)$. The set A is called the *domain* of the function, and the set of all y that are associated with members of A is called the *range* of the function.

Often the rule is expressed as an equation. For example, the domain A might be all positive real numbers, and the function $f(x) = \log_e(x)$ associates each element of A with its natural logarithm. The range of this function is the set of all real numbers.

For a second example, the domain A might be the set of real numbers in the interval $[0, 1]$ and the function $f(x) = x^4 \times (1 - x)^6$. The range of this function is the set of real numbers in the interval $[0, .4^4 \times .6^6]$.

Note that the variable name is merely a cypher, or a place holder. $f(x) = x^2$ and $f(z) = z^2$ are the same function, where the rule of the function is associate each number with its square. The function is the rule by which the association is made. We could refer to the function as f without the variable name, but usually we will refer to it as $f(x)$. The notation $f(x)$ is used for two things. First, it represents the specific value associated by the function f to the point x. Second, it represents the

*

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

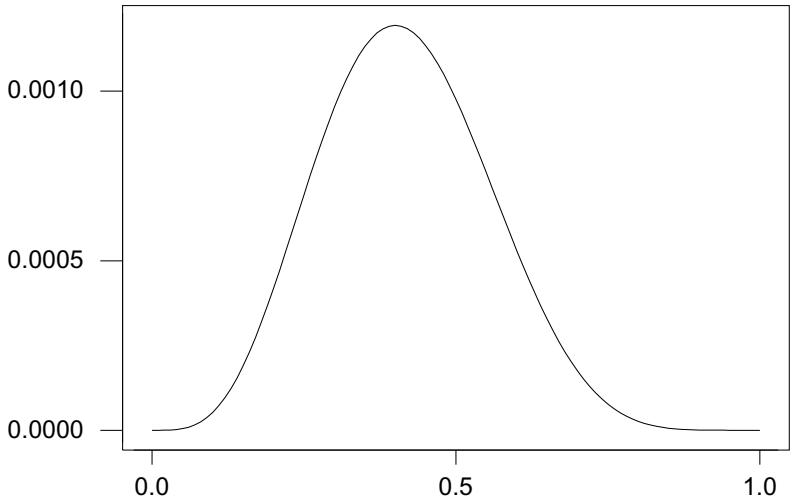


Figure A.1 Graph of function $f(x) = x^4 \times (1-x)^6$.

function by giving the rule which it uses. Generally, there is no confusion as it is clear from the context which meaning we are using.

Combining Functions

We can combine two functions algebraically. Let f and g be functions having the same domain A, and let k_1 and k_2 be constants. The function $h = k_1 \times f$ associates a number x with $y = k_1 f(x)$. Similarly the function $s = k_1 f \pm k_2 g$ associates the number x with $y = k_1 \times f(x) \pm k_2 \times g(x)$. The function $u = f \times g$ associates a number x with $y = f(x) \times g(x)$. Similarly the function $v = \frac{f}{g}$ associates the number x with $y = \frac{f(x)}{g(x)}$.

If function g has domain A and function f has domain that is a subset of the range of the function g, then the composite function (function of a function) $w = f(g)$ associates a number x with $y = f(g(x))$.

Graph of a Function

The graph of the function f is the graph of the equation $y = f(x)$. The graph consists of all points $(x, f(x))$ where $x \in A$ plotted in the coordinate plane. The graph of the function f defined on the closed interval $A = [0, 1]$ where $f(x) = x^4 \times (1-x)^6$ is shown in Figure A.1. The graph of the function g defined on the open interval $A = (0, 1)$, where $g(x) = x^{-\frac{1}{2}} \times (1-x)^{-\frac{1}{2}}$ is shown in the Figure A.2.

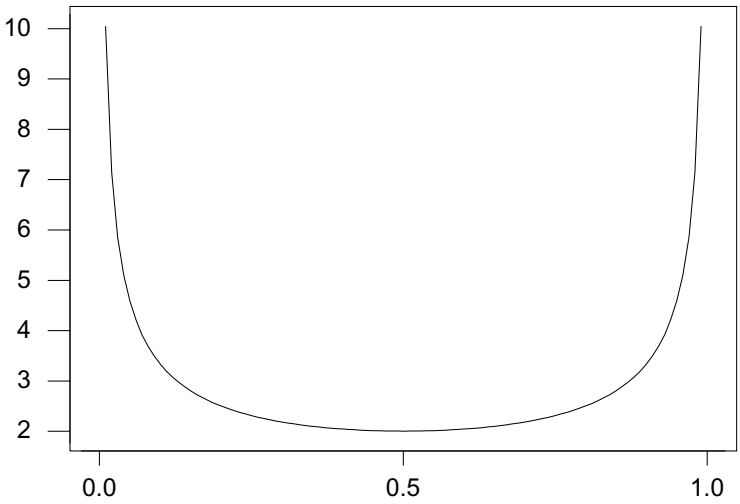


Figure A.2 Graph of function $f(x) = x^{-\frac{1}{2}} \times (1-x)^{-\frac{1}{2}}$.

Limit of a Function

The limit of a function at a point is one of the fundamental tools of calculus. We write

$$\lim_{x \to a} f(x) = b$$

to indicate that b is the limit of the function f when x approaches a. Intuitively, this means that as we take x values closer and closer to (but not equal to) a, their corresponding values of $f(x)$ are getting closer and closer to b. We note that the function $f(x)$ does not have to be defined at a to have a limit at a. For example, 0 is not in the domain A of the function $f(x) = \frac{sinx}{x}$ because division by 0 is not allowed. Yet

$$\lim_{x \to 0} \frac{\sin(x)}{x} = 1$$

as seen in Figure A.3. We see that if we want to be within a specified closeness to $y = 1$, we can find a degree of closeness to $x = 0$ such that all points x that are within that degree of closeness to $x = 0$ and are in the domain A will have $f(x)$ values within that specified closeness to $y = 1$.

We should note that a function may not have a limit at a point a. For example, the function $f(x) = cos(1/x)$ does not have a limit at $x = 0$. This is shown in Figure A.4, which shows the function at three scales. No matter how close we get to $x = 0$, the possible $f(x)$ values always range from -1 to 1.

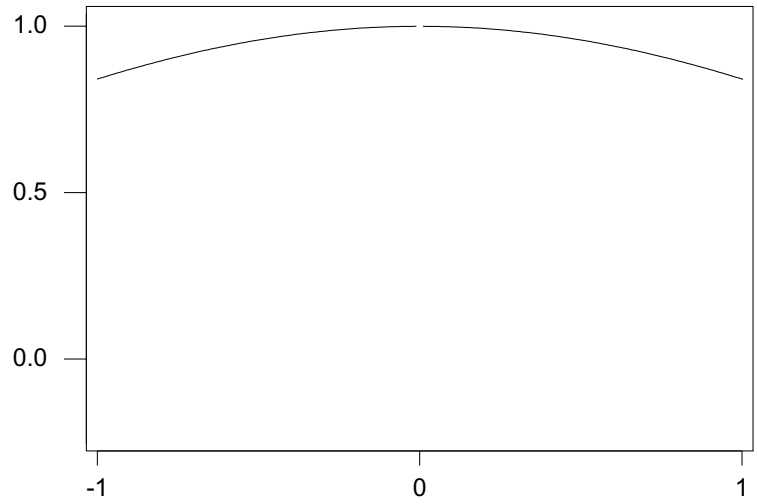


Figure A.3 Graph of $f(x) = \frac{\sin(x)}{x}$ on $A = (-1, 0) \cup (0, 1)$. Note that f is not defined at $x = 0$.

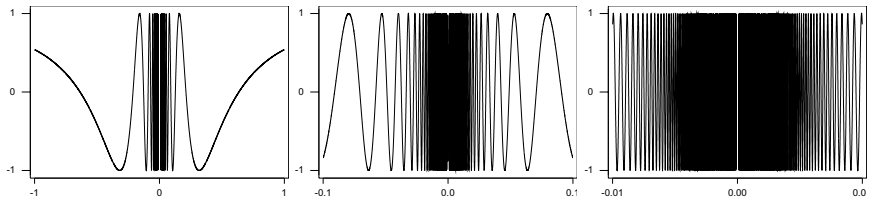


Figure A.4 Graph of $f(x) = cos\left(\frac{1}{x}\right)$ at three scales. Note that f is defined at all real numbers except for $x = 0$.

Theorem 1 *Limit Theorems:*
Let $f(x)$ and $g(x)$ be functions that each have limit at a, and let k_1 and k_2 be scalars.

1. *Limit of a sum (difference) of functions*

$$\lim_{x \to a}[k_1 \times f(x) \pm k_2 \times g(x)] = k_1 \times \lim_{x \to a} f(x) \pm k_2 \times \lim_{x \to a} g(x)\,.$$

2. *Limit of a product of functions*

$$\lim_{x \to a}[f(x) \times g(x)] = \lim_{x \to a} f(x) \times \lim_{x \to a} g(x)\,.$$

3. Limit of a quotient of functions

$$\lim_{x \to a} \left[\frac{f(x)}{g(x)} \right] = \left[\frac{\lim_{x \to a} f(x)}{\lim_{x \to a} g(x)} \right] .$$

4. Limit of a power of a function

$$\lim_{x \to a} [f^n(x)] = [\lim_{x \to a} f(x)]^n .$$

Let $g(x)$ be a function that has limit at a equal to b, and let $f(x)$ be a function that has a limit at b. Let $w(x) = f(g(x))$ be a composite function.

5. Limit of a composite function

$$\lim_{x \to a} w(x) = \lim_{x \to a} f(g(x) = f(\lim_{x \to a} g(x) = f(g(b)) .$$

CONTINUOUS FUNCTIONS

A function $f(x)$ is *continuous* at point a if and only if

$$\lim_{x \to a} f(x) = f(a).$$

This says three things. First, the function has a limit at $x = a$. Second, a is in the domain of the function, so $f(a)$ is defined. Third, the limit of the function at $x = a$ is equal to the value of the function at $x = a$. If we want $f(x)$ to be some specified closeness to $f(a)$, we can find a degree of closeness so that for all x within that degree of closeness to a, $f(x)$ is within the specified closeness to $f(a)$.

A function that is continuous at all values in an interval is said to be continuous over the interval. Sometimes a continuous function is said to be a function that "can be graphed over the interval without lifting the pencil." Strictly speaking, this is not true for all continuous functions. However, it is true for all functions with formulas made from polynomial, exponential, or logarithmic terms.

Theorem 2 *Let $f(x)$ and $g(x)$ be continuous functions, and let k_1 and k_2 be scalars. Then:*

1. *A linear function of continuous functions*

$$s(x) = k_1 \times f(x) + k_2 \times g(x)\,,$$

2. *A product of continuous functions*

$$u(x) = f(x) \times g(x)\,,$$

3. *A quotient of continuous functions*

$$v(x) = \frac{f(x)}{g(x)}\,,$$

4. *And a composite function of continuous functions*

$$w(x) = f(g(x))\,,$$

are all continuous functions on their range of definition.

Minima and Maxima of Continuous Functions

One of the main achievements of calculus is that it gives us a method for finding where a continuous function will achieve minimum and/or maximum values.

Suppose $f(x)$ is a continuous function defined on a continuous domain A. The function achieves a local maximum at the point $x = c$ if and only if $f(x) \leq f(c)$ for *all* points $x \in A$ that are sufficiently close to c. Then $f(c)$ is called a local maximum of the function. The largest local maximum of a function in the domain A is called the global maximum of the function.

Similarly the function achieves a local minimum at point $x = c$ if and only if $f(x) \geq f(c)$ for *all* points $x \in A$ that are sufficiently close to c, and $f(c)$ is called a local minimum of the function. The smallest local minimum of a function in the domain A is called the global minimum of the function.

A continuous function defined on a domain A that is a closed interval $[a, b]$, always achieves a global maximum (and minimum). It can occur at either one of the endpoints $x = a$ or $x = b$, or an interior point $c \in (a, b)$. For example, the function $f(x) = x^4 \times (1 - x)^6$ defined on $A = [0, 1]$ achieves a global maximum at $x = \frac{4}{6}$ and a global minimum at $x = 0$ and $x = 1$ as can be seen in Figure A.1.

A continuous function defined on a domain A that is an open interval (a, b) may or may not achieve either a global maximum or minimum. For example, the function $f(x) = \frac{1}{x^{1/2} \times (x-1)^{1/2}}$ defined on the open interval $(0, 1)$ achieves a global minimum at $x = .5$, but it does not achieve a global maximum as can be seen from Figure A.2.

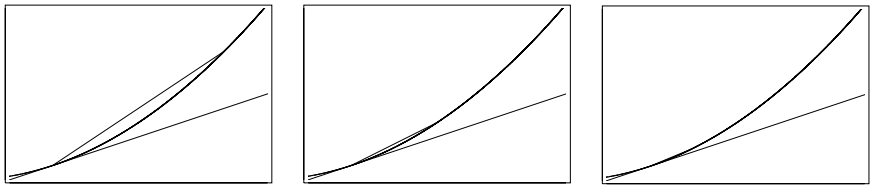


Figure A.5 The derivative at a point is the slope of the tangent to the curve at that point.

DIFFERENTIATION

The first important use of the concept of a limit is finding the derivative of a continuous function. The process of finding the derivative is known as *differentiation*, and it is extremely useful in finding values of x where the function takes a minimum or maximum.

We assume that $f(x)$ is a continuous function whose domain is an interval of the real line. The derivative of the function at $x = c$, a point in the interval is

$$f'(c) = \lim_{h \to 0} \left(\frac{f(c+h) - f(c)}{h} \right)$$

if this limit exists. When the derivative exists at $x = c$, we say the function $f(x)$ is *differentiable* at $x = c$. If this limit does not exist, the function $f(x)$ does not have a derivative at $x = c$. The limit is not easily evaluated, as plugging in $h = 0$ leaves the quotient $\frac{0}{0}$ which is undefined. We also use the notation for the derivative at point c

$$f'(c) = \frac{d}{dx} f(x) \Big|_{x=c} .$$

We note that the derivative at point $x = c$ is the slope of the curve $y = f(x)$ evaluated at $x = c$. It gives the "instantaneous rate of change" in the curve at $x = c$. This is shown in Figure A.5, where $f(x)$, the line joining the point $(c, f(c))$ and point $(c + h, f(c + h))$ for decreasing values of h and its tangent at c are graphed.

The Derivative Function

When the function $f(x)$ has a derivative at all points in an interval, the function

$$f'(x) = \lim_{h \to 0} \left(\frac{f(x+h) - f(x)}{h} \right)$$

is called the *derivative function*. In this case we say that $f(x)$ is a *differentiable function*. The derivative function is sometimes denoted $\frac{dy}{dx}$. The derivatives of some elementary functions are given in the following table:

$f(x)$	$f'(x)$
$a \times x$	a
x^b	$b \times x^{b-1}$
e^x	e^x
$\log_e(x)$	$\frac{1}{x}$
$\sin(x)$	$\cos(x)$
$\cos(x)$	$-\sin(x)$
$\tan(x)$	$-\sec^2(x)$

The derivatives of more complicated functions can be found from these using the following theorems:

Theorem 3 *Let $f(x)$ and g be differentiable functions on an interval, and let k_1 and k_2 be constants.*

1. *The derivative of a constant times a function is the constant times the derivative of the function. Let $h(x) = k_1 \times f(x)$. Then $h(x)$ is also a differentiable function on the interval, and*

$$h'(x) = k_1 \times f'(x)\,.$$

2. *The sum (difference) rule*
   Let $s(x) = k_1 \times f(x) \pm k_2 \times g(x)$. Then $s(x)$ is also a differentiable function on the interval, and

$$s'(x) = k_1 \times f'(x) \pm k_2 \times g'(x)\,.$$

3. *The product rule.*
   Let $u(x) = f(x) \times g(x)$. Then $u(x)$ is a differentiable function, and

$$u'(x) = f(x) \times g'(x) + f'(x) \times g(x)\,.$$

4. *The quotient rule.*
   Let $v(x) = \frac{f(x)}{g(x)}$. Then $v(x)$ is also a differentiable function on the interval, and

$$v'(x) = \frac{g(x) \times f'(x) - f(x) \times g'(x)}{(g(x))^2}\,.$$

Theorem 4 *The chain rule.*
Let $f(x)$ and $g(x)$ be differentiable functions (defined over appropriate intervals) and let $w(x) = f(g(x))$. Then $w(x)$ is a differentiable function and

$$w'(x) = f'(g(x)) \times g'(x)\,.$$

Higher Derivatives

The second derivative of a differentiable function $f(x)$ at a point $x = c$ is the derivative of the derivative function $f'(x)$ at the point. The second derivative is given by

$$f''(c) = \lim_{h \to 0} \left(\frac{f'(c+h) - f'(c)}{h} \right)$$

if it exists. If the second derivative exists for all points x in an interval, then $f''(x)$ is the second derivative function over the interval. Other notation for the second derivative at point c and for the second derivative function are

$$f''(c) = f^{(2)}(c) = \frac{d}{dx} f'(x) \Big|_{x=c} \quad \text{and} \quad f^{(2)}(x) = \frac{d_2}{dx^2} f(x) \, .$$

Similarly the k^{th} derivative is the derivative of the $k - 1^{th}$ derivative function

$$f^{(k)}(c) = \lim_{h \to 0} \left(\frac{f^{(k-1)}(c+h) - f^{(k-1)}(c)}{h} \right)$$

if it exists.

Critical Points

For a function $f(x)$ that is differentiable over an open interval (a, b), the derivative function $f'(x)$ is the slope of the curve $y = f(x)$ at each x-value in the interval. This gives a method of finding where the minimum and maximum values of the function occur. The function will achieve its minimum and maximum at points where the derivative equals 0. When $x = c$ is a solution of the equation

$$f'(x) = 0 \, ,$$

c is called a critical point of the function $f(x)$. The critical points may lead to local maximum or minimum, global maximum or minimum, or they may be points of inflection. A point of inflection is where the function changes from being concave to convex, or vice versa.

Theorem 5 *First derivative test: If $f(x)$ is a continuous differentiable function over an interval (a, b) having derivative function $f'(x)$ which is defined on the same interval. Suppose c is a critical point of the function. By definition, $f'(c) = 0$.*

1. *The function achieves a unique local maximum at $x = c$ if, for all points x that are sufficiently close to c*
   when $x < c$ then $f'(x) > 0$ and
   when $x > c$ then $f'(x) < 0$.
2. *Similarly the function achieves a unique local minimum at $x = c$ if, for all points x that are sufficiently close to c*
   when $x < c$ then $f'(x) < 0$ and
   when $x > c$ then $f'(x) > 0$.

3. *The function has a point of inflection at critical point $x = c$ if, for all points x that are sufficiently close to c, either*
   when $x < c$ then $f'(x) < 0$ and
   when $x > c$ then $f'(x) < 0$,
   or
   when $x < c$ then $f'(x) > 0$ and
   when $x > c$ then $f'(x) > 0$.
   At a point of inflection, the function either stops increasing, and then resumes increasing, or it stops decreasing, and then resumes decreasing.

For example, the function $f(x) = x^3$, and its derivative $f'(x) = 3 \times x^2$ are shown in Figure A.6. We see that the derivative function $f'(x) = 3x^2$ is positive for $x < 0$, so the function $f(x) = x^3$ is increasing for $x < 0$. The derivative function is positive for $x > 0$ so the function is also increasing for $x > 0$. However at $x = 0$, the derivative function equals 0, so the original function is not increasing at $x = 0$. Thus the function $f(x) = x^3$ has a point of inflection at $x = 0$.

Theorem 6 *Second derivative test: If $f(x)$ is a continuous differentiable function over an interval (a, b) having first derivative function $f'(x)$ and second derivative function $f^{(2)}(x)$ both defined on the same interval. Suppose c is a critical point of the function. By definition, $f'(c) = 0$.*

1. *The function achieves a maximum at $x = c$ if $f^{(2)}(c) < 0$*
2. *The function achieves a minimum at $x = c$ if $f^{(2)}(c) > 0$*

INTEGRATION

The second main use of calculus is finding the area under a curve using *integration*. It turns out that *integration* is the inverse of *differentiation*. Suppose $f(x)$ is a function defined on an interval $[a, b]$. Let the function $F(x)$ be an *antiderivative* of $f(x)$. That means the derivative function $F'(x) = f(x)$. Note that the antiderivative of $f(x)$ is not unique. The function $F(x) + c$ will also be an antiderivative of $f(x)$. The antiderivative is also called the *indefinite integral*.

The Definite Integral: Finding the Area under a Curve

Suppose we have a nonnegative[1] continuous function $f(x)$ defined on a closed interval $[a, b]$. $f(x) \geq 0$ for all $x \in [a, b]$. Suppose we partition the the interval $[a, b]$ using the partition $x_0, x_1, \ldots, x_n$, where $x_0 = a$ and $x_n = b$ and $x_i < x_{i+1}$. Note that the partition does not have to have equal length intervals. Let the minimum and maximum value of $f(x)$ in each interval be

$$l_i = \sup_{x \in [x_{i-1}, x_i]} f(x) \quad \text{and} \quad m_i = \inf_{x \in [x_{i-1}, x_i]} f(x)$$

[1]The requirement that $f(x)$ be nonnegative is not strictly necessary. However since we are using the definite integral to find the area under probability density functions that are nonnegative, we will impose the condition.

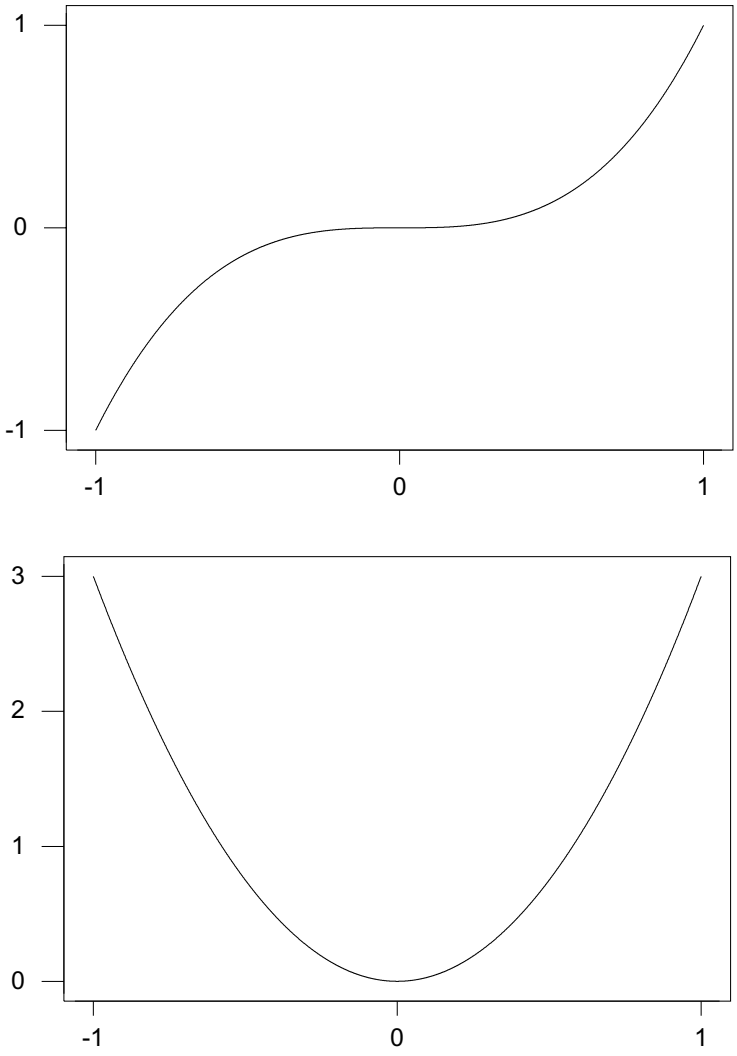


Figure A.6 Graph of $f(x) = x^3$ and its derivative. The derivative function is negative where the original function is increasing, and it is positive where the original function is increasing We see the original function has a point of inflection at $x = 0$.

where $\sup$ is the least upper bound, and $\inf$ is the greatest lower bound. Then the area under the curve $y = f(x)$ between $x = a$ and $x = b$ lies between the lower sum

$$L_{x_0,\ldots,x_n} = \sum_{i=1}^{n} l_i \times (x_i - x_{i-1})$$

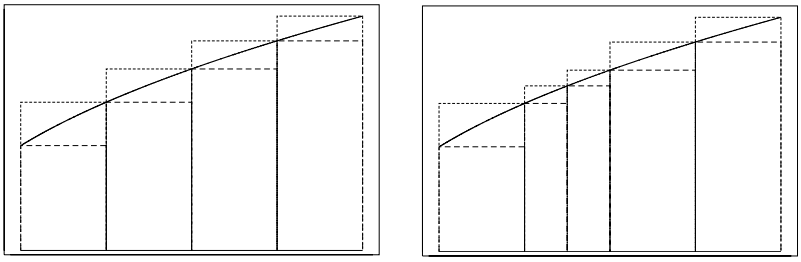


Figure A.7 Lower and upper sums over a partition and its refinement. The lower sum has increased and the upper sum has decreased in the refinement. The area under the curve is always between the lower and upper sums.

and the upper sum

$$M_{x_0,\ldots,x_n} = \sum_{i=1}^{n} m_i \times (x_i - x_{i-1})$$

We can refine the partition by adding one more x value to it. Let $x'_1, \ldots, x'_{n+1}$ be a refinement of the partition $x_1, \ldots, x_n$. Then $x'_0 = x_0$, $x'_{n+1} = x_n$, $x'_i = x_i$ for all $i < k$, and $x'_{i+i} = x_i$ for all $i > k$. x_k is the new value added to the partition. In the lower and upper sum, all the bars except for the k^{th} are unchanged. The k^{th} bar has been replaced by two bars in the refinement. Clearly,

$$M_{x'_0,\ldots,x'_{n+1}} \leq M_{x_0,\ldots,x_n}$$

and

$$L_{x'_0,\ldots,x'_{n+1}} \geq L_{x_0,\ldots,x_n} .$$

The lower and upper sums for a partition and its refinement are shown in Figure A.7. We see that refining a partition must make tighter bounds on the area under the curve.

Next we will show that for any continuous function defined on a closed interval $[a, b]$,we can find a partition $x_0, \ldots, x_n$ for some n that will make the difference between the upper sum and the lower sum as close to zero as we wish. Suppose $\epsilon > 0$ is the number we want the difference to be less than. We draw lines $\delta = \frac{\epsilon}{[b-a]}$ apart parallel to the horizontal (x) axis. (Since the function is defined on the closed interval, its maximum and minimum are both finite.) Thus a finite number of the horizontal lines will intercept the curve $y = f(x)$ over the interval $[a, b]$. Where one of the lines intercepts the curve, draw a vertical line down to the horizontal axis. The x values where these vertical lines hit the horizontal axis are the points for our partition. For example, the function $f(x) = 1 + \sqrt{4 - x^2}$ is defined on the interval $[0, 2]$. The difference between the upper sum and the lower sum for the partition for that ϵ is given by

$$\begin{aligned} M_{x_0,\ldots,x_n} - L_{x_0,\ldots,x_n} &= \delta \times [(x_1 - x_0) + (x_2 - x_1) + \ldots + (x_n - x_{n-1})] \\ &= \delta \times [b - a] \\ &= \epsilon. \end{aligned}$$

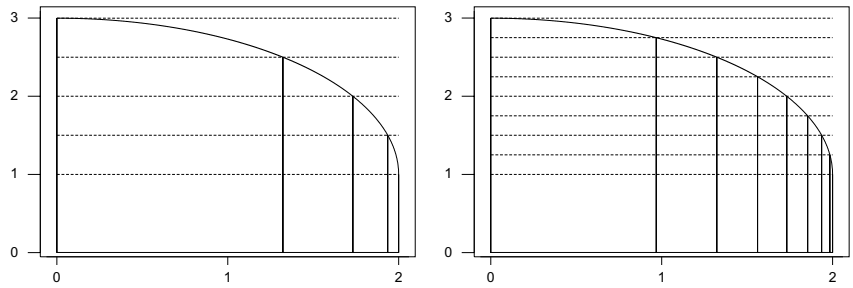


Figure A.8 The partition induced for the function $f(x) = 1 + \sqrt{4 - x^2}$ where $\epsilon_1 = 1$ and its refinement where $\epsilon_2 = \frac{1}{2}$.

We can make this difference as small as we want to by choosing $\epsilon > 0$ small enough.

Let $\epsilon_k = \frac{1}{k}$ for $k = 1, \ldots, \infty$. This gives us a sequence of partitions such that $\lim_{k\to\infty} \epsilon_k = 0$. Hence

$$\lim_{k\to\infty} M_{x_0,\ldots,x_{n_k}} - L_{x_0,\ldots,x_{n_k}} = 0\,.$$

The partitions for ϵ_1 and ϵ_2 are shown in Figure A.8. Note that $\delta_k = \frac{1}{2k}$.

That means that the area under the curve is the least upper bound for the lower sum, and the greatest lower bound for the upper sum. We call it the definite integral and denote it

$$\int_a^b f(x)dx\,.$$

Note the variable x in the formula above is a dummy variable:

$$\int_a^b f(x)dx = \int_a^b f(y)dy\,.$$

Basic Properties of Definite Integrals

Theorem 7 *Let $f(x)$ and $g(x)$ be functions defined on the interval $[a, b]$, and let c be a constant. Then the following properties hold.*

1. *The definite integral of a constant times a function is the constant times the definite integral of the function:*

$$\int_a^b cf(x)dx = c\int_a^b f(x)dx\,.$$

2. *The definite integral of a sum of two functions is a sum of the definite integrals of the two functions:*

$$\int_a^b (f(x) + g(x))dx = \int_a^b f(x)dx + \int_a^b g(x)dx\,.$$

Fundamental Theorem of Calculus

The methods of finding extreme values by differentiation and finding area under a curve by integration were known before the time of Newton and Liebniz. Newton and Liebniz independently discovered the fundamental theorem of calculus that connects differentiation and integration. Because each was unaware of the others work, they are both credited with the discovery of the calculus.

Theorem 8 *Fundamental theorem of calculus. Let $f(x)$ be a continuous function defined on a closed interval. Then:*

1. *The function has antiderivative in the interval.*
2. *If a and b are two numbers in the closed interval such that $a < b$, and $F(x)$ is any antiderivative function of $f(x)$, then*

$$\int_a^b f(x)dx = F(b) - F(a)\,.$$

Proof:
For $x \in (a, b)$, define the function

$$I(x) = \int_a^x f(x)dx\,.$$

This function shows the area under the curve $y = f(x)$ between a and x. Note that the area under the curve is additive over an extended region from a to $x + h$:

$$\int_a^{x+h} f(x)dx = \int_a^x f(x)dx + \int_x^{x+h} f(x)dx\,.$$

By definition, the derivative of the function $I(x)$ is

$$I'(x) = \lim_{h\to 0} \frac{I(x+h) - I(x)}{h} = \lim_{h\to 0} \frac{\int_x^{x+h} f(x)dx}{h}\,.$$

In the limit as h approaches 0,

$$\lim_{h\to 0} f(x') = f(x)$$

for all values $x' \in [x, x + h)$. Thus

$$I'(x) = \lim_{h\to 0} \frac{h \times f(x)}{h} = f(x)\,.$$

In other words, $I(x)$ is an antiderivative of $f(x)$. Suppose $F(x)$ is any other antiderivative of $f(x)$. Then

$$F(x) = I(x) + c$$

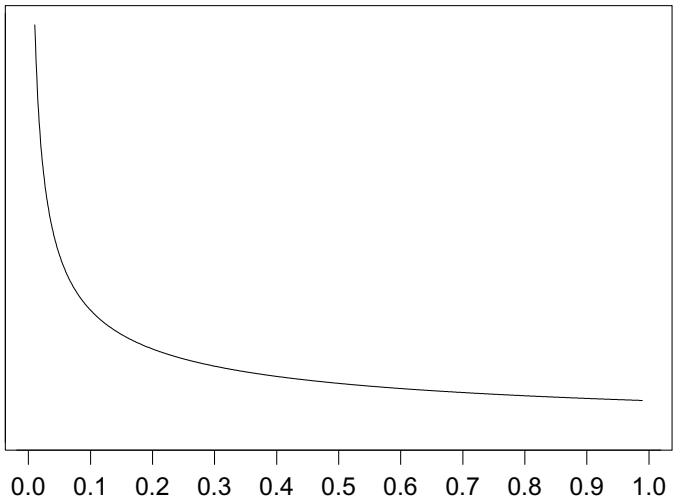


Figure A.9 The function $f(x) = x^{-1/2}$.

for some constant c. Thus $F(b) - F(a) = I(b) - I(a) = \int_a^b f(x)dx$*, and the theorem is proved.*

For example, suppose $f(x) = e^{-2x}$ for $x \geq 0$. Then $F(x) = -\frac{1}{2} \times e^{-2x}$ is an antiderivative of $f(x)$. The area under the curve between 1 and 4 is given by

$$\int_1^4 f(x)dx = F(4) - F(1) = -\frac{1}{2} \times e^{-2\times 4} + \frac{1}{2} \times e^{-2\times 1}\,.$$

Definite Integral of a Function $f(x)$ Defined on an Open Interval

Let $f(x)$ be a function defined on the open interval (a, b). In this case, the antiderivative $F(x)$ is not defined at a and b. We define

$$F(a) = \lim_{x\to a} F(x) \quad \text{and} \quad F(b) = \lim_{x\to b} F(x)$$

provided those limits exist. Then we define the definite integral with the same formula as before

$$\int_a^b f(x) = F(b) - F(a)$$

For example, let $f(x) = x^{-1/2}$. This function is defined over the half-open interval $(0, 1]$. It is not defined over the closed interval $[0, 1]$ because it is not defined at the endpoint $x = 0$. This curve is shown in Figure A.9. We see the curve has a

vertical asymptote at $x = 0$. We will define

$$\begin{aligned} F(0) &= \lim_{x \to 0} F(x) \\ &= \lim_{x \to 0} 2x^{1/2} \\ &= 0\,. \end{aligned}$$

Then

$$\int_0^1 x^{-1/2} = 2x^{1/2}\Big|_0^1 = 2\,.$$

Theorem 9 *Integration by parts. Let $F(x)$ and $G(x)$ be differentiable functions defined on an interval $[a, b]$. Then*

$$\int_a^b F'(x) \times G(x)dx = F(x) \times G(x)|_a^b - \int_a^b F(x) \times G'(x)dx\,.$$

Proof: Integration by parts is the inverse of finding the derivative of the product $F(x) \times G(x)$:

$$\frac{d}{dx}[F(x) \times G(x)] = F(x) \times G'(x) + F(x) \times G'(x)\,.$$

Integrating both sides, we see that

$$F(b) \times G(b) - F(a) \times G(a) = \int_a^b F(x) \times G'(x)dx + \int_a^b F'(x) \times G(x)dx\,.$$

Theorem 10 *Change of variable formula. Let $x = g(y)$ be a differentiable function on the interval $[a, b]$. Then*

$$\int_a^b f(g(y))g'(y)dy = \int_{g(a)}^{g(b)} f(y)dy$$

The change of variable formula is the inverse of the chain rule for differentiation. The derivative of the function of a function $F(g(y))$ is

$$\frac{d}{dx}[F(g(y)] = F'(g(y)) \times g'(y)\,.$$

Integrating both sides from $y = a$ to $y = b$ gives

$$F(g(b)) - F(g(a)) = \int_a^b F'(g(y)) \times g'(y)dy\,.$$

The left-hand-side equals $\int_{ga)}^{g(b)} F'(y)dy$. Let $f(x) = F'(x)$, and the theorem is proved.

MULTIVARIATE CALCULUS

Partial Derivatives

In this section we consider the calculus of two or more variables. Suppose we have a function of two variables $f(x, y)$. The function is continuous at the point (a, b) if and only if

$$\lim_{(x,y)\to(a,b)} f(x, y) = f(a, b)\,.$$

The first *partial derivatives* at the point (a, b) are defined to be

$$\left.\frac{\partial f(x, y)}{\partial x}\right|_{(a,b)} = \lim_{h\to 0} \frac{f(a + h, b) - f(a, b)}{h}$$

and

$$\left.\frac{\partial f(x, y)}{\partial y}\right|_{(a,b)} = \lim_{h\to 0} \frac{f(a, b + h) - f(a, b)}{h}$$

provided these limits exist. In practice, the first partial derivative in the x-direction is found by treating y as a constant and differentiating the function with respect to x, and vice versa, to find the first partial derivative in the y-direction.

If the function $f(x, y)$ has first partial derivatives for all points (x, y) in a continuous two-dimensional region, then the first partial derivative function with respect to x is the function that has value at point (x, y) equal to the partial derivative of $f(x, y)$ with respect to x at that point. It is denoted

$$f_x(x, y) = \left.\frac{\partial f(x, y)}{\partial x}\right|_{(x,y)}\,.$$

The first partial derivative function with respect to y is defined similarly. The first derivative functions $f_x(x, y)$ and $f_y(x, y)$ give the instantaneous rate of change of the function in the x-direction and y-direction, respectively.

The second *partial derivatives* at the point (a, b) are defined to be

$$\left.\frac{\partial^2 f(x, y)}{\partial x^2}\right|_{(a,b)} = \lim_{h\to 0} \frac{f_x(x + h, y) - f_x(x, y)}{h}$$

and

$$\left.\frac{\partial^2 f(x, y)}{\partial y^2}\right|_{(a,b)} = \lim_{h\to 0} \frac{f_y(x, y + h) - f_y(x, y)}{h}\,.$$

The second *cross partial derivatives* at (a, b) are

$$\left.\frac{\partial^2 f(x, y)}{\partial x \partial y}\right|_{(a,b)} = \lim_{h\to 0} \frac{f_y(x + h, y) - f_y(x, y)}{h}$$

and

$$\left.\frac{\partial^2 f(x, y)}{\partial y \partial x}\right|_{(a,b)} = \lim_{h\to 0} \frac{f_x(x, y + h) - f_x(x, y)}{h}\,.$$

For all the functions that we consider, the *cross partial derivatives* are equal, so it doesn't matter which order we differentiate.

If the function $f(x, y)$ has second partial derivatives (including cross partial derivatives) for all points (x, y) in a continuous two-dimensional region, then the second partial derivative function with respect to x is the function that has value at point (x, y) equal to the second partial derivative of $f(x, y)$ with respect to x at that point. It is denoted

$$f_{xx}(x, y) = \left.\frac{\partial f_x(x, y)}{\partial x}\right|_{(x,y)} .$$

The second partial derivative function with respect to y is defined similarly. The second cross partial derivative functions are

$$f_{xy}(x, y) = \left.\frac{\partial f_x(x, y)}{\partial y}\right|_{(x,y)}$$

and

$$f_{yx}(x, y) = \left.\frac{\partial f_y(x, y)}{\partial x}\right|_{(x,y)} .$$

The two cross partial derivative functions are equal.

Partial derivatives of functions having more than 2 variables are defined in a similar manner.

Finding Minima and Maxima of a Multivariate Function

A univariate functions with a continuous derivative achieves minimum or maximum at an interior point x only at points where the derivative function $f'(x) = 0$. However, not all such points were minimum or maximum. We had to check either the first derivative test, or the second derivative test to see whether the critical point was minimum, maximum, or point of inflection.

The situation is more complicated in two dimensions. Suppose a continuous differentiable function $f(x, y)$ is defined on a two dimensional rectangle. It is not enough that both $f_x(x, y) = 0$ and $f_y(x, y) = 0$.

The directional derivative of the function $f(x, y)$ in direction θ at a point measures the rate of change of the function in the direction of the line through the point that has angle θ with the positive x-axis. It is given by

$$D_\theta f(x, y) = f_x(x, y)\cos(\theta) + f_y(x, y)\sin(\theta) .$$

The function achieves a maximum or minimum value at points (x, y) where $D_\theta f(x, y) = 0$ for all θ.

Multiple Integrals

Let $f(x, y) > 0$ be a nonnegative function defined over a closed a rectangle $a_1 \le x \le b_1$ and $a_2 \le y \le b_2$. Let $x_0, \ldots, x_n$ partition the interval $[a_1, b_1]$, and let $y_1, \ldots, y_m$

partition the interval a_2, b_2. Together these partition the rectangle into $j = m \times n$ rectangles. The volume under the surface $f(x, y)$ over the rectangle A is between the upper sum

$$U = \sum_{j=1}^{mn} f(t_j, u_j)$$

and the lower sum

$$U = \sum_{j=1}^{mn} f(v_j, w_j)\,,$$

where (t_j, u_j) is the point where the function is maximized in the j^{th} rectangle, and (v_j, w_j) is the point where the function is minimized in the j^{th} rectangle. Refining the partition always lowers the upper sum and raises the lower sum. We can always find a partition that makes the upper sum arbitrarily close to the lower sum. Hence the total volume under the surface denoted

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y) dx dy$$

is the least upper bound of the lower sum and the greatest lower bound of the upper sum.

B

Use of Statistical Tables

BINOMIAL DISTRIBUTION

Table B.1 contains values of the *binomial* (n, π) probability distribution for $n = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15,$ and 20 and for $\pi = .05, .10, \ldots, .95$. Given the parameter π, the *binomial* probability is obtained by the formula

$$P(Y = y|\pi) = \left(\begin{array}{c} n \\ \pi \end{array} \right) n\pi^y(1-\pi)^{n-y}. \tag{B.1}$$

When $\pi \leq .5$, use the π value along the top row to find the correct column of probabilities. Go down to the correct n. The probabilities correspond to the y values found in the left-hand column. For example, to find $P(Y = 6)$ when Y has the *binomial* $(n = 10, \pi = .3)$ distribution, go down the table to $n = 10$ and find the row $y = 6$ on the *left* side. Look across the top to find the column labelled .30. The value in the table at the intersection of that row and column is $P(Y = 6) = .0368$ in this example.

When $\pi > .5$ use the π value along the bottom row to find the correct column of probabilities. Go down to the correct n. The probabilities correspond to the y values found in the right hand column. For example to find $P(Y = 3)$ when y has the *binomial* $(n = 8, \pi = .65)$ distribution, go down the table to $n = 8$ and find the row $y = 3$ on the *right* side. Look across the bottom to find the column labelled .65. The

*

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

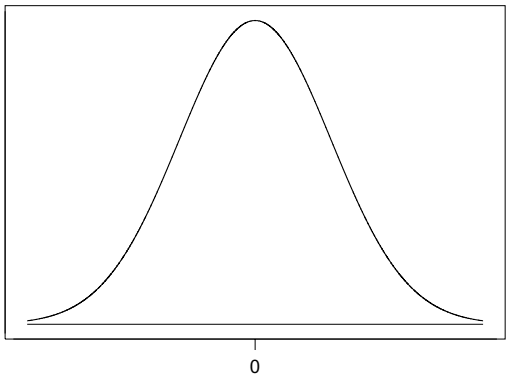


Figure B.1 Standard normal density.

value in the table at the intersection of that row and column is $P(Y = 3) = .0808$ in this example.

STANDARD NORMAL DISTRIBUTION

This section contains two tables. Table B.2 contains the area under the standard normal density. Table B.3 contains the ordinates (height) of the standard normal density. The standard normal density has mean equal to 0 and variance equal to 1. Its density is given by the formula

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \tag{B.2}$$

We see that the standard normal density is symmetric about 0. The graph of the standard normal density is shown in Figure B.1.

Area Under Standard Normal Density

Table B.2 tabulates the area under the standard normal density function between 0 and z for nonnegative values of z from 0.0 to 3.99 in steps of $.01$. We read down the z column until we come to the value that has the correct *units* and tenths digits of z. This is the correct row. We look across the top row to find the *hundredth* digit of z. This is the correct column. The tabulated value at the intersection of the correct row and correct column is $P(0 \leq Z \leq z)$ where Z has the *normal* $(0, 1)$ distribution. For example, to find $P(0 \leq Z \leq 1.23)$ we go down the z column to 1.2 for the correct row and across top to 3 for correct column. We find the tabulated value at the intersection of this row and column. For this example $P(0 \leq Z \leq 1.23) = .3907$.

Because the standard normal density is symmetric about 0,

$$P(-z \le Z \le 0) = P(0 \le Z \le z)\,.$$

Also, since it is a density function, the total area underneath it equals 1.0000, so the total area to the right of 0 must equal .5000. We can proceed to find

$$P(Z > z) = .5000 - P(Z \le z)\,.$$

Finding Any Normal Probability

We can standardize any normal random variable to a standard normal random variable having mean 0 and variance 1. For instance, if W is a normal random variable having mean m and variance s^2, we standardize by subtracting the mean and dividing by the standard deviation.

$$Z = \frac{W - m}{s}\,.$$

This lets us find any normal probability by using the standard normal tables.

Example 25 *Suppose W has the normal distribution with mean 120 and variance 225. (The standard deviation of W is 15.) Suppose we wanted to find the probability*

$$P(W \le 129)\,.$$

We can subtract the mean from both sides of an inequality without changing the inequality:

$$P(W - 120 \le 129 - 120)\,.$$

We can divide both sides of an inequality by the standard deviation (which is positive) without changing the inequality:

$$P\left(\frac{W - 120}{15} \le \frac{9}{15}\right)\,.$$

On the left-hand side we have the standard normal Z, and on the right-hand side we have the number .60. Therefore

$$P(W \le 129) = P(Z \le .60) = .5000 + .2258 = .7258\,.$$

Ordinates of the Standard Normal Density

Figure B.3 shows the ordinate of the standard normal table at z. We see the ordinate is the height of the curve at z. Table B.3 contains the ordinates of the standard normal density for nonnegative z values from 0.00 to 3.99 in steps of .01. Since the standard normal density is symmetric about 0, $f(-z) = f(z)$, we can find ordinates of negative z values.

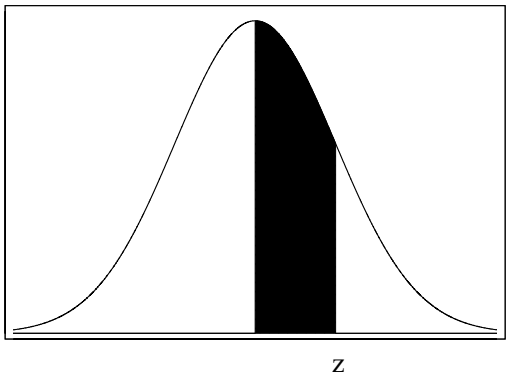


Figure B.2 Shaded area under standard normal density. These values are shown in Table B.2.

This table is used to find values of the likelihood when we have a discrete prior distribution for μ. We go down the z column until we find the value that has the *units* and *tenths* digits. This gives us the correct row. We go across the top until we find the *hundredth* digit. This gives us the correct column. The value at the intersection of this row and column is the ordinate of the standard normal density at the value z. For instance, if we want to find the height of the standard normal density at $z = 1.23$ we go down z column to 1.2 to find the correct row, and across the top to 3 to find the correct column. The ordinate of the standard normal at $z = 1.23$ is equal to .1872. (Note: You can verify this is correct by plugging $z = 1.23$ into Equation B.2.)

Example 26 *Suppose the distribution of* Y *given* μ *is* normal $(\mu, \sigma^2 = 1)$. *Also suppose there are 4 possible values of* μ. *They are 3,4,5,and 6. We observe y=5.6. We calculate*

$$z_i = \left(\frac{5.6 - \mu_i}{1}\right) .$$

The likelihood is found by looking up the ordinates of the normal distribution for the z_i *values. We can put them in the following table.*

μ_i	z_i	*Likelihood*
3	*2.60*	*.136*
4	*1.6*	*.1109*
5	*.6*	*.3332*
6	*-.4*	*.3683*

Table B.1 Binomial probability table

n	y	π .05	.10	.15	.20	.25	.30	.35	.40	.45	.50	y
2	0	.9025	.81	.7225	.64	.5625	.49	.4225	.36	.3025	.25	2
	1	.0950	.18	.2550	.32	.3750	.42	.4550	.48	.4950	.50	1
	2	.0025	.01	.0225	.04	.0625	.09	.1225	.16	.2025	.25	0
3	0	.8574	.729	.6141	.512	.4219	.343	.2746	.216	.1664	.125	3
	1	.1354	.243	.3251	.384	.4219	.441	.4436	.432	.4084	.375	2
	2	.0071	.027	.0574	.096	.1406	.189	.2389	.288	.3341	.375	1
	3	.0001	.001	.0034	.008	.0156	.027	.0429	.064	.0911	.125	0
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625	4
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500	3
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750	2
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500	1
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625	0
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313	5
	1	.2036	.3281	.3915	.4096	.3955	.3601	.3124	.2592	.2059	.1563	4
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125	3
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125	2
	4	.0000	.0005	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1563	1
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0313	0
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156	6
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0937	5
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344	4
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125	3
	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344	2
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0937	1
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156	0
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078	7
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547	6
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641	5
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734	4
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734	3
	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641	2
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547	1
	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078	0
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039	8
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313	7
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094	6
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188	5
	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734	4
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188	3
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094	2
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0313	1
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039	0
		.95	.90	.85	.80	.75 π	.70	.65	.60	.55	.50	y

Table B.1 (Continued)

n	y	π .05	.10	.15	.20	.25	.30	.35	.40	.45	.50	y
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020	9
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176	8
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703	7
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641	6
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461	5
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461	4
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641	3
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703	2
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176	1
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020	0
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010	10
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098	9
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439	8
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172	7
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051	6
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461	5
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051	4
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172	3
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439	2
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098	1
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	0
11	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054	10
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269	9
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806	8
	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611	7
	5	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256	6
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256	5
	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611	4
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806	3
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269	2
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054	1
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005	0
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002	12
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029	11
	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161	10
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537	9
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208	8
	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934	7
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256	6
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934	5
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208	4
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537	3
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161	2
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029	1
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	0
		.95	.90	.85	.80	.75	.70	.65	.60	.55	.50	
						π						y

Table B.1 (Continued)

n	y	π .05	.10	.15	.20	.25	.30	.35	.40	.45	.50	
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000	15
	1	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005	14
	2	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032	13
	3	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139	12
	4	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417	11
	5	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916	10
	6	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527	9
	7	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964	8
	8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964	7
	9	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527	6
	10	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916	5
	11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417	4
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139	3
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032	2
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	1
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	0
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000	20
	1	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000	19
	2	.1887	.2852	.2293	.1369	.0669	.0278	.0100	.0031	.0008	.0002	18
	3	.0596	.1901	.2428	.2054	.1339	.0716	.0323	.0123	.0040	.0011	17
	4	.0133	.0898	.1821	.2182	.1897	.1304	.0738	.0350	.0139	.0046	16
	5	.0022	.0319	.1028	.1746	.2023	.1789	.1272	.0746	.0365	.0148	15
	6	.0003	.0089	.0454	.1091	.1686	.1916	.1712	.1244	.0746	.0370	14
	7	.0000	.0020	.0160	.0545	.1124	.1643	.1844	.1659	.1221	.0739	13
	8	.0000	.0004	.0046	.0222	.0609	.1144	.1614	.1797	.1623	.1201	12
	9	.0000	.0001	.0011	.0074	.0271	.0654	.1158	.1597	.1771	.1602	11
	10	.0000	.0000	.0002	.0020	.0099	.0308	.0686	.1171	.1593	.1762	10
	11	.0000	.0000	.0000	.0005	.0030	.0120	.0336	.0710	.1185	.1602	9
	12	.0000	.0000	.0000	.0001	.0008	.0039	.0136	.0355	.0727	.1201	8
	13	.0000	.0000	.0000	.0000	.0002	.0010	.0045	.0146	.0366	.0739	7
	14	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0049	.0150	.0370	6
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0049	.0148	5
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0046	4
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011	3
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	2
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	1
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	0
		.95	.90	.85	.80	.75	.70	.65	.60	.55	.50	
						π						y

Table B.2 Area under standard normal density

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999

Table B.3 Ordinates of standard normal density

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Table B.4 Critical values of the *Student's t* distribution

degrees of freedom	*upper tail area*							
	.20	.10	.05	.025	.01	.005	.001	.0005
1	1.376	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	1.061	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	.979	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	.920	1.476	2.015	2.571	3.365	4.032	5.893	6.868
6	.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	.846	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	.845	1.290	1.660	1.984	2.364	2.626	3.174	3.390
∞	.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

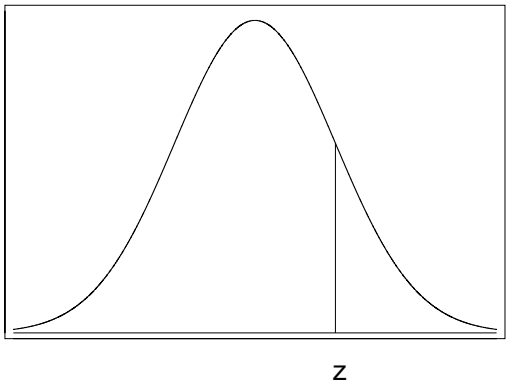


Figure B.3 Ordinates of standard normal density function. These values are shown in Table B.3.

STUDENT'S *t* DISTRIBUTION

Figure B.4 shows the *Student's t* distribution for several different degrees of freedom, along with the standard *normal*$(0, 1)$ distribution. We see the *Student's t* family of distributions are similar to the standard *normal* in that they are symmetric bell shaped curves, however they have more weight in the tails. The heaviness of the tails of the *Student's t* decreases as the degrees of freedom increase[1].

The *Student's t* distribution is used when we use the unbiased estimate of the standard deviation $\hat{\sigma}$ instead of the true unknown standard deviation σ in the standardizing formula

$$z = \frac{y - \mu}{\sigma_y}$$

and y is a normally distributed random variable. We know that z will have the *normal*$(0, 1)$ distribution. The similar formula

$$t = \frac{y - \mu}{\hat{\sigma}_y}$$

will have the *Student's t* distribution with k degrees of freedom. The degrees of freedom k will equal the sample size minus the number of parameters estimated in the equation for $\hat{\sigma}$. For instance, if we are using $\bar{y}$ the sample mean, the estimated standard deviation $\hat{\sigma}_{\bar{y}} = \frac{\hat{\sigma}}{n}$ where

$$\hat{\sigma} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

[1]The *normal*$(0, 1)$ distribution corresponds to the *Student's t* distribution with ∞ degrees of freedom

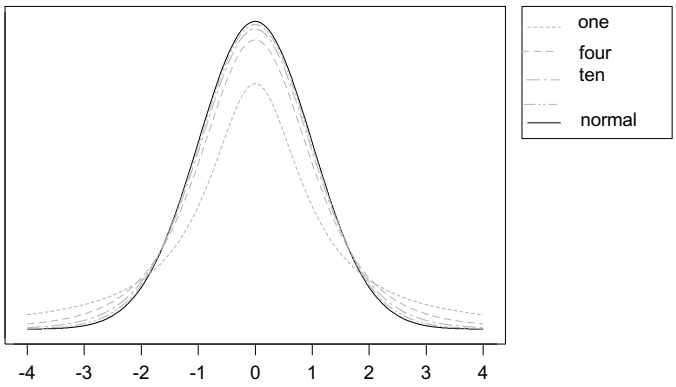


Figure B.4 Student's t densities for selected degrees of freedom together with the standard *normal* $(0, 1)$ density which corresponds to Student's t with ∞ degrees of freedom.

and we observe that to use the above formula we have to first estimate $\bar{y}$. Hence, in the single sample case we will have $k = n - 1$ degrees of freedom.

Table B.4 contains the tail areas for the *Student's t* distribution family. The *degrees of freedom* are down the left column, and the tabulated tail areas are across the rows for the specified tail probabilities.

C

Using the Included Minitab Macros

Minitab macros for performing Bayesian analysis and for doing Monte Carlo simulations are included. The address may be downloaded from the Web page for this text on the site <www.wiley.com>. The Minitab Macros are zipped up in a package called *Bolstad Minitab Macros.zip*. Some Minitab worksheets are also included at that site.

To use the Minitab macros, define a directory named BAYESMAC on your hard disk. The best place is inside the Minitab directory, which on is often called MTBWIN on PC's running Microsoft Windows. For example, on my PC, BAYESMAC is inside MTBWIN which is within program files which is on drive C. The correct path I need to invoke to use these macros is *C:/progra ~ 1/MTBWIN/BAYESM ~1/* (Note that the the filenames are truncated at six characters.) You should also define a directory BAYESMTW for the Minitab worksheets containing the data sets. The best place is also inside the Minitab directory, so you can find it easily.

*

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

Table C.1 Minitab commands for sampling Monte Carlo study

Minitab Commands	Meaning
%<*insert path*>sscsample.mac c1 100;	"data are in c1, N=100"
strata c2 3;	"there are 3 strata stored in c2"
cluster c3 20;	"there are 20 clusters stored in c3"
type 1;	1=simple, 2=stratified, 3=cluster
isize 20;	"sample size n=20"
mcarlo 200;	"Monte Carlo sample size 200"
output c6 c7 c8 c9;	"c6 contains sample means, c7-c9 contain numbers in each strata"

CHAPTER 2: SCIENTIFIC DATA GATHERING

Sampling Methods

We use the *sscsample.mac* to perform a small-scale Monte Carlo study on the efficiency of simple, stratified, and cluster random sampling on the population data contained in *sscsample.mtw*. In the "file" menu pull down "open worksheet" command. When the dialog box opens, find the directory BAYESMTW and type in *sscsample.mtw* in the filename box and click on "open". In the "edit" menu pull down "command line editor" and type the commands from Table C.1 into the command line editor:

Experimental Design

We use the Minitab macro *Xdesign.mac* to perform a small-scale Monte Carlo study, comparing *completely randomized design* and *randomized block design* in their effectiveness for assigning experimental units into treatment groups. Type the commands from Table C.2 into the command line editor.

CHAPTER 6: BAYESIAN INFERENCE FOR DISCRETE RANDOM VARIABLES

Binomial Proportion with Discrete Prior

BinoDP.mac is used to find the posterior when we have *binomial* (n, π) observation, and we have a discrete prior for π. For example, suppose π has the discrete distribution with three possible values, .3, .4, and .5. Suppose the prior distribution is given in Table C.3. and we want to find the posterior distribution after $n = 6$ trials and observing $y = 5$ successes. In the "edit" menu pull down "command line editor" and type the commands from Table C.4 into the command line editor.

Table C.2 Minitab commands for experimental design Monte Carlo study

Minitab Commands	Meaning
let k1=.8	"correlation between other and response variables"
random 80 c1 c2; normal 0 1.	"generate 80 other and response variables in c1 and c2 respectively"
let c2=sqrt(1-k1**2)*c2+k1*c1	"give them correlation k1"
desc c1 c2	"summary statistics"
corr c1 c2	
plot c2*c1	"shows relationship"
%<*insert path*>Xdesign.mac c1 c2;	"other variable in c1, response in c2"
size 20;	"treatment groups of 20 units"
treatments 4;	"4 treatment groups"
mcarlo 500;	"Monte Carlo sample size 500"
output c3 c4 c5.	"c3 contains other means, c4 contains response means, c5 contains treatment groups 1-4 from completely randomized design 5-8 from randomized block design"
code (1:4) 1 (5:8) 2 c5 c6	
desc c4; by c6.	"summary statistics "

Table C.3 Discrete prior distribution for π

π	$f(\pi)$
.3	.2
.4	.3
.5	.5

CHAPTER 8: BAYESIAN INFERENCE FOR BINOMIAL PROPORTION

Beta(a, b) Prior for π

BinoBP.mac is used to find the posterior when we have *binomial* (n, π) observation, and we have a *beta* (a, b) prior for π. The *beta* family of priors is conjugate for

Table C.4 Minitab commands for Bayesian inference on π with a discrete prior

Minitab Commands	Meaning
set c1	"puts π in c1"
.3 .4 .5	
end	
set c2	"puts $g(\pi)$ in c2"
.2 .3 .4	
end	
%<*insert path*>BinoDP.mac 6;	"n=6 trials"
prior c1 c2	"π in c1, prior $g(\pi)$ in c2"
observation 5;	"y=5 successes observed"
likelihood c3;	"store likelihood in c3"
posterior c4.	"store posterior $g(\pi\|y = 5)$ in c4"

Table C.5 Minitab commands for Bayesian inference on π with a *beta* prior

Minitab Commands	Meaning
%<*insert path*>BinoBP.mac 12 ;	"n=12 trials"
beta 3 3;	"the beta prior"
prior c1 c2;	"stores π and the prior $g(\pi)$"
observation 4 ;	"y=4 was observed"
likelihood c3;	"store likelihood in c3"
posterior c4;	"store posterior $g(\pi\|y = 4)$ in c4"

binomial (n, π) observations, so the posterior will be another member of the family, *beta* (a', b') where $a' = a + y$ and $b' = b + n - y$. For example, suppose we have $n = 12$ trials, and observe $y = 4$ successes, and we use a *beta* $(3, 3)$ prior for π. In the "edit" menu pull down "command line editor" and type the commands from Table C.5 into the command line editor. We can find the posterior mean and standard deviation from the output. We can determine an (equal tail area) credible interval for π by looking at the values of y1 that correspond to the desired tail area values of invf.

General Continuous Prior for π

BinoGCP.mac is used to find the posterior when we have *binomial* (n, π) observation, and we have a general continuous prior for π. Note, π must go from 0 to 1 in steps of .001, and $g(\pi)$ must be defined at each of the π values. For example, suppose we have $n = 12$ trials, and observe $y = 4$ successes, and we use a general continuous prior for π stored in c2. In the "edit" menu pull down "command line editor" and type the

Table C.6 Minitab commands for Bayesian inference on π with a continuous prior

Minitab Commands	Meaning
%<*insert path*>BinoGCP.mac 12 ;	"n=12 trials"
prior c1 c2	"inputs π in c1, prior $g(\pi)$ in c2"
observation 4;	"y=4 successes observed"
likelihood c3;	"store likelihood in c3"
posterior c4.	"store posterior $g(\pi\|y=4)$ in c4"

Table C.7 Minitab commands to integrate posterior density of π

Minitab Commands	Meaning
%<*insert path*>tintegral.mac c1 c4;	"integrates posterior density"
output k1 c6.	"stores definite integral over range in k1"
	"stores definite integral function in c6"
print c1 c6	

commands from Table c.6 into the command line editor. The output of *BinoGCP.mac* does not print out the posterior mean and standard deviation. Neither does it print out the values that give the tail areas of the integrated density function that we need to determine credible interval for π. Instead we use the macro *tintegral.mac* which numerically integrates a function over its range to determine these things. We can find the integral of the posterior density $g(\pi|y)$ using this macro. In the "edit" menu pull down "command line editor" and type the commands from Table C.7 into the command line editor. To find a 95% credible interval (with equal tail areas) we find the values in c1 that correspond to .025 and .975 in c6 respectively. We can also find the posterior mean and variance by numerically evaluating

$$m' = \int_0^1 \pi g(\pi|y)d\pi$$

and

$$(s')^2 = \int_0^1 (\pi - m')^2 g(\pi|y)d\pi$$

using the macro *tintegral.mac*. In the "edit" menu pull down "command line editor" and type the commands from Table C.8 into the command line editor.

CHAPTER 10: BAYESIAN INFERENCE FOR NORMAL MEAN

Discrete Prior for μ

NormDP.mac is used to find the posterior when we have a column of *normal* (μ, σ^2) observations and σ^2 is known, and we have a discrete prior for μ. For example,

Table C.8 Minitab commands to find posterior mean and variance

Minitab Commands	Meaning
let c7=c1*c4	"$\pi \times g(\pi\|y)$
%<*insert path*>tintegral.mac c1 c7;	"finds posterior mean"
output k1 c8.	
let c9=(c1-k1)**2 * c4	
%<*insert path*>tintegral.mac c1 c9;	"finds posterior variance"
output k2 c10.	
let k3=sqrt(k2)	"finds posterior st. deviation"
print k1-k3	

Table C.9 Discrete prior distribution for μ

μ	$f(\pi)$
2	.2
2.5	.2
3	.4
3.5	.2
4	.1

suppose μ has the discrete distribution with 5 possible values, 2 2.5, 3, 3.5 and ,4. Suppose the prior distribution is given in Table C.9. and we want to find the posterior distribution after a random sample of $n = 5$ observations from a *normal* $(\mu, \sigma^2 = 1)$ that are 1.52, 0.02, 3.35, 3.49 1.82 . In the "edit" menu pull down "command line editor" and type the commands from Table C.10 into the command line editor.

Normal(m, s^2) Prior for μ

NormNP.mac is used when we have a column c5 containing a random sample of n observations from a *normal* (μ, σ^2) distribution (with σ^2 known) and we use a *normal* (m, s^2) prior distribution. The normal family of priors is conjugate for *normal* (μ, σ^2) observations, so the posterior will be another member of the family, $normal[m', (s')^2]$ where the new constants are given by

$$\frac{1}{(s')^2} = \frac{1}{s^2} + \frac{n}{\sigma^2}$$

and

$$m' = \frac{\frac{1}{b^2}}{\frac{1}{(b')^2}} \times m + \frac{\frac{n}{\sigma^2}}{\frac{1}{(b')^2}} \times \bar{y}$$

Table C.10 Minitab commands for Bayesian inference on μ with discrete prior

Minitab Commands	Meaning
set c1	puts "μ in c1"
2:4/.5	
end	
set c2	"puts $g(\mu)$ in c2
.1 .2 .4 .2 .1	
end	
set c5	"puts $data$ in c5
1.52, 0.02, 3.35, 3.49 1.82	
end	
%<*insert path*>NormDP.mac c5 1;	"observed data in c5, known $\sigma = 1$"
prior c1 c2	"μ in c1, prior $g(\mu)$ in c2"
likelihood c3;	"store likelihood in c3"
posterior c4.	"store posterior $g(\mu\|data)$ in c4"

Table C.11 Minitab commands for Bayesian inference on μ with *normal* prior

Minitab Commands	Meaning
set c5	"puts $data$ in c5
2.99, 5.56, 2.83, and 3.47	
end	
%<*insert path*>NormNP.mac c5 1;	"observed data in c5, known $\sigma = 1$"
norm 3 2	"prior mean 3, prior std 2"
prior c1 c2	"store μ in c1, prior $g(\mu)$ in c2"
likelihood c3;	"store likelihood in c3"
posterior c4.	"store posterior $g(\mu\|data)$ in c4"

For example, suppose we have a normal random sample of 4 observations from *normal* $(\mu, \sigma^2 = 1)$ which are 2.99, 5.56, 2.83, and 3.47. Suppose we use a *normal* $(3, 2^2)$ prior for μ. In the "edit" menu pull down "command line editor" and type the commands from Table C.11 into the command line editor. We can determine an (equal tail area) credible interval for μ either by looking at the values of y1 corresponding to the desired values of invf printed out by *NormNP.mac*. We can find the posterior mean and variance from the output.

Table C.12 Minitab commands for Bayesian inference on μ with continuous prior

Minitab Commands	Meaning
set c5	"puts $data$ in c5
2.99, 5.56, 2.83, and 3.47	
end	
%<*insert path*>NormGCP.mac c5 1;	"observed data in c5, known $\sigma = 1$"
prior c1 c2	"μ in c1, prior $g(\mu)$ in c2"
likelihood c3;	"store likelihood in c3"
posterior c4.	"store posterior $g(\mu\|data)$ in c4"

Table C.13 Minitab commands to integrate posterior density of μ

Minitab Commands	Meaning
%<*insert path*>tintegral.mac c1 c4;	"integrates posterior density"
output k1 c6.	"stores definite integral over range in k1"
	"stores definite integral function in c6"
print c1 c6	

General Continuous Prior for μ

NormGCP.mac is used when we have a column c5 containing a random sample of n observations from a *normal* (μ, σ^2) distribution (with σ^2 known) and we have column c1 containing values of μ, and a column c2 containing values from a continuous prior $g(\mu)$.

For example, suppose we have a normal random sample of 4 observations from *normal* $(\mu, \sigma^2 = 1)$ which are 2.99, 5.56, 2.83, and 3.47. In the "edit" menu pull down "command line editor" and type the following commands from Table C.12 into the command line editor. The output of *NormGCP.mac* does not print out the posterior mean and standard deviation. Neither does it print out the values that give the tail areas of the integrated density function that we need to determine credible interval for μ. Instead we use the macro *tintegral.mac* which numerically integrates a function over its range to determine these things. We can find the integral of the posterior density $g(\mu|data)$ using this macro. In the "edit" menu pull down "command line editor" and type the commands from Table C.13 into the command line editor. To find a 95% credible interval (with equal tail areas) we find the values in c1 that correspond to .025 and .975 in c6 respectively. We can find the posterior mean and variance by numerically evaluating

$$m' \quad = \quad \int \mu g(\mu|data) d\mu$$

Table C.14 Minitab commands to find posterior mean and variance of μ

Minitab Commands	Meaning
let c7=c1*c4	"$\mu \times g(\mu|data)$
%<*insert path*>tintegral.mac c1 c7;	"finds posterior mean"
output k1 c8.	
let c8=(c1-k1)**2 * c4	
%<*insert path*>tintegral.mac c1 c8;	"finds posterior variance"
output k2 c9.	
let k3=sqrt(k2)	"finds posterior st. deviation"
print k1-k3	

and

$$(s')^2 \quad = \quad \int (\mu - m')^2 g(\mu|data) d\mu$$

using the macro *tintegral.mac*. In the "edit" menu pull down "command line editor" and type the commands from Table C.14 into the command line editor.

D

Using the Included R Functions

OBTAINING AND USING R AND THE R FUNCTIONS

R functions for performing Bayesian analysis and for doing Monte Carlo simulations are included. The address may be downloaded from the Web page for this text on the site <www.wiley.com>. The R functions are zipped up in a package called *Bolstad R Functions.zip*.

The latest version of R may always be found at *http://lib.stat.cmu.edu/R/CRAN/*. Compiled versions of R for Linux, Mac OS (System 8.6 to 9.1 and Mac OS X), Mac OS X (Darwin/X11) and Windows (95 and later), and the source code (for those who wish to compile R themselves) may also be found at this address.

To install R for Windows, double click on the file *rw1070.exe* and follow the installer functions. In the following discussion it is assumed that you have copied the file $Bolstad.R.Functions.zip$ to a location on your computer. You can find it in this way:

1. Start R from the Start menu or by double clicking on the icon on your desktop.
2. Pull down the 'Packages' menu and select the item 'Install package from local zip file...'

*

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

3. Use the dialog box to locate *Bolstad.R.Functions.zip*, select it and click on 'Open'.

R will now recognize the package `Bolstad.R.Functions` as a package it can load. To use the functions in the package `Bolstad.R.Functions`, either type `library(bolstad.R.Functions)` at the command prompt or select the item 'Load package...' from the 'Packages' menu. To see the list of functions contained within the package, type `library(help=Bolstad.R.Functions)`. This should bring up the following list:

Function Name	Description
binobp	binomial sampling with a beta prior
binodp	binomial sampling with a discrete prior
binogcp	binomial sampling with a general continuous prior
normdp	Bayesian inference on a normal mean with a discrete prior
normgcp	Bayesian inference on a normal mean with a general continuous prior
normnp	Bayesian inference on a normal mean with a normal prior
sintegral	numerical integration using Simpson's Rule
sscsample	simple, stratified and cluster sampling
sscsample.data	A stratified and clustered data set
xdesign	carry out simulations using the default parameters

Help on each of the R functions is available once you have loaded the `bolstad` package. There are a number of ways to access help files under R. The traditional way is to use the `help` or `?` function. For example, to see the help file on the `binodp` function, type `help(binodp)` or `?binodp`. HTML-based help is also available. To use HTML help, select 'R language(html)' from the 'Help' menu. Click on the 'Packages' link, and then the link for 'bolstad'. This will bring up an index page where you may select the help file for the function you're interested in.

Each help file has a standard layout, which is as follows:

Title: a brief title that gives some idea of what the function is supposed to do or show

Description: a fuller description of the what the function is supposed to do or show

Usage: the formal calling syntax of the function

Arguments: a description of each of the arguments of the function

Values: a description of the values (if any) returned by the function

See also: a reference to related functions

Examples: some examples of how the function may be used. These examples may be copied and pasted into R

The R language has two special features that may make it confusing to users of other programming and statistical languages: default or optional arguments, and variable ordering of arguments. An R function may have arguments for which the author has specified a default value. Let's take the function `binobp` as an example. The syntax of `binobp` is `binobp(x, n, a = 1, b = 1, ret = FALSE)`. The function takes five arguments `x`, `n`, `a`, `b`, and `ret`. However, the author has specified default values for `a`, `b`, and `ret`, namely `a = 1`, `b = 1` and `ret = FALSE`. This means that the user only has to supply the arguments `x` and `n`. Therefore the arguments `a`, `b` and `ret` are said to be optional or default. In this example, by default, a *beta*($a = 1$, $b = 1$) prior is used and the prior, likelihood, and posterior distributions (along with some associated information) are not returned (`ret = FALSE`.) Hence the simplest example for `binobp` is given as `binobp(6,8)`. If the user wanted to change the prior used, say to *beta*(5,6), then they would type `binobp(6,8,5,6)`. There is a slight catch here, which leads into the next feature. Assume that the user wanted to use a *beta*(1,1) prior, but wanted to return the output. One might be tempted to type `binobp(6,8,FALSE)`. This is incorrect. R will think that the value `FALSE` is the value being assigned to the parameter `a`, and convert it from a logical value, `FALSE`, to the numerical equivalent, `0`, which will of course give an error because the parameters of the beta distribution must be greater than zero. The correct way to make such a call is to use named arguments, such as `binobp(6,8,ret=FALSE)`. This specifically tells R which argument is to be assigned the value `FALSE`. This feature also makes the calling syntax more flexible because it means that the order of the arguments does not need to be adhered to. For example, `binobp(n=8, x=6, ret=FALSE, a=1, b=3)` would be a perfectly legitimate function call.

CHAPTER 2: SCIENTIFIC DATA GATHERING

In this chapter we use the function `sscsample` to perform a small-scale Monte Carlo study on the efficiency of simple, stratified, and cluster random sampling on the population data contained in `sscsample.data`. Make sure the `bolstad` package is loaded by typing

```
library(bolstad)
```

first. Type the following commands into the R console:

```
sscsample(20,200)
```

This calls the `sscsample` function and asks for 200 samples of size 20 to be drawn from the dataset `sscsample.data`. To return the means and the samples themselves, type

```
res<-sscsample(20,200,ret=T)
```

This will store all 200 samples and their means in an R list structure called `res`. The means of the sample may be accessed by typing

```
res\$means
```

The samples themselves are stored in the columns of a 20 $\times$ 200 matrix called `res$samples`. To access the i^{th} sample, where $i = 1, \ldots, 200$, type

```
res\$samples[,i]
```

For example, to access the 50^{th} sample, type

```
res\$samples[,50]
```

Experimental Design

We use the function `xdesign` to perform a small-scale Monte Carlo study comparing *completely randomized design* and *randomized block design* in their effectiveness for assigning experimental units into treatment groups. Suppose we want to carry out our study with four treatment groups, each of size 20, and with a correlation of 0.8 between the response and the blocking variable. Type the following commands into the command line editor:

```
xdesign()
```

Suppose we want to carry out our study with five treatment groups, each of size 25, and with a correlation of -0.6 between the response and the blocking variable. We also want to store the results of the simulation in a variable called `res`. Type the following commands into the command line:

```
res<-xdesign(corr=-0.6,size=25,n.treatments=5)
```

`res` is a list containing three member vectors of length `2×n.treatments×n.rep`. Each block of `n.rep` elements contains the simulated means for each Monte Carlo replicate with in a specific treatment group. The first `n.treatments` blocks correspond to the *completely randomized design*, and the second `n.treatments` blocks correspond to *randomized block design*

- `block.means`: a vector of the means of the blocking variable
- `treat.means`: a vector of the means of the response variable
- `ind`: a vector indicating which means belong to which treatment group

An example of using these results might be

```
boxplot(block.means~ind,data=res)
boxplot(treat.means~ind,data=res)
```

CHAPTER 6: BAYESIAN INFERENCE FOR DISCRETE RANDOM VARIABLES

Binomial Proportion with Discrete Prior

`binodp` is used to find the posterior when we have a *binomial* (n, θ) observation, and we have a discrete prior for θ. For example, suppose θ has the discrete distribution with three possible values, .3, .4, and .5. Suppose the prior distribution is

θ	$f(\theta)$
.3	.2
.4	.3
.5	.5

and we want to find the posterior distribution after $n = 6$ trials and observing $y = 5$ successes. Type the following commands into the command line editor:

```
theta<-c(0.3,0.4,0.5)
theta.prior<-c(0.2,0.3,0.5)
binodp(5,6,theta=theta,theta.prior=theta.prior)
```

CHAPTER 8: BAYESIAN INFERENCE FOR BINOMIAL PROPORTION

Beta(a, b) Prior for π

`binobp` is used to find the posterior when we have a *binomial* (n, θ) observation, and we have a *beta* (a, b) prior for θ. The *beta* family of priors is conjugate for *binomial* (n, θ) observations, so the posterior will be another member of the family, *beta* (a', b') where $a' = a + y$ and $b' = b + n - y$. For example, suppose we have $n = 12$ trials, and observe $y = 4$ successes, and use a *beta* $(3, 3)$ prior for θ. Type the following command:

```
binobp(4,12,3,3)
```

into the R console. This should give the following output:

```
> binobp(4,12,3,3)
Posterior Mean           :  0.3888889
Posterior Variance       :  0.0125081
Posterior Std. Deviation :  0.1118397


Prob.   Quantile
------  ---------
0.005   0.1370832
0.01    0.1552348
```

```
0.025   0.184437
0.05    0.2119082
0.5     0.3846872
0.95    0.5802946
0.975   0.6167163
0.99    0.6577095
0.995   0.6845936
```

We can find the posterior mean and standard deviation from the output. We can determine an (equal tail area) credible interval for θ by taking the appropriate quantiles that correspond to the desired tail area values of the interval. For example, for 95% credible interval we take the quantiles with probability 0.025 and 0.975, respectively. These are 0.184 and 0.617.

General Continuous Prior for π

`binogcp` is used to find the posterior when we have a *binomial* (n, θ) observation, and we have a general continuous prior for θ. Note, θ must go from 0 to 1 in steps of at least .01, and $g(\theta)$ must be defined at each of the θ values. For example, suppose we have $n = 12$ trials and observe $y = 4$ successes. In this example our continuous prior for θ is a $N(\mu = 0.5, \sigma = 0.25)$. Type the following commands into the R console:

```
binogcp(4,12,density="normal",params=c(0.5,0.25))
```

This example is perhaps not quite general as it uses some of the built in functionality of `binogcp`. In this second example we use a "user-defined" general continuous prior. Let the probability density function be a triangular distribution defined by

$$g(\theta) = \begin{cases} 4\theta & ,0 \leq \theta \leq 0.5 \\ 4 - 4\theta & 0.5 < \theta \leq 1 \end{cases}$$

.

Type the following commands into the R console:

```
theta<-seq(0,1,by=0.001)
theta.prior<-rep(0,length(theta))
theta.prior[theta<=0.5]<-4*theta[theta<=0.5]
theta.prior[theta>0.5]<-4-4*theta[theta>0.5]
results<-binogcp(4,12,"user",theta=theta,
                 theta.prior=theta.prior,ret=TRUE)
```

The output of `binogcp` does not print out the posterior mean and standard deviation. Neither does it print out the values that give the tail areas of the integrated density function that we need to determine credible interval for θ. Instead, we use the function `sintegral`, which numerically integrates a function over its range to determine these things. We can find the integral of the posterior density $g(\theta|y)$ using this macro. Type the following command into the R console:

```
cdf<-sintegral(theta,results\$posterior,
                    n.pts=length(theta),ret=TRUE)
plot(cdf,type="l",xlab=expression(theta[0])
              ,ylab=expression(Pr(theta<=theta[0])))
```

These commands created a new variable `cdf`, which is a list containing values `x` and `y`, where `cdf$y` is equal to $\Pr(Y \leq x)$, i.e. the cumulative density function (cdf.) To find a 95% credible interval (with equal tail areas) we find the values of `cdf$x` that correspond to .025 and .975 in `cdf$y` respectively.

```
d<-abs(cdf\$y-0.025)
lb<-cdf\$x[max((1:length(cdf\$y))[d==min(d)])]


d<-abs(cdf\$y-0.975)
ub<-cdf\$x[min((1:length(cdf\$y))[d==min(d)])]


cat(paste("Approximate 95% credible interval : ["
           ,round(lb,4)," ",round(ub,4),"]\n",sep=""))
```

We can also find the posterior mean and variance by numerically evaluating

$$m' = \int_0^1 \theta g(\theta|y)d\theta$$

and

$$(s')^2 = \int_0^1 (\theta - m')^2 g(\theta|y)d\theta$$

using the function `sintegral`.Type the following commands into the R console:

```
dens<-theta*results\$posterior post.mean<-sintegral(theta,dens)


dens<-(theta-post.mean)^2*results\$posterior
post.var<-sintegral(theta,dens)


post.sd<-sqrt(post.var)
```

Of course we can use these values to calculate an approximate 95% credible interval using standard theory:

```
lb<-post.mean-qnorm(0.975)*post.sd
ub<-post.mean+qnorm(0.975)*post.sd

cat(paste("Approximate 95% credible interval : ["
   ,round(lb,4)," ",round(ub,4),"]\n",sep=""))
```

CHAPTER 10: BAYESIAN INFERENCE FOR NORMAL MEAN

Discrete Prior for μ

`normdp` is used to find the posterior when we have a vector of *normal* (μ, σ^2) observations and σ^2 is known, and we have a discrete prior for μ. For example, suppose μ has the discrete distribution with five possible values: 2, 2.5, 3, 3.5, and 4. Suppose the prior distribution is

μ	$f(\theta)$
2	.1
2.5	.2
3	.4
3.5	.2
4	.1

and we want to find the posterior distribution after a random sample of $n = 5$ observations from a *normal* $(\mu, \sigma^2 = 1)$ that are 1.52, 0.02, 3.35, 3.49, and 1.82. Type the following commands into the R console:

```
mu<-seq(2,4,by=0.5)
mu.prior<-c(0.1,0.2,0.4,0.2,0.1)
y<-c(1.52,0.02,3.35,3.49,1.82)
normdp(y,1,uniform=F,n.mu=5,mu,mu.prior)
```

Normal(m, s^2) Prior for μ

`normnp` is used when we have a vector containing a random sample of n observations from a *normal* (μ, σ^2) distribution (with σ^2 known) and we use a *normal* (m, s^2) prior distribution. The normal family of priors is conjugate for *normal* (μ, σ^2) observations, so the posterior will be another member of the family, $normal[m', (s')^2]$ where the new constants are given by

$$\frac{1}{(s')^2} = \frac{1}{s^2} + \frac{n}{\sigma^2}$$

and

$$m' = \frac{\frac{1}{b^2}}{\frac{1}{(b')^2}} \times m + \frac{\frac{n}{\sigma^2}}{\frac{1}{(b')^2}} \times \bar{y}\,.$$

For example, suppose we have a normal random sample of four observations from *normal* $(\mu, \sigma^2 = 1)$ that are 2.99, 5.56, 2.83, and 3.47. Suppose we use a *normal* $(3, 2^2)$ prior for μ. Type the following commands into the R console:

```
y<-c(2.99,5.56,2.83,3.47)
normnp(y,1,3,2)
```

This gives the following output:

```
Posterior mean           : 3.6705882
Posterior std. deviation : 0.4850713


Prob.   Quantile
------  ---------
0.005   2.4211275
0.01    2.5421438
0.025   2.7198661
0.05    2.872717
0.5     3.6705882
0.95    4.4684594
0.975   4.6213104
0.99    4.7990327
0.995   4.920049
```

We can find the posterior mean and standard deviation from the output. We can determine an (equal tail area) credible interval for μ by taking the appropriate quantiles that correspond to the desired tail area values of the interval. For example, for 99% credible interval we take the quantiles with probability 0.005 and 0.995, respectively. These are 2.42 and and 4.92.

General Continuous Prior for μ

`normgcp` is used when we have a vector containing a random sample of n observations from a *normal* (μ, σ^2) distribution (with σ^2 known) and we have vector containing values of μ, and vector containing values from a continuous prior $g(\mu)$.

For example, suppose we have a random sample of four observations from a *normal* $(\mu, \sigma^2 = 1)$ distribution. The values are 2.99, 5.56, 2.83, and 3.47. Suppose we have a triangular prior defined over -3 to 3:

$$g(\mu) = \begin{cases} \frac{1}{3} + \frac{\mu}{9} & , -3 \leq \mu \leq 00 \\ \frac{1}{3} - \frac{\mu}{9} & 0 < \mu \leq 3 \end{cases} .$$

Type the following commands into the R console:

```
y<-c(2.99,5.56,2.83,3.47)
mu<-seq(-3,3,by=0.1)
mu.prior<-rep(0,length(mu))
mu.prior[mu<=0]<-1/3+mu[mu<=0]/9
mu.prior[mu>0]<-1/3-mu[mu>0]/9
results<-normgcp(y,1,density="user",mu=mu,
                 mu.prior=mu.prior,ret=T)
```

The output of `normgcp` does not print out the posterior mean and standard deviation. Neither does it print out the values that give the tail areas of the integrated

density function that we need to determine credible interval for μ. Instead we use the macro `sintegral` which numerically integrates a function over its range to determine these things. We can find the integral of the posterior density $g(\mu|data)$ using this macro. Type the following commands into the R console:

```
cdf<-sintegral(mu,results\$posterior
               ,n.pts=length(mu),ret=TRUE)
plot(cdf,type="l",xlab=expression(mu[0])
            ,ylab=expression(Pr(mu<=mu[0])))
```

These commands created a new variable `cdf`, which is a list containing values `x` and `y`, where `cdf$y` is equal to $\Pr(Y \leq x)$, i.e. the cumulative density function (cdf.) To find a 95% credible interval (with equal tail areas), we find the values of `cdf$x` that correspond to .025 and .975 in `cdf$y`, respectively.

```
d<-abs(cdf$y-0.025)
lb<-cdf$x[max((1:length(cdf$y))[d==min(d)])]


d<-abs(cdf$y-0.975)
ub<-cdf$x[min((1:length(cdf$y))[d==min(d)])]


cat(paste("Approximate 95% credible interval : ["
          ,round(lb,4)," ",round(ub,4),"]\n",sep=""))
```

We can also find the posterior mean and variance by numerically evaluating

$$m' \quad = \quad \int \mu g(\mu|data) d\mu$$

and

$$(s')^2 \quad = \quad \int (\mu - m')^2 g(\mu|data) d\mu$$

using the function `sintegral`. Type the following commands into the R console:

```
dens<-mu*results$posterior
post.mean<-sintegral(mu,dens)


dens<-(mu-post.mean)^2*results$posterior
post.var<-sintegral(mu,dens)


post.sd<-sqrt(post.var)
```

Of course, we can use these values to calculate an approximate 95% credible interval using standard theory:

```
lb<-post.mean-qnorm(0.975)*post.sd
ub<-post.mean+qnorm(0.975)*post.sd


cat(paste("Approximate 95% credible interval : ["
   ,round(lb,4)," ",round(ub,4),"]\n",sep=""))
```

E

Answers to Selected Exercises

Chapter 3: Displaying and Summarizing Data

3.1 (a) Stem-and-leaf plot for sulphur dioxide (SO_2) data

		leaf unit 1
0	3 3
0	5 7 99
1	1334
1	6789
2	33
2	56789
3	
3	5
4	34
4	6

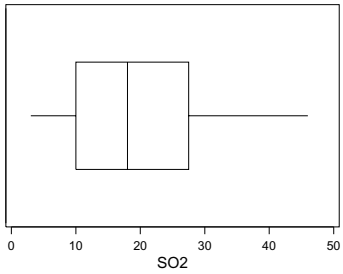
(b) Median $Q_2 = X_{[13]} = 18$,

Lower quartile $Q_1 = X_{[\frac{26}{4}]} = \frac{X_6 + X_7}{2} = 10$, and

Upper quartile $Q_3 = X_{[\frac{78}{4}]} = \frac{X_{19} + X_{20}}{2} = 27.5$

*
[0] *Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

(c) Boxplot of SO_2 data

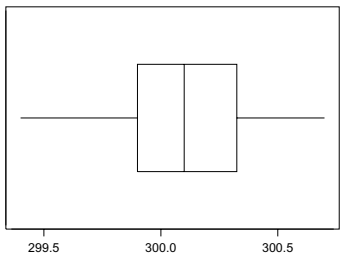


3.3 (a) Stem-and-leaf plot for distance measurements data

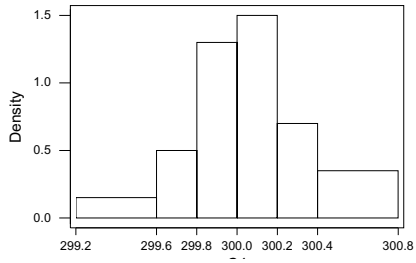
	leaf unit .01
299.4	0
299.5	0
299.6	0
299.7	00
299.8	000
299.9	000000
300.0	0000000
300.1	00000000
300.2	0000000
300.3	00
300.4	00000
300.5	000
300.6	00
300.7	00

(b) Median=300.1 $Q_1 = 299.9$ $Q_3 = 300.35$

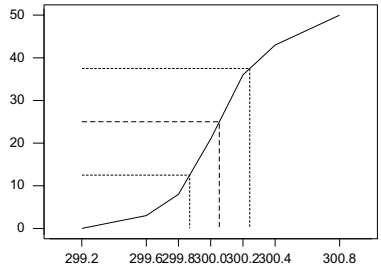
(c) Boxplot of distance measurement data



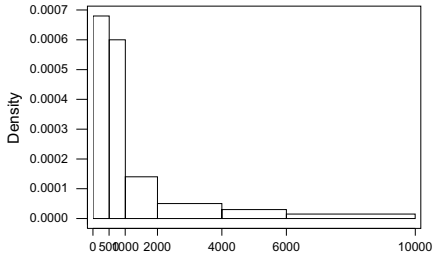
(d) Histogram of distance measurement data



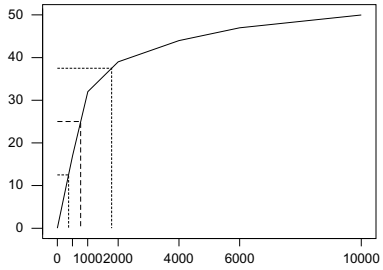
(e) Cumulative frequency polygon of distance measurement data



3.5 (a) Histogram of liquid cash reserve

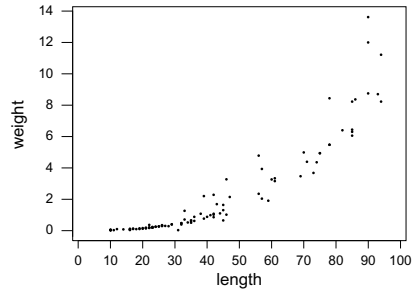


(b) Cumulative frequency polygon of liquid cash reserve

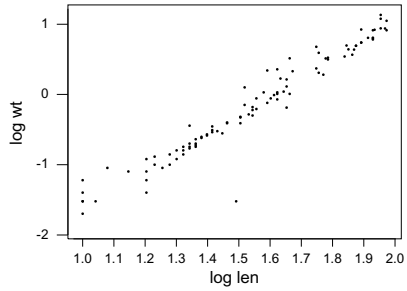


(c) Grouped mean = 1600

3.7 (a) Plot of weight versus length (slug data)



(b) Plot of log(weight) versus log(length)



(c) The point $(1.5, -1.5)$ does not seem to fit the pattern. This corresponds to observation 90. Dr. Harold Henderson at AgResearch New Zealand has told me that there are two possible explanations for this point. Either the digits of length were transposed at recording or the decimal place for weight was misplaced.

Chapter 4: Logic, Probability, and Uncertainty

4.1 (a) $P(\tilde{A}) = .6$

(b) $P(A \cap B) = .2$

(c) $P(A \cup B) = .7$

4.3 (a) $P(\tilde{A} \cap B) = .24$, $P(B) = .4$, therefore $P(A \cap B) = .16$. $P(A \cap B) = P(A) \times P(B)$, therefore they are independent.

(b) $P(A \cup B) = .4 + .4 - .16 = .64$

4.5 (a) $\Omega = \{1, 2, 3, 4, 5, 6\}$

(b) $A = \{2, 4, 6\}$, $P(A) = \frac{3}{6}$

(c) $B = \{3, 6\}$, $P(B) = \frac{2}{6}$

(d) $A \cap B = \{6\}$, $P(A \cap B) = \frac{1}{6}$

(e) $P(A \cap B) = P(A) \times P(B)$, therefore they are independent.

4.7 (a)

$$A = \left\{ \begin{array}{ccc} (1,1) & (1,3) & (1,5) \\ (2,2) & (2,4) & (2,6) \\ (3,1) & (3,3) & (3,5) \\ (4,2) & (4,4) & (4,6) \\ (5,1) & (5,3) & (5,5) \\ (6,2) & (6,4) & (6,6) \end{array} \right\}$$

$P(A) = \frac{18}{36}$

(b)

$$B = \left\{ \begin{array}{cccccc} (1,2) & (1,5) & (2,1) & (2,4) & (3,3) & (3,6) \\ (4,2) & (4,5) & (5,1) & (5,4) & (6,3) & (6,6) \end{array} \right\}$$

$P(B) = \frac{12}{36}$

(c) $A \cap B = \{(1,5)(2,4)(3,3)(4,2)(5,1)(6,6)\}$
$P(A \cap B) = \frac{6}{36}$

(d) $P(A \cap B) = P(A) \times P(B)$, yes they are independent.

4.9 Let D be "the person has the disease" and let T be "The test result was positive."

$$P(D|T) = \frac{P(D \cap T)}{P(T)} = .0875$$

Chapter 5: Discrete Random Variables

5.1 (a) $P(1 < Y \leq 3) = .4$

(b) $E(Y) = 1.6$

(c) $Var(Y) = 1.44$

(d) $E(W) = 6.2$

(e) $Var(W) = 5.76$

5.3 (a) The filled-in table:

y_i	$f(y_i)$	$y_i \times f(y_i)$	$y_i^2 \times f(y_i)$
0	.0102	.0000	.0000
1	.0768	.0768	.0768
2	.2304	.4608	.9216
3	.3456	1.0368	3.1104
4	.2592	1.0368	4.1472
5	.0778	.3890	1.9450
Sum	1.0000	3.0000	10.2000

i. $E(Y) = 3$

ii. $Var(Y) = 10.2 - 3^2 = 1.2$

(b) Using formulas

i. $E(Y) = 5 \times .6 = 3$

ii. $Var(Y) = 5 \times .6 \times .4 = 1.2$

5.5 The filled-in table:

X	Y					$f(x)$
	1	2	3	4	5	
1	.02	.04	.06	.08	.05	.25
2	.08	.02	.10	.02	.03	.25
3	.05	.05	.03	.02	.10	.25
4	.10	.04	.05	.03	.03	.25
$f(y)$	.25	.15	.24	.15	.21	

(a) The marginal distribution of X is found by summing across rows.

(b) The marginal distribution of Y is found by summing down columns.

(c) No they are not. The entries in the joint probability table aren't all equal to the products of the marginal probabilities.

(d) $P(X = 3|Y = 1) = \frac{.05}{.25} = .20$

Chapter 6: Bayesian Inference for Discrete Random Variables

6.1 (a) Bayesian universe:

$$\left\{\begin{array}{l} (0,0)\ (0,1) \\ (1,0)\ (1,1) \\ (2,0)\ (2,1) \\ (3,0)\ (3,1) \\ (4,0)\ (4,1) \\ (5,0)\ (5,1) \\ (6,0)\ (6,1) \\ (7,0)\ (7,1) \\ (8,0)\ (8,1) \\ (9,0)\ (9,1) \end{array}\right\}$$

(b) The filled-in table:

X	$prior$	$Y=0$	$Y=1$
0	$\frac{1}{9}$	$\frac{1}{9}\times\frac{9}{9}$	$\frac{1}{9}\times\frac{0}{9}$
1	$\frac{1}{9}$	$\frac{1}{9}\times\frac{8}{9}$	$\frac{1}{9}\times\frac{1}{9}$
2	$\frac{1}{9}$	$\frac{1}{9}\times\frac{7}{9}$	$\frac{1}{9}\times\frac{2}{9}$
3	$\frac{1}{9}$	$\frac{1}{9}\times\frac{6}{9}$	$\frac{1}{9}\times\frac{3}{9}$
4	$\frac{1}{9}$	$\frac{1}{9}\times\frac{5}{9}$	$\frac{1}{9}\times\frac{4}{9}$
5	$\frac{1}{9}$	$\frac{1}{9}\times\frac{4}{9}$	$\frac{1}{9}\times\frac{5}{9}$
6	$\frac{1}{9}$	$\frac{1}{9}\times\frac{3}{9}$	$\frac{1}{9}\times\frac{6}{9}$
7	$\frac{1}{9}$	$\frac{1}{9}\times\frac{2}{9}$	$\frac{1}{9}\times\frac{7}{9}$
8	$\frac{1}{9}$	$\frac{1}{9}\times\frac{1}{9}$	$\frac{1}{9}\times\frac{8}{9}$
9	$\frac{1}{9}$	$\frac{1}{9}\times\frac{0}{9}$	$\frac{1}{9}\times\frac{9}{9}$

which simplifies to

X	$prior$	$Y=0$	$Y=1$
0	$\frac{1}{9}$	$\frac{9}{81}$	$\frac{0}{81}$
1	$\frac{1}{9}$	$\frac{8}{81}$	$\frac{1}{81}$
2	$\frac{1}{9}$	$\frac{7}{81}$	$\frac{2}{81}$
3	$\frac{1}{9}$	$\frac{6}{81}$	$\frac{3}{81}$
4	$\frac{1}{9}$	$\frac{5}{81}$	$\frac{4}{81}$
5	$\frac{1}{9}$	$\frac{4}{81}$	$\frac{5}{81}$
6	$\frac{1}{9}$	$\frac{3}{81}$	$\frac{6}{81}$
7	$\frac{1}{9}$	$\frac{2}{81}$	$\frac{7}{81}$
8	$\frac{1}{9}$	$\frac{1}{81}$	$\frac{8}{81}$
9	$\frac{1}{9}$	$\frac{0}{81}$	$\frac{9}{81}$
		$\frac{45}{81}$	$\frac{45}{81}$

(c) The marginal distribution was found by summing down the columns.

(d) The reduced Bayesian universe is

$$\left\{\begin{array}{c}(0,1)\\(1,1)\\(2,1)\\(3,1)\\(4,1)\\(5,1)\\(6,1)\\(7,1)\\(8,1)\\(9,1)\end{array}\right\}.$$

(e) The posterior probability distribution is found by dividing the joint probabilities on the reduced Bayesian universe, by the sum of the joint probabilities over the reduced Bayesian universe.

(f) The simplified table is

X	*prior*	*likelihood*	*prior* $\times$ *likelihood*	*posterior*
0	$\frac{1}{9}$	$\frac{0}{9}$	$\frac{0}{81}$	$\frac{0}{45}$
1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{81}$	$\frac{1}{45}$
2	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{2}{81}$	$\frac{2}{45}$
3	$\frac{1}{9}$	$\frac{3}{9}$	$\frac{3}{81}$	$\frac{3}{45}$
4	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{4}{81}$	$\frac{4}{45}$
5	$\frac{1}{9}$	$\frac{5}{9}$	$\frac{5}{81}$	$\frac{5}{45}$
6	$\frac{1}{9}$	$\frac{6}{9}$	$\frac{6}{81}$	$\frac{6}{45}$
7	$\frac{1}{9}$	$\frac{7}{9}$	$\frac{7}{81}$	$\frac{7}{45}$
8	$\frac{1}{9}$	$\frac{8}{9}$	$\frac{8}{81}$	$\frac{8}{45}$
9	$\frac{1}{9}$	$\frac{9}{9}$	$\frac{9}{81}$	$\frac{9}{45}$
Sum			$\frac{45}{81}$	1

6.3 Looking at the two draws together, the simplified table is

X	*prior*	*likelihood*	*prior* $\times$ *likelihood*	*posterior*
0	$\frac{1}{9}$	$\frac{0}{9} \times 1$	$\frac{0}{81}$	$\frac{0}{120}$
1	$\frac{1}{9}$	$\frac{1}{9} \times \frac{8}{8}$	$\frac{8}{648}$	$\frac{8}{120}$
2	$\frac{1}{9}$	$\frac{2}{9} \times \frac{7}{8}$	$\frac{14}{648}$	$\frac{8}{120}$
3	$\frac{1}{9}$	$\frac{3}{9} \times \frac{6}{8}$	$\frac{18}{648}$	$\frac{8}{120}$
4	$\frac{1}{9}$	$\frac{4}{9} \times \frac{5}{8}$	$\frac{20}{648}$	$\frac{8}{120}$
5	$\frac{1}{9}$	$\frac{5}{9} \times \frac{4}{8}$	$\frac{20}{648}$	$\frac{8}{120}$
6	$\frac{1}{9}$	$\frac{6}{9} \times \frac{3}{8}$	$\frac{18}{648}$	$\frac{8}{120}$
7	$\frac{1}{9}$	$\frac{7}{9} \times \frac{2}{8}$	$\frac{14}{648}$	$\frac{8}{120}$
8	$\frac{1}{9}$	$\frac{8}{9} \times \frac{1}{8}$	$\frac{8}{648}$	$\frac{8}{120}$
9	$\frac{1}{9}$	$\frac{9}{9} \times \frac{0}{8}$	$\frac{0}{648}$	$\frac{8}{120}$
Sum			$\frac{120}{648}$	1

6.5

$$P(\text{"}Blackjack\text{"}) = P(A) \times P(F|A) + P(F) \times P(A|F)$$

(They are disjoint ways of getting "Blackjack.")

$$P(\text{"}Blackjack\text{"}) = \frac{16}{204} \times \frac{64}{203} + \frac{64}{204} \times \frac{16}{203} = 0.0494543$$

Chapter 7: Continuous Random Variables

7.1 (a) $E(X) = \frac{3}{8} = .375$

(b) $Var(X) = \frac{15}{8^2 \times 9} = 0.0260417$

7.3 The uniform distribution is also the *beta* $(1, 1)$ distribution.

(a) $E(X) = \frac{1}{2} = .5$

(b) $Var(X) = \frac{1}{2^2 \times 3} = .08333$

(c) $P(X \leq .25) = \int_0^{.25} 1 \; dx = .25$

(d) $P(.33 < X < .75) = \int_{.33}^{.75} 1 \; dx = .42$

7.5 (a) $P(0 \leq Z < .65) = .2422$

(b) $P(Z \geq .54) = .2946$

(c) $P(-.35 \leq Z \leq 1.34) = .5467$

7.7 (a) $P(Y \leq 130) = .8944$

(b) $P(Y \geq 135) = .0304$

(c) $P(114 \leq Y \leq 127) = .5826$

7.9 (a) $E(Y) = \frac{10}{10+12} = .4545$

(b) $Var(Y) = \frac{10 \times 12}{(22)^2 \times (23)} = .0107797$

(c) $P(Y > .5) = .3308$

Chapter 8: Bayesian Inference for Binomial Proportion

8.1 (a) *binomial* $(n = 150, \pi)$ distribution

(b) *beta* $(30, 122)$

8.3 (a) a and b are the simultaneous solutions of

$$\frac{a}{a+b} = .5$$

and

$$\frac{a \times b}{(a+b)^2 \times (a+b+1)} = .15^2$$

Solution is $a = 5.05$ and $b = 5.05$

(b) The equivalent sample size of her prior is 11.11

(c) *beta* $(26.05, 52.05)$

8.5 (a) *binomial* $(n = 116, \pi)$

(b) *beta* $(18, 103)$

(c)

$$E(\pi|y) = \frac{18}{18+103}$$

and

$$Var(\pi|y) = \frac{18 \times 103}{(121)^2 \times (122)}$$

(d) *normal*$(.149, .0322^2)$

(e) (.086,.212)

8.7 (a) *binomial* $(n = 174, \pi)$

(b) *beta* $(11, 168)$

(c)

$$E(\pi|y) = \frac{11}{11+168} = .0614$$

and

$$Var(\pi|y) = \frac{11 \times 168}{(179)^2 \times (180)} = .0003204$$

(d) *normal*$(.061, .0179^2)$

(e) (.026,.097)

Chapter 9: Comparing Bayesian and Frequentist Inferences for Proportion

9.1 (a) *binomial* $(n = 30, \pi)$

(b) $\hat{\pi}_f = \frac{8}{30} = .267$

(c) *beta* $(9, 23)$

(d) $\hat{\pi}_B = \frac{9}{32} = .281$

9.3 (a) $\hat{\pi}_f = \frac{11}{116} = .095$

(b) *beta* $(12, 115)$

(c) $E(\pi|y) = .094$ and $Var(\pi|y) = .0006684$
The Bayesian estimator $\hat{\pi}_B = .094$.

(d) (.044,.145)

(e) The null value $\pi = .10$ lies in the credible interval, so it remains a credible value at the 5% level

9.5 (a) $\hat{\pi}_f = \frac{24}{176} = .136$

(b) *beta* $(25, 162)$

(c) $E(\pi|y) = .134$ and $Var(\pi|y) = .0006160$
The Bayesian estimator $\hat{\pi}_B = .134$.

(d)

$$P(\pi \geq .15) = .255$$

This is greater than level of significance .05, so we can't reject the null hypothesis $H_0 : \pi \geq .15$.

Chapter 10: Bayesian Inference for Normal Mean

10.1 (a) posterior distribution

value	posterior probability
991	.0000
992	.0000
993	.0000
994	.0000
995	.0000
996	.0010
997	.0674
998	.4980
999	.3987
1000	.0346
1001	.0003
1002	.0000
1003	.0000
1004	.0000
1005	.0000
1006	.0000
1007	.0000
1008	.0000
1009	.0000
1010	.0000

(b) $P(\mu < 1000) = .965$.

10.3 (a) The posterior precision equals

$$\frac{1}{(s')^2} = \frac{1}{10^2} + \frac{10}{3^2} = 1.1211$$

The posterior variance equals $(s')^2 = \frac{1}{1.1211} = .89197$. The posterior standard deviation equals $s' = \sqrt{.89197} = .9444$. The posterior mean equals

$$m' = \frac{\frac{1}{10^2}}{1.1211} \times 30 + \frac{\frac{10}{3^2}}{1.1211} \times 36.93 = 36.87$$

The posterior distribution of μ is *normal*$(36.87, .9444^2)$.

(b) Test

$$H_0 : \mu \leq 35 \quad \text{versus} \quad H_1 : \mu > 35$$

Note that the alternative hypothesis is what we are trying to determine. The null hypothesis is that mean yield is unchanged from that of the standard process.

(c)

$$P(\mu \leq .35) = P\left(\frac{\mu - 36.87}{.944} \leq \frac{35 - 36.87}{.944}\right)$$

$$= P(Z \leq -2.012) = .022$$

This is less than the level of significance $\alpha = .05\%$, so we reject the null hypothesis and conclude the yield of the revised process is greater than .35.

10.5 (a) The posterior precision equals

$$\frac{1}{(s')^2} = \frac{1}{200^2} + \frac{4}{40^2} = .002525$$

The posterior variance equals $(s')^2 = \frac{1}{002525} = 396.0$ The posterior standard deviation equals $s' = \sqrt{396.0} = 19.9$. The posterior mean equals

$$m' = \frac{\frac{1}{200^2}}{.002525} \times 1000 + \frac{\frac{4}{40^2}}{.002525} \times 970 = 970.3$$

The posterior distribution of μ is *normal*$(970.3, .19.9^2)$.

(b) The 95 % credible interval for μ is is (931.3, 1009.3)

(c) The posterior distribution of θ is *normal*$(1392, 16.6)$

(d) The 95 % credible interval for θ is (1360,1425)

Chapter 11: Comparing Bayesian and Frequentist Inferences for Mean

11.1 (a) posterior precision

$$\frac{1}{(s')^2} = \frac{1}{10^2} + \frac{10}{2^2} = 2.51$$

The posterior variance $(s')^2 = \frac{1}{2.51} = .3984$ and the posterior standard deviation $s' = \sqrt{.3984} = .63119$. The posterior mean

$$m' = \frac{\frac{1}{10^2}}{2.51} \times 75 + \frac{\frac{10}{2^2}}{2.51} \times 79.410 = 79.39$$

The posterior distribution is *normal*$(79.39, .63119^2)$

(b) The 95 % Bayesian credible interval is (78.16,80.63)

(c) Calculate the posterior probability of the null hypothesis.

$$P(\mu \geq 80) = .168$$

This is greater than the level of significance, so we cannot reject the null hypothesis.

11.3 (a) posterior precision

$$\frac{1}{(s')^2} = \frac{1}{80^2} + \frac{25}{80^2} = .0040625$$

The posterior variance $(s')^2 = \frac{1}{.0040625} = 246.154$ and the posterior standard deviation $s' = \sqrt{246.154} = 15.69$. The posterior mean

$$m' = \frac{\frac{1}{80^2}}{.0040625} \times 325 + \frac{\frac{25}{80^2}}{.0040625} \times 401.96 = 399\,.$$

The posterior distribution is *normal*$(399, 15.69^2)$.

(b) The 95 % Bayesian credible interval is (368,429).

(c) We observe that the null value (350) lies outside the credible interval, so we reject the null hypothesis $H_0 : \mu = 350$ at the 5% level of significance. We can conclude that $\mu \neq 350$.

(d) We calculate the posterior probability of the null hypothesis.

$$P(\mu \leq 350) = .0009$$

This is less than the level of significance, so we reject the null hypothesis and conclude that $\mu > 350$.

Chapter 12: Bayesian Inference for Difference between Means

12.1 (a) The posterior distribution of μ_A is *normal*$(119.4, 1.888^2)$, the posterior distribution of μ_B is *normal*$(122.7, 1.888^2)$, and they are independent.

(b) The posterior distribution of $\mu_d = \mu_A - \mu_B$ is *normal*$(-3.271, 2.671^2)$.

(c) The 95% credible interval for $\mu_A - \mu_B$ is $(-8.506, 1.965)$.

(d) We note that the null value 0 lies inside the credible interval. Hence we cannot reject the null hypothesis.

12.3 (a) The posterior distribution of μ_1 is *normal*$(14.96, .3778^2)$, the posterior distribution of μ_2 is *normal*$(15.55, .3778^2)$, and they are independent.

(b) The posterior distribution of $\mu_d = \mu_1 - \mu_1$ is *normal*$(-.5847, .5343^2)$.

(c) The 95% credible interval for $\mu_1 - \mu_1$ is $(-1.632, .462)$.

(d) We note that the null value 0 lies inside the credible interval. Hence we cannot reject the null hypothesis.

12.5 (a) The posterior distribution of μ_1 is *normal*$(10.283, .816^2)$, the posterior distribution of μ_2 is *normal*$(9.186, .756^2)$, and they are independent.

(b) The posterior distribution of $\mu_d = \mu_1 - \mu_2$ is *normal*$(1.097, 1.113^2)$.

(c) The 95% credible interval for $\mu_1 - \mu_2$ is $(-1.08, 3.28)$.

(d) We calculate the posterior probability of the null hypothesis

$$P(\mu_1 - \mu_2 \leq 0) = .162$$

This is greater than the level of significance, so we cannot reject the null hypothesis.

12.7 (a) The posterior distribution of μ_1 is *normal*$(1.51999, .000009444^2)$.

(b) The posterior distribution of μ_2 is *normal*$(1.52001, .000009444^2)$.

(c) The posterior distribution of $\mu_d = \mu_1 - \mu_2$ is *normal*$(-.00002, .000013^2)$.

(d) A 95% credible interval for μ_d is $(-.000046, .000006)$.

(e) We observe that the null value 0 lies inside the credible interval so we cannot reject the null hypothesis.

12.9 (a) The posterior distribution of π_1 is *beta* $(172, 144)$.

(b) The posterior distribution of π_2 is *beta* $(138, 83)$.

(c) The approximate posterior distribution of $\pi_1 - \pi_2$ is *normal*$(-.080, .0429^2)$

(d) The 99 % Bayesian credible interval for $\pi_1 - \pi_2$ is $(-.190, .031)$.

(e) We observe that the null value 0 lies inside the credible interval, so we cannot reject the null hypothesis that the proportions of New Zealand women who are in paid employment are equal for the two age groups.

12.11 (a) The posterior distribution of π_1 is *beta* $(70, 246)$.

(b) The posterior distribution of π_2 is *beta* $(115, 106)$.

(c) The approximate posterior distribution of $\pi_1 - \pi_2$ is *normal*$(-.299, .0408^2)$.

(d) We calculate the posterior probability of the null hypothesis:

$$P(\pi_1 - \pi_2 \geq 0) = P(Z \geq 7.31) = .0000$$

We reject the null hypothesis and conclude that the proportion of New Zealand women in the younger group who have been married before age 22 is less than the proportion of New Zealand women in the older group who have been married before age 22.

12.13 (a) The posterior distribution of π_1 is *beta* $(137, 179)$.

(b) The posterior distribution of π_2 is *beta* $(136, 85)$.

(c) The approximate posterior distribution of $\pi_1 - \pi_2$ is *normal*$(-.182, .0429^2)$.

(d) The 99 % Bayesian credible interval for $\pi_1 - \pi_2$ is $(-.292, -.071)$.

(e) We calculate the posterior probability of the null hypothesis:

$$P(\pi_1 - \pi_2 \geq 0) = P(Z \geq 4.238) = .0000$$

We reject the null hypothesis and conclude that the proportion of New Zealand women in the younger group who have given birth before age 25 is less than the proportion of New Zealand women in the older group who have given birth before age 25.

12.15 (a) The posterior distribution of μ_d after the first experiment is *normal* $(-2.970, .680^2)$. We will use that as the prior for the second experiment. The posterior distribution of μ_d given both data from both experiments is *normal*$(-3.69, .33^2)$.

(b) The 95 % credible interval is $(-4.34, -3.04)$. We note that this is considerably shorter than when we analyzed the experiments separately.

(c) We observe that the null value 0 lies outside the credible interval, so we reject the null hypothesis. The ^{13}C measurements are different depending on which chamber the fluid goes to. This means that the ^{13}C test could be used to determine to which chamber the fluid went.

Chapter 13: Bayesian Inference for Simple Linear Regression

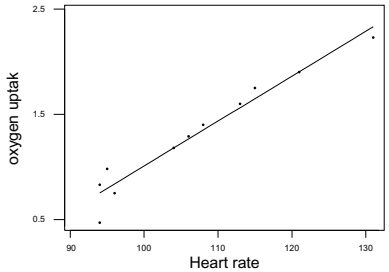
13.1 (b) The least squares slope

$$B = \frac{145.610 - 107 \times 1.30727}{11584.1 - 107^2} = 0.0426514$$

The least squares y-intercept equals

$$A_0 = 1.30727 - .0426514 \times 107 = -3.25643$$

(c) The scatterplot of oxygen uptake on heart rate with least squares line



(d) The estimated variance about the least squares line is found by taking the sum of squares of residuals and dividing by $n-2$ and equals $\hat{\sigma}^2 = .1303^2$

(e) The *likelihood* of β is proportional to a *normal*$(B, \frac{\sigma^2}{SS_x})$ where B is the least squares slope and $SS_x = n \times (\overline{x^2} - \bar{x}^2) = 1486$ and $\sigma^2 = .13^2$. The prior for β is *normal*$(0, 1^2)$. The posterior precision will be

$$\frac{1}{(s')^2} = \frac{1}{1^2} + \frac{SS_x}{.13^2} = 87930\,,$$

the posterior variance will be $(s')^2 = \frac{1}{87930} = .000011373$ and the posterior mean is

$$m' = \frac{\frac{1}{1^2}}{87930} \times 0 + \frac{\frac{SS_x}{.13^2}}{87930} \times .0426514 = .0426509$$

The posterior distribution of β is *normal*$(.0426, .00337^2)$

(f) A 95 % Bayesian credible interval for β is (.036, .049).

(g) We observe that the null value 0 lies outside the credible interval, so we reject the null hypothesis.

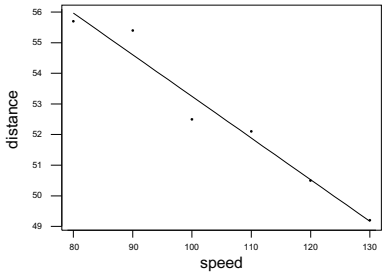
13.3 (b) The least squares slope

$$B = \frac{5479.83 - 105 \times 52.5667}{11316.7 - 105^2} = -0.136000$$

The least squares y-intercept equals

$$A_0 = 52.5667 - -0.136000 \times 105 = 66.8467$$

(c) The scatterplot of distance on speed with least squares line



(d) The estimated variance about the least squares line is found by taking the sum of squares of residuals and dividing by $n - 2$ and equals $\hat{\sigma}^2 = .571256^2$.

(e) The *likelihood* of β is proportional to a *normal*$(B, \frac{\sigma^2}{SS_x})$ where B is the least squares slope and $SS_x = n \times (\overline{x^2} - \bar{x}^2) = 1750$ and $\sigma^2 = .57^2$. The prior for β is *normal*$(0, 1^2)$. The posterior precision will be

$$\frac{1}{(s')^2} = \frac{1}{1^2} + \frac{SS_x}{.57^2} = 5387.27$$

the posterior variance $(s')^2 = \frac{1}{5387.27} = .000185623$ and the posterior mean is

$$m' = \frac{\frac{1}{1^2}}{5387.27} \times 0 + \frac{\frac{SS_x}{.57^2}}{5387.27} \times (-0.136000) = -.135975$$

The posterior distribution of β is *normal*$(-.136, .0136^2)$.

(f) A 95 % Bayesian credible interval for β is $(-.163, -0.109)$.

(g) We calculate the posterior probability of the null hypothesis.

$$P(\beta \geq 0) = P(Z \geq 9.98) = .0000$$

This is less than the level of significance, so we reject the null hypothesis and conclude that $\beta < 0$.

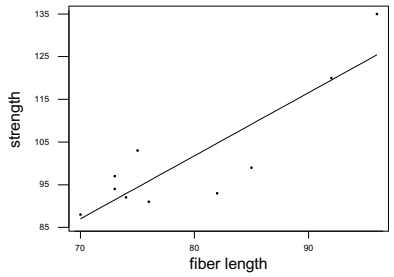
13.5 (b) The least squares slope

$$B = \frac{8159.3 - 79.6 \times 101.2}{6406.4 - 79.6^2} = 1.47751\,.$$

The least squares y-intercept equals

$$A_0 = 101.2 - 1.47751 \times 79.6 = -16.4095\,.$$

(c) The scatterplot of score on cans with least squares line



(d) The estimated variance about the least squares line is found by taking the sum of squares of residuals and dividing by $n-2$ and equals $\hat{\sigma}^2 = 7.667^2$.

(e) The *likelihood* of β is proportional to a *normal*$(B, \frac{\sigma^2}{SS_x})$ where B is the least squares slope and $SS_x = n \times (\overline{x^2} - \bar{x}^2) = 702.400$ and $\sigma^2 = 7.7^2$. The prior for β is *normal*$(0, 10^2)$. The posterior precision will be

$$\frac{1}{(s')^2} = \frac{1}{10^2} + \frac{SS_x}{7.7^2} = 11.8569$$

the posterior variance $(s')^2 = \frac{1}{11.8569} = .0843394$ and the posterior mean is

$$m' = \frac{\frac{1}{10^2}}{11.8569} \times 0 + \frac{\frac{SS_x}{7.7^2}}{11.8569} \times 1.47751 = 1.47626$$

The posterior distribution of β is *normal*$(1.48, .29^2)$

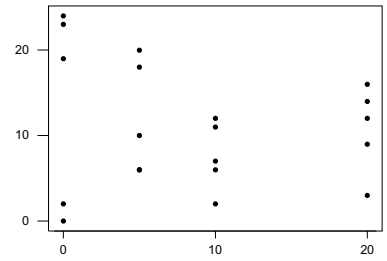
(f) A 95 % Bayesian credible interval for β is (.91, 2.05)

(g) We calculate the posterior probability of the null hypothesis:

$$P(\beta \leq 0) = P(Z \leq -5.08) = .0000\,.$$

This is less than the level of significance, so we reject the null hypothesis and conclude $\beta > 0$.

13.7 (a) The scatterplot of number of ryegrass plants on the weevil infestation rate where the ryegrass was infected with endophyte. Doesn't look linear. Has dip at infestation rate of 10.

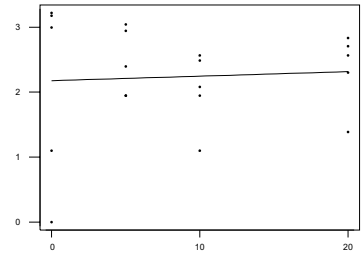


(c) The least squares slope

$$B = \frac{19.9517 - 8.75 \times 2.23694}{131.250 - 8.75^2} = .00691966\,.$$

The least squares y-intercept equals

$$A_0 = 2.23694 - .00691966 \times 8.75 = 2.17640\,.$$



(d) $\hat{\sigma}^2 = .850111^2$

(e) The *likelihood* of β is proportional to a *normal*$(B, \frac{\sigma^2}{SS_x})$ where B is the least squares slope and $SS_x = n \times (\overline{x^2} - \bar{x}^2) = 1093.75$ and $\sigma^2 = .850111^2$. The prior for β is *normal*$(0, 1^2)$. The posterior precision will be

$$\frac{1}{(s')^2} = \frac{1}{1^2} + \frac{SS_x}{.850111^2} = 1514.45$$

the posterior variance $(s')^2 = \frac{1}{1514.45} = .000660307$ and the posterior mean is

$$m' = \frac{\frac{1}{1^2}}{1514.45} \times 0 + \frac{\frac{SS_x}{.311469^2}}{1514.45} \times .00691966 = .00691509\,.$$

The posterior distribution of β is *normal*$(.0069, .0257^2)$

13.9 (a) To find the posterior distribution of $\beta_1 - \beta_2$, we take the difference between the posterior means, and add the posterior variances since they are independent. The posterior distribution of $\beta_1 - \beta_2$ is *normal*$(1.012, .032^2)$.

(b) The 95 % credible interval for $\beta_1 - \beta_2$ is (.948,1.075).

(c) We calculate the posterior probability of the null hypothesis:

$$P(\beta_1 - \beta_2 \leq 0) = P(Z \leq -31) = .0000\,.$$

This is less than the level of significance, so we reject the null hypothesis and conclude $\beta_1 - \beta_2 > 0$. This means that infection by endophyte offers ryegrass some protection against weevils.

Chapter 14: Robust Bayesian Methods

14.1 (a) The posterior $g_0(\pi|y = 10)$ is *binomial*$(7 + 10, 13 + 190)$.

(b) The posterior $g_1(\pi|y = 10)$ is *binomial*$(1 + 10, 1 + 190)$.

(c) The posterior probability $P(I = 0|y = 10) = .163$.

(d) The marginal posterior $g(\pi|y = 10) = .163 \times g_0(\pi|y = 10) + .837 \times g_1(\pi|y = 10)$. This is a mixture of the two beta posteriors where the proportions are the posterior probabilities of I.

14.3 (a) The posterior $g_0(\mu|y_1, \ldots, y_6)$ is *normal*$(1.10270, .000377964)$.

(b) The posterior $g_1(\mu|y_1, \ldots, y_6)$ is *normal*$(1.10314, .000407909)$.

(c) The posterior probability $P(I = 0|y_1, \ldots, y_6) = .123$.

(d) The marginal posterior $g(\mu|y_1, \ldots, y_6) = .123 \times g_0(\pi|y = 10) + .877 \times g_1(\pi|y_1, \ldots, y_6)$. This is a mixture of the two normal posteriors where the proportions are the posterior probabilities of I.

References

1. Bayes, T. (1763), An essay towards solving a problem in the doctrine of chances, *Philo. Trans. of the Roy. Soc.* 53, 370-418. (Reprinted in Biometrika 45 (1958), 293-315.
2. Berry, D. (1996), *Statistics: A Bayesian Perspective*, Duxbury, Belmont, CA.
3. Box, G., and Tiao, G. (1992), *Bayesian Inference in Statistical Analysis*, Wiley Classics Library, John Wiley & Sons, New York.
4. De Finetti, B. (1991), *Theory of Probability, Volume 1 and Volume 2*, Wiley Classics Library, John Wiley & Sons, New York.
5. Jaynes, E. T. (1995), *Probability Theory: The Logic of Science*, http:/bayes.wustl.edu/.
6. Lee, P. (1989), *Bayesian Statistics: An Introduction*, Edward Arnold, London.
7. O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics, Vol. 2B, Bayesian Inference*, Edward Arnold, London.
8. Press, S. J. (1989), *Bayesian Statistics: Principles, Models, and Applications*, John Wiley & Sons, New York.
9. Wald, A. (1950), *Statistical Decision Functions*, Wiley, New York.
10. Bolstad, W.M., Hunt, L.A., and McWhirter, J.L. (2001), Sex, Drugs, and Rock & Roll Survey in a First-Year Service Course in Statistics, *The American Statistician* Vol. 55, 145-149.

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2

11. McBride, G., Till, D., Ryan, T., Ball, A., Lewis, G., Palmer, S., and Weinstein, P. (2002)., Freshwater Microbiology Research Programme Pathogen Occurrence and Human Risk Assessment Analysis.

12. Petchet, Fiona (2000), Radiocarbon dating fish bone from the Houhora archeological site, New Zealand, *Archeol. Oceania* 35, 104-115.

13. Petchet, Fiona and Higham, T. (2000), Bone diagenesis and radiocarbon dating of fish bones at the Shag River mouth site, New Zealand, *Journal of Archeological Science* 27, 135-150.

14. Stigler, S.M. (1977), Do robust estimators work with real data? (With discussion.), *The Annals of Statistics* 5, 1055-1098.

15. Stuiver, M., Reimer, P.J., Braziunas, S. (1998), High precision radiocarbon age calibration for terrestial and marine samples, *Radiocarbon* 40, 1127-1151.

Index

[0]*Introduction to Bayesian Statistics*. By William M. Bolstad
ISBN 0-471-27020-2