Part III

Estimation

Topic 12

Overview of Estimation

> *Inference is the problem of turning data into knowledge, where knowledge often is expressed in terms of entities that are not present in the data per se but are present in models that one uses to interpret the data. Statistical rigor is necessary to justify the inferential leap from data to knowledge, and many difficulties arise in attempting to bring statistical principles to bear on massive data. Overlooking this foundation may yield results that are, at best, not useful, or harmful at worst. In any discussion of massive data and inference, it is essential to be aware that it is quite possible to turn data into something resembling knowledge when actually it is not. Moreover, it can be quite difficult to know that this has happened.* - page 2, Frontiers in Massive Data Analysis by the National Research Council, 2013.

The balance of this book is devoted to developing formal procedures of statistical inference. In this introduction to inference, we will be basing our analysis on the premise that the data have been collected according to carefully planned procedures informed by the appropriate probability models. We will focus our presentation on parametric estimation and hypothesis testing based on a given family of probability models chosen in line with the science under investigation and with the data collection procedures.

12.1 Introduction

In the simplest possible terms, the goal of **estimation theory** is to answer the question:

What is that number?

What is the length, the reaction rate, the fraction displaying a particular behavior, the temperature, the kinetic energy, the Michaelis constant, the speed of light, mutation rate, the melting point, the probability that the dominant allele is expressed, the elasticity, the force, the mass, the free energy, the mean number of offspring, the focal length, mean lifetime, the slope and intercept of a line?

The next step is to perform an experiment that is well designed to estimate one (or more) numbers. However, before we can embark on such a design, we must be informed by the principles of estimation in order to have an understanding of the properties of a good estimator and to present our uncertainties concerning the estimate. Statistics has provided two distinct approaches - typically called **classical** or frequentist and **Bayesian**. We shall give an overview of both approaches. However, this text will emphasize the classical approach.

Let's begin with a definition:

Definition 12.1. *A* **statistic** *is a function of the data that does not depend on any unknown parameter.*

Example 12.2. *We have to this point, seen a variety of statistics.*

- *sample mean,* $\bar{x}$

- *sample variance,* s^2
- *sample standard deviation,* s
- *sample median, sample quartiles* Q_1, Q_3*, percentiles and other quantiles*
- *standardized scores* $(x_i - \bar{x})/s$
- *order statistics* $x_{(1)}, x_{(2)}, \dots x_{(n)}$*, including sample maximum and minimum*
- *sample moments*

$$\overline{x^m} = \frac{1}{n}\sum_{k=1}^{n} x_k^m, \quad m = 1, 2, 3, \dots.$$

Here, we will look at a particular type of **parameter estimation**, in which we consider $X = (X_1, \dots, X_n)$, independent random variables chosen according to one of a family of probabilities P_θ where θ is element from the **parameter space** Θ. Based on our analysis, we choose an **estimator** $\hat{\theta}(X)$. If the data $\mathbf{x}$ takes on the values $x_1, x_2, \dots, x_n$, then

$$\hat{\theta}(x_1, x_2, \dots, x_n)$$

is called the **estimate** of θ. Thus we have three closely related objects,

1. θ - the parameter, an element of the parameter space Θ. This is a number or a vector.
2. $\hat{\theta}(x_1, x_2, \dots, x_n)$ - the estimate. This again is a number or a vector obtained by evaluating the estimator on the data $\mathbf{x} = (x_1, x_2, \dots, x_n)$.
3. $\hat{\theta}(X_1, \dots, X_n)$ - the estimator. This is a random variable. We will analyze the distribution of this random variable to decide how well it performs in estimating θ.

The first of these three objects is a number. The second is a statistic. The third can be analyzed and its properties described using the theory of probability. Keeping the relationship among these three objects in mind is essential in understanding the fundamental issues in statistical estimation.

Example 12.3. *For Bernoulli trials* $X = (X_1, \dots, X_n)$*, each* $X_i, i = 1, \dots, n$ *can take only two values 0 and 1. We have*

1. *p, a single parameter, the probability of success, with parameter space* $[0, 1]$*. This is the probability that a single Bernoulli takes on the value 1.*
2. $\hat{p}(x_1, \dots, x_n)$ *is the* **sample proportion** *of successes in the data set.*
3. $\hat{p}(X_1, \dots, X_n)$*, the sample mean of the random variables*

$$\hat{p}(X_1, \dots, X_n) = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}S_n$$

is an estimator of p. In this case the X_i *are Bernoulli trials. Consequently, we can give the distribution of this estimator because* S_n *is a binomial random variable.*

Example 12.4. *Given pairs of observations* $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ *that display a general linear pattern, we use* **ordinary least squares regressn** *for*

1. *parameters - the slope* β *and intercept* α *of the regression line. So, the parameter space is* $\mathbb{R}^2$*, pairs of real numbers.*

2. *They are estimated using the statistics $\hat{\beta}$ and $\hat{\alpha}$ in the equations*

$$\hat{\beta}(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\mathrm{var}(\mathbf{x})}, \quad \bar{y} = \hat{\alpha}(\mathbf{x}, \mathbf{y}) + \hat{\beta}(\mathbf{x}, \mathbf{y})\bar{x}.$$

3. *Later, when we consider statistical inference for linear regression, we will analyze the distribution of the estimators.*

Exercise 12.5. *Let $X = (X_1, \ldots, X_n)$ be independent uniform random variables on the interval $[0, \theta]$ with θ unknown. Give some estimators of θ from the statistics above.*

12.2 Classical Statistics

In classical statistics, the **state of nature** is assumed to be fixed, but unknown to us. Thus, one goal of estimation is to determine which of the P_θ is the source of the data. The **estimate** is a statistic

$$\hat{\theta} : \text{data} \to \Theta.$$

Introduction to estimation in the classical approach to statistics is based on two fundamental questions:

- How do we determine estimators?
- How do we evaluate estimators?

We can ask if this estimator in any way systematically under or over estimate the parameter, if it has large or small variance, and how does it compare to a notion of best possible estimator. How easy is it to determine and to compute and how does the procedure improve with increased sample size?

The raw material for our analysis of any estimator is the **distribution of the random variables** that underlie the data under any possible value θ of the parameter. To simplify language, we shall use the term **density function** to refer to both continuous and discrete random variables. Thus, to each parameter value $\theta \in \Theta$, there exists a density function which we denote

$$\mathbf{f}_X(\mathbf{x}|\theta).$$

We focus on experimental designs based on a **simple random sample**. To be more precise, the data are assumed to be a sample from a sequence of random variables

$$X_1(\omega), \ldots, X_n(\omega),$$

drawn from a family of distributions having common density $f_X(x|\theta)$ where the parameter value θ is unknown and must be estimated. Because the random variables are independent, the **joint density** is the product of the **marginal densities**.

$$\mathbf{f}_X(\mathbf{x}|\theta) = \prod_{k=i}^{n} f_X(x_k|\theta) = f_X(x_1|\theta) f_X(x_2|\theta) \cdots f_X(x_n|\theta).$$

In this circumstance, the data $\mathbf{x}$ are known and the parameter θ is unknown. Thus, we write the density function as

$$L(\theta|\mathbf{x}) = \mathbf{f}_X(\mathbf{x}|\theta)$$

and call L the **likelihood function**.

Because the algebra and calculus of the likelihood function are a bit unfamiliar, we will look at several examples.

Example 12.6 (Parametric families of densities)**.**

1. *For Bernoulli trials with a known number of trials n but unknown success probability parameter p has joint density*

$$\begin{aligned}\mathbf{f}_X(\mathbf{x}|p) &= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n} = p^{\sum_{k=1}^n x_k}(1-p)^{\sum_{k=1}^n (1-x_k)} \\ &= p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k} = p^{n\bar{x}}(1-p)^{n(1-\bar{x})}\end{aligned}$$

2. *Normal random variables with known variance σ_0 but unknown mean μ has joint density*

$$\begin{aligned}\mathbf{f}_X(\mathbf{x}|\mu) &= \frac{1}{\sigma_0\sqrt{2\pi}}\exp\left(-\frac{(x_1-\mu)^2}{2\sigma_0^2}\right)\cdot\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left(-\frac{(x_2-\mu)^2}{2\sigma_0^2}\right)\cdots\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left(-\frac{(x_n-\mu)^2}{2\sigma_0^2}\right) \\ &= \frac{1}{(\sigma_0\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma_0^2}\sum_{k=1}^n(x_k-\mu)^2\right)\end{aligned}$$

3. *Normal random variables with unknown mean μ and variance σ has density*

$$\mathbf{f}_X(\mathbf{x}|\mu,\sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^n(x_k-\mu)^2\right).$$

4. *Beta random variables with parameters α and β has joint desity*

$$\begin{aligned}\mathbf{f}_X(x|\alpha,\beta) &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n(x_1\cdot x_2\cdots x_n)^{\alpha-1}((1-x_1)\cdot(1-x_2)\cdots(1-x_n))^{\beta-1} \\ &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n\left(\prod_{i=1}^n x_i\right)^{\alpha-1}\left(\prod_{i=1}^n(1-x_i)\right)^{\beta-1}\end{aligned}$$

Exercise 12.7. *Give the likelihood function for n observations of independent $\Gamma(\alpha,\beta)$ random variables.*

The choice of a **point estimator** $\hat{\theta}$ is often the first step. For the next three topics, we consider two approaches for determining estimators - method of moments and maximum likelihood. In between the introduction of these two estimation procedures, we will develop analyses of the quality of the estimator. With this in view, we will provide methods for approximating the bias and the variance of the estimators. Typically, this information is, in part, summarized though what is know as an **interval estimator**. This is a procedure that determines a subset of the parameter space with high probability that it contains the real state of nature. We see this most frequently in the use of **confidence intervals**.

12.3 Bayesian Statistics

For a few tosses of a coin always that always turn up tails, the estimate $\hat{p} = 0$ for the probability of heads did not seem reasonable to Thomas Bayes. He wanted a way to place our uncertainly of the value for p into the procedure for estimation.

Today, the **Bayesian approach to statistics** takes into account not only the density

$$\mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)$$

for the data collected for any given experiment but also external information to determine a **prior density** π on the parameter space Θ. Thus, in this approach, both the parameter and the data are modeled as random. Estimation is based on Bayes formula. We now want to take Bayes theorem, previously derived for a finite partition and obtain a formula useful for Bayesian estimation. The first step will yield a formula based on a discrete mixture. We will then need to introduce the notion of a continuous mixture to give us the final formula, (12.4).

r

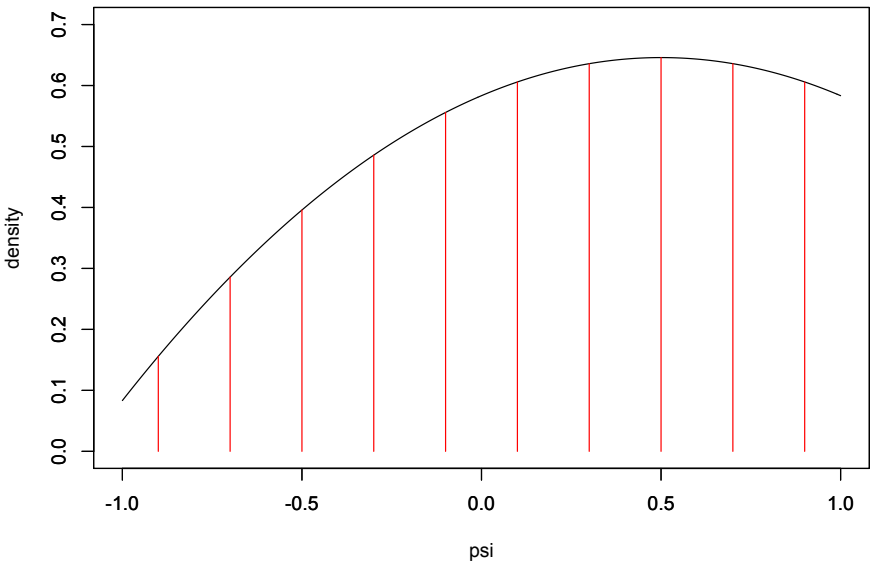

Figure 12.1: Plot of density $\pi(\psi)$ in black for continuous mixture and the approximating discrete mixture with weights $\tilde{\pi}(\tilde{\psi})$ proportional to the heights of the red vertical lines.

Let $\tilde{\Theta}$ be a random variable having the given prior density π. In the case in which both $\tilde{\Theta}$ and the data take on only a finite set of values, $\tilde{\Theta}$ is a discrete random variable and π is a mass function

$$\pi\{\psi\} = P\{\tilde{\Theta} = \psi\}.$$

Let $C_\psi = \{\tilde{\Theta} = \psi\}$ be the event that $\tilde{\Theta}$ takes on the value ψ and $A = \{X = \mathbf{x}\}$ be the values taken on by the data. Then $\{C_\psi, \psi \in \Theta\}$ form a partition of the probability space. Bayes formula is

$$P(C_\theta|A) = \frac{P(A|C_\theta)P(C_\theta)}{\sum_\psi P(A|C_\psi)P(C_\psi)} \quad \text{or} \tag{12.1}$$

$$f_{\Theta|X}(\theta|\mathbf{x}) = P\{\tilde{\Theta} = \theta|X = \mathbf{x}\} = \frac{P\{X=\mathbf{x}|\tilde{\Theta}=\theta\}P\{\tilde{\Theta}=\theta\}}{\sum_\psi P\{X=\mathbf{x}|\tilde{\Theta}=\psi\}P\{\tilde{\Theta}=\psi\}} = \frac{\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi\{\theta\}}{\sum_\psi \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi\{\psi\}}.$$

Given data $\mathbf{x}$, the function of θ, $f_{\Theta|X}(\theta|\mathbf{x}) = P\{\tilde{\Theta} = \theta|X = \mathbf{x}\}$ is called the **posterior density**.

Remark 12.8. *As we learned in tte section on Random Variables and Distribution functions, the expression*

$$\sum_\psi \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi\{\psi\}$$

is the **mixture** *of the densities $\mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)$ for ψ in the finite set with weights $\pi(\psi)$. Typically the parameter space Θ is continuous and so we want to use the density π of a continuous random variable. To determine an expression for a continuous mixture, we will be guided by the ideas used deriving the formula for the expected value for a continuous random variable based on the formula for a discrete random variable. Beginning with the property*

$$\int_\theta \pi(\psi)d\psi = 1 \quad \textit{we have, for a Riemann sum,} \quad \sum_{\tilde{\psi}} \pi(\tilde{\psi})\Delta\psi \approx 1.$$

Now, write

$$\tilde{\pi}\{\tilde{\psi}\} = \pi(\tilde{\psi})\Delta\psi. \tag{12.2}$$

Then $\tilde{\pi}$ is (approximately) the density function for a discrete random variable. If we take a mixture of $\mathbf{f}_{X|\Theta}(\mathbf{x}|\tilde{\psi})$ with weights $\tilde{\pi}(\tilde{\psi})$, we have the mixture density

$$\sum_{\tilde{\psi}} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\tilde{\pi}\{\tilde{\psi}\} = \sum_{\tilde{\psi}} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\tilde{\psi})\Delta\psi.$$

This last sum is a Riemann sum and so taking limits as $\Delta\psi \to 0$, we have that the Riemann sum converges to the definite integral. This gives us the continuous mixture

$$\mathbf{f}_X(\mathbf{x}) = \int_\Theta \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi. \tag{12.3}$$

Exercise 12.9. *Show that the expression (12.3) for $\mathbf{f}_X(\mathbf{x})$ is a valid density function*

Returning to the expression (12.1), substituting (12.2), we have in the interval from θ to $\theta + \Delta\theta$,

$$f_{\Theta|X}(\theta|\mathbf{x})\Delta\theta = \frac{\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)\Delta\theta}{\sum_\psi \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\Delta\psi}.$$

After dividing by $\Delta\theta$ and taking a limit as $\Delta\psi \to 0$, we have, for π, a density for a continuous random variable, that the sum in Bayes formula becomes an integral for a continuous mixture,

$$f_{\Theta|X}(\theta|\mathbf{x}) = \frac{\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)}{\int \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi} \tag{12.4}$$

Sometimes we shall write (12.4) as

$$f_{\Theta|X}(\theta|\mathbf{x}) = c(\mathbf{x})\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)$$

where $c(\mathbf{x})$, the reciprocal of continuous mixture (12.3) in the denominator of (12.4), is the value necessary so that the integral of the posterior density $f_{\Theta|X}(\theta|\mathbf{x})$ with respect to θ equals 1. We might also write

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto \mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta) \tag{12.5}$$

where $c(\mathbf{x})$ is the constant of proportionality.

Estimation, e.g., point and interval estimates, in the Bayesian approach is based on the data and an analysis using the posterior density. For example, one way to estimate θ is to use the mean of the posterior distribution, or more briefly, the **posterior mean**,

$$\hat{\theta}(\mathbf{x}) = E[\theta|\mathbf{x}] = \int \theta f_{\Theta|X}(\theta|\mathbf{x})\, d\theta.$$

Example 12.10. *As suggested in the original question of Thomas Bayes, we will make independent flips of a biased coin and use a Bayesian approach to make some inference for the probability of heads. We first need to set a prior distribution for $\tilde{P}$. The beta family $Beta(\alpha, \beta)$ of distributions takes values in the interval $[0, 1]$ and provides a convenient prior density π. Thus,*

$$\pi(p) = c_{\alpha,\beta}p^{(\alpha-1)}(1-p)^{(\beta-1)}, \quad 0 < p < 1.$$

Any density on the interval $[0, 1]$ that can be written a a power of p times a power of $1 - p$ times a constant chosen so that

$$1 = \int_0^1 \pi(p)\, dp$$

is a member of the beta family. This distribution has

$$\text{mean } \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \text{variance } \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{12.6}$$

Thus, the mean is the ratio of α and $\alpha+\beta$. If the two parameters are each multiplied by a factor of k, then the mean does not change. However, the variance is reduced by a factor close to k. The prior gives a sense of our prior knowledge of the mean through the ratio of α to $\alpha+\beta$ and our uncertainly through the size of α and β

If we perform n Bernoulli trials, $\mathbf{x}=(x_1,\dots,x_n)$, then the joint density

$$\mathbf{f}_X(\mathbf{x}|p)=p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k}.$$

Thus the posterior distribution of the parameter $\tilde{P}$ given the data $\mathbf{x}$, using (12.5), we have.

$$\begin{aligned} f_{\tilde{P}|X}(p|\mathbf{x}) \propto \mathbf{f}_{X|\tilde{P}}(\mathbf{x}|p)\pi(p) &= p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k}\cdot c_{\alpha,\beta}p^{(\alpha-1)}(1-p)^{(\beta-1)}. \\ &= c_{\alpha,\beta}p^{\alpha+\sum_{k=1}^n x_k-1}(1-p)^{\beta+n-\sum_{k=1}^n x_k-1}. \end{aligned}$$

Consequently, the posterior distribution is also from the beta family with parameters

$$\alpha+\sum_{k=1}^n x_k \quad \text{and} \quad \beta+n-\sum_{k=1}^n x_k=\beta+\sum_{k=1}^n(1-x_k).$$

$$\alpha+\#\text{ successes} \quad \text{and} \quad \beta+\#\text{ failures}.$$

Notice that the posterior mean can be written as

$$\begin{aligned} \frac{\alpha+\sum_{k=1}^n x_k}{\alpha+\beta+n} &= \frac{\alpha}{\alpha+\beta+n}+\frac{\sum_{k=1}^n x_k}{\alpha+\beta+n} \\ &= \frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{\alpha+\beta+n}+\frac{1}{n}\sum_{k=1}^n x_k\cdot\frac{n}{\alpha+\beta+n} \\ &= \frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{\alpha+\beta+n}+\bar{x}\cdot\frac{n}{\alpha+\beta+n}. \end{aligned}$$

This expression allow us to see that the posterior mean can be expresses as a weighted average $\alpha/(\alpha+\beta)$ from the prior mean and $\bar{x}$, the sample mean from the data. The relative weights are

$$\alpha+\beta \textit{ from the prior} \quad \textit{and} \quad n\textit{, the number of observations.}$$

Thus, if the number of observations n is small compared to $\alpha+\beta$, then most of the weight is placed on the prior mean $\alpha/(\alpha+\beta)$. As the number of observations n increase, then

$$\frac{n}{\alpha+\beta+n}$$

increases towards 1. The weight result in a shift the posterior mean away from the prior mean and towards the sample mean $\bar{x}$.

This brings forward two central issues in the use of the Bayesian approach to estimation.

- *If the number of observations is small, then the estimate relies heavily on the quality of the choice of the prior distribution π. Thus, an unreliable choice for π leads to an unreliable estimate.*
- *As the number of observations increases, the estimate relies less and less on the prior distribution. In this circumstance, the prior may simply be playing the roll of a catalyst that allows the machinery of the Bayesian methodology to proceed.*

Exercise 12.11. *Show that this answer is equivalent to having α heads and β tails in the data set before actually flipping coins.*

Example 12.12. *If we flip a coin $n = 14$ times with 8 heads, then the classical estimate of the success probability p is 8/14=4/7. For a Bayesian analysis with a beta prior distribution, using (12.6) we have a beta posterior distribution with the following parameters.*

prior				data		posterior			
α	β	mean	variance	heads	tails	α	β	mean	variance
6	6	1/2	1/52=0.0192	8	6	14	12	14/(12+14)=7/13	168/18542=0.0092
9	3	3/4	3/208=0.0144	8	6	17	9	17/(17+9) =17/26	153/18252=0.0083
3	9	1/4	3/208=0.0144	8	6	11	15	11/(15+11)=11/26	165/18542=0.0090

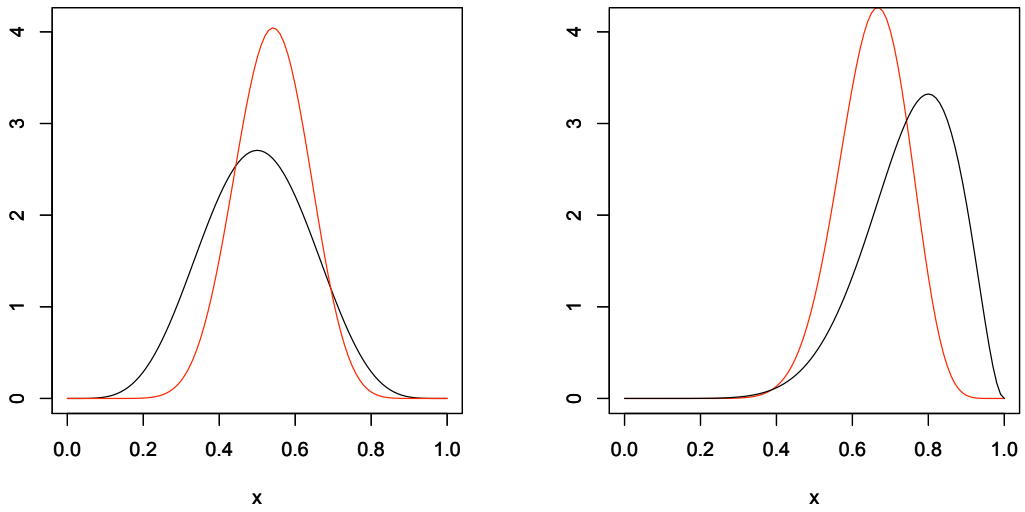

Figure 12.2: Example of prior (black) and posterior (red) densities based on 14 coin flips, 8 heads and 6 tails. Left panel: Prior is $Beta(6, 6)$, Right panel: Prior is $Beta(9, 3)$. Note how the peak is narrowed. This shows that the posterior variance is smaller than the prior variance. In addition, the peal moves from the prior towards $\hat{p} = 4/7$, the sample proportion of the number of heads.

In his original example, Bayes chose was the uniform distribution ($\alpha = \beta = 1$) for his prior. In this case the posterior mean is

$$\frac{1}{2+n}\left(1+\sum_{k=1}^{n} x_k\right).$$

For the example above

prior				data		posterior			
α	β	mean	variance	heads	tails	α	β	mean	variance
1	1	1/2	1/12=0.0833	8	6	9	7	9/(9+7)=9/16	63/4352=0.0144

The Bayesian approach is amenable to **sequential updating**. For example, if we collect **independent** data in three batches, say $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, then the density for the entire data set $\mathbf{x}$ can be written

$$f_{X|\Theta}(\theta|\mathbf{x}) = f_{X_3|\Theta}(\mathbf{x}_3|\theta) \cdot f_{X_2|\Theta}(\mathbf{x}_2|\theta) \cdot f_{X_1|\Theta}(\mathbf{x}_1|\theta).$$

To set the notation, write

- $X = (X_1, X_2, X_3)$ for the the sequential sets of random variables associated to the observations,
- $f_{\Theta|X_1}(\theta|\mathbf{x}_1)$ for the posterior density based on the data $\mathbf{x}_1$, and

- $f_{\Theta|X_1,X_2}(\theta|\mathbf{x}_1,\mathbf{x}_2)$ for the posterior density based on the data $(\mathbf{x}_1,\mathbf{x}_2)$,

Then, the posterior density

$$\begin{aligned} f_{\Theta|X}(\theta|\mathbf{x}) &\propto f_{X|\Theta}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3|\theta)\pi(\theta) = f_{X_3|\Theta}(\mathbf{x}_3|\theta)\cdot f_{X_2|\Theta}(\mathbf{x}_2|\theta)\cdot f_{X_1|\Theta}(\mathbf{x}_1|\theta)\pi(\theta) \\ &= f_{X_3|\Theta}(\mathbf{x}_3|\theta)\cdot f_{X_2|\Theta}(\mathbf{x}_2|\theta)\cdot f_{\Theta|X_1}(\theta|\mathbf{x}_1) \\ &= f_{X_3|\Theta}(\mathbf{x}_3|\theta)\cdot f_{\Theta|X_1,X_2}(\theta|\mathbf{x}_1,\mathbf{x}_2) \end{aligned}$$

Thus,

- The posterior density $f_{\Theta|X_1}(\theta|\mathbf{x}_1) \propto f_{X_1|\Theta}(\mathbf{x}_1|\theta)\pi(\theta)$ serves are the prior density for $(\mathbf{x}_2,\mathbf{x}_3)$.
- The posterior density $f_{\Theta|X_1,X_2}(\theta|\mathbf{x}_1,\mathbf{x}_2) \propto f_{X_2|\Theta}(\mathbf{x}_2|\theta)\cdot f_{\Theta|X_1}(\theta|\mathbf{x}_1)$ serves are the prior density for $\mathbf{x}_3$.

Of course, this strategy can be used for any number of sequential updates.

Example 12.13. *Extending the example on the original use of Bayes estimation, the observations $\mathbf{x}_1$ consist of 8 heads and 6 tails, .$\mathbf{x}_2$ consist of 8 heads and 4 tails, and $\mathbf{x}_3$ consist of 9 heads and 4 tails, We start with a $Beta(1,1)$ prior and so all of the subsequent posteriors will have a $Beta(\alpha,\beta)$ distribution. The data and the parameter values are shown in the table below.*

prior			data			posterior		
	α	β	observations	heads	tails		α	β
$(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$	1	1	$\mathbf{x}_1$	8	6	$\mathbf{x}_1$	9	7
$(\mathbf{x}_2,\mathbf{x}_3)$	9	7	$\mathbf{x}_2$	8	4	$(\mathbf{x}_1,\mathbf{x}_2)$	17	11
$\mathbf{x}_3$	17	11	$\mathbf{x}_3$	9	4	$(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$	26	15

Notice that the posterior for one stage serves as the prior for the next.

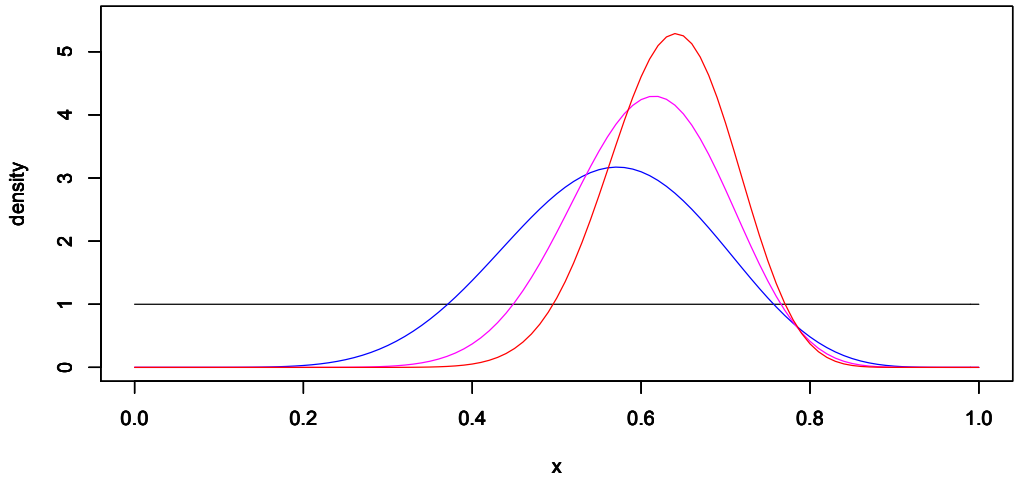

Figure 12.3: Bayesian updating. With a beta distributed prior and Bernoulli trials, the posterior densities are also beta distributed. Successive updates are shown in blue, magenta, and red.

Example 12.14. Reliability engineering *emphasizes dependability of a product by assessing the ability of a system or component to function. We introduce the Bayesian perspective to reliability through a the consideration of the reliability of simple devise. Our analysis is based on an extension the ideas of Bernoulli trials example above.*

A devise consists of two independent units. Let A_i, $i = 1, 2$ be the event that the i-th unit is operating appropriately and define the probability $p_i = P(A_i)$. Then, the devise works with probability $p_1p_2 = P(A_1 \cap A_2)$. For each unit, we

place independent $Beta(\alpha_i, \beta_i)$ prior distributions. Next we test n_i units of type i. Repeating the steps in a previous exercise, we find that y_i units are functioning, then the posterior distribution are also in the beta family,

$$Beta(\alpha_1 + y_1, \beta_1 + n_1 - y_1) \quad \textit{and} \quad Beta(\alpha_2 + y_2, \beta_2 + n_2 - y_2),$$

respectively. This results in a joint posterior density

$$f_{P_1,P_2|Y_1,Y_2}(p_1, p_2|y_1, y_2) = c(\alpha_1, \beta_1, n_1)c(\alpha_2, \beta_2, n_2)\, p_1^{\alpha_1+x_1}(1-p_1)^{\beta_1+n_1-y_1} \cdot p_2^{\alpha_2+x_2}(1-p_2)^{\beta_1+n_2-y_2}.$$

To find the posterior distribution of $p = p_1p_2$ that the devise functions, we integrate to find the cumulative distribution function.

$$F_{P|Y_1,Y_2}(p|y_1, y_2) = \int\int_{\{p_1p_2 \le p\}} f_{P_1,P_2|Y_1,Y_2}(p_1, p_2|y_1, y_2)\, dp_2 dp_1.$$

We can also simulate using the `rbeta` *command to estimate values for this distribution functions.*

To provide a concrete example, assume a uniform prior ($\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1$) and test twenty units of each type ($n_1 = n_2 = 20$). If 15 and 17 of the devises work ($y_1 = 15, y_2 = 17$), then the posteriors distributions are

$$Beta(16, 6) \quad \textit{and} \quad Beta(18, 4),$$

We simulate this in R *to find the distribution of the posterior probability of $p = p_1p_2$.*

```
> p1<-rbeta(10000,16,6);p2<-rbeta(10000,18,4)
> p<-p1*p2
```

We then give a table of deciles for the posterior distribution function and present a histogram.

```
> data.frame(quantile(p,d))
     quantile.p..d.
0%        0.2825593
10%       0.4660896
20%       0.5094321
30%       0.5422747
40%       0.5712765
50%       0.5968341
60%       0.6209610
70%       0.6477835
80%       0.6776208
90%       0.7187307
100%      0.9234834
> hist(p)
```

The posterior density $f_{P|Y_1,Y_2}(p|y_1, y_2)$ is non-negative throughout the interval from 0 to 1, but is very small for values near 0 and 1. Indeed, none of the 10,000 simulations give a posterior probability below 0.282 or above 0.923. We could take the mean of the simulated sample as a point estimate $\hat{p}$ for p.

```
> mean(p);sd(p)
[1] 0.5935356
[1] 0.09661065
```

This is very close to the means from the beta distributions.

$$E\hat{p} = Ep_1p_2 = Ep_1Ep_2 = \frac{16}{22} \cdot \frac{18}{22} = \frac{72}{121} = 0.595.$$

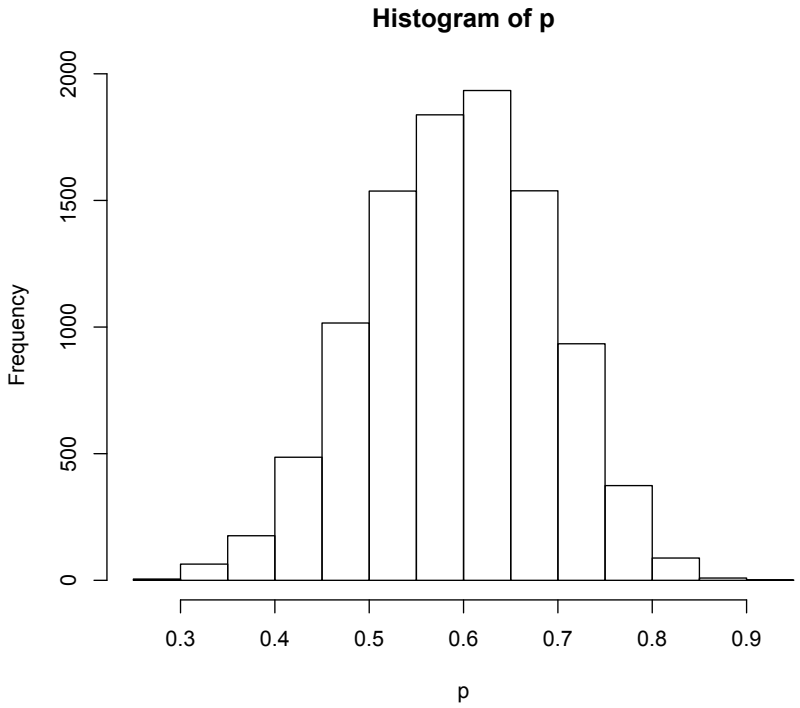

Figure 12.4: Histogram of simulated posterior distribution of the reliability $p = p_1 p_2$ where p_1 and p_2 have independent beta distributions based on the prior distributions and the data.

Exercise 12.15. *The simulation variance is also indicated. Compare this answer with the answer given by the delta method.*

Example 12.16. *Suppose that the prior density is a normal random variable with mean θ_0 and variance $1/\lambda$. This way of giving the variance may seem unusual, but we will see that λ is a measure of* **information**. *Thus, low variance means high information. Our data $\mathbf{x}$ are a realization of independent normal random variables with unknown mean θ. We shall choose the variance to be 1 to set a scale for the size of the variation in the measurements that yield the data $\mathbf{x}$. We will present this example omitting some of the algebraic steps to focus on the central ideas.*

The prior density is

$$\pi(\theta) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(\theta - \theta_0)^2\right).$$

We rewrite the density for the data to empathize the difference between the parameter θ for the mean and the $\bar{x}$, the sample mean.

$$\begin{aligned} f_{X|\Theta}(\mathbf{x}|\theta) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right). \end{aligned}$$

The posterior density is proportional to the product $f_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)$, Because the posterior is a function of θ, we

need only keep track of the terms which involve θ. Consequently, we write the posterior density as

$$f_{\Theta|X}(\theta|\mathbf{x}) = c(\mathbf{x}) \exp\left(-\frac{1}{2}(n(\theta-\bar{x})^2 + \lambda(\theta-\theta_0)^2)\right)$$
$$= \tilde{c}(\mathbf{x}) \exp(-\frac{n+\lambda}{2}(\theta-\theta_1(\mathbf{x}))^2).$$

where

$$\theta_1(\mathbf{x}) = \frac{\lambda}{\lambda+n}\theta_0 + \frac{n}{\lambda+n}\bar{x}. \tag{12.7}$$

Notice that the posterior distribution is normal with mean $\theta_1(\mathbf{x})$ that results from the weighted average with relative weights

$$\lambda \quad \textit{from the information from the prior} \quad \textit{and} \quad n \quad \textit{from the data.}$$

The variance in inversely proportional to the total information $\lambda + n$. Thus, if n is small compared to λ, then $\theta_1(\mathbf{x})$ is near θ_0. If n is large compared to λ, $\theta_1(\mathbf{x})$ is near $\bar{x}$.

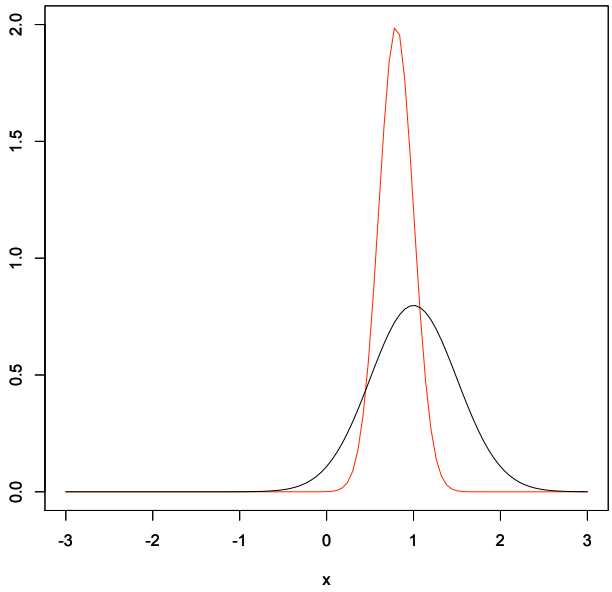

Figure 12.5: Example of prior (black) and posterior (red) densities for a normal prior distribution and normally distributed data. In this figure the prior density is $N(1, 1/2)$. Thus, $\theta_0 = 1$ and $\lambda = 2$. Here the data consist of 3 observations having sample mean $\bar{x} = 2/3$. Thus, the posterior mean from equation (12.7) is $\theta_1(\mathbf{x}) = 4/5$ and the variance is 1/(2+3) = 1/5.

Exercise 12.17. *Fill in the steps in the derivation of the posterior density in the example above.*

Exercise 12.18. *Use sequential updating for the normal family of distribution in the example above. The prior and the summary statistics are below.*

prior			data				posterior		
	μ	σ^2	observations	n	$\bar{x}$	s^2		μ	σ^2
$(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$	0	4	$\mathbf{x}_1$	6	1.216	1.042	$\mathbf{x}_1$		
$(\mathbf{x}_2, \mathbf{x}_3)$			$\mathbf{x}_2$	3	1.911	0.432	$(\mathbf{x}_1, \mathbf{x}_2)$		
$\mathbf{x}_3$			$\mathbf{x}_3$	3	0.811	0.348	$(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$		

Show that the answer is the same if we aggregate the data first.

For these two examples, we see that the prior distribution and the posterior distribution are members of the same parameterized family of distributions, namely the beta family and the normal family. In these two cases, we say that the prior density and the density of the data form a **conjugate pair**. In the case of coin tosses, we find that the beta and the Bernoulli families form a conjugate pair. In Example 12.11, we learn that the normal density is conjugate to itself.

Typically, the computation of the posterior density is much more computationally intensive that what was shown in the two examples above. The choice of conjugate pairs is enticing because the posterior density is a determined from a simple algebraic computation.

Bayesian statistics is seeing increasing use in the sciences, including the life sciences, as we see the explosive increase in the amount of data. For example, using a classical approach, mutation rates estimated from genetic sequence data are, due to the paucity of mutation events, often not very precise. However, we now have many data sets that can be synthesized to create a prior distribution for mutation rates and will lead to estimates for this and other parameters of interest that will have much smaller variance than under the classical approach.

Exercise 12.19. *Show that the gamma family of distributions is a conjugate prior for the Poisson family of distributions. Give the posterior mean based on n observations.*

12.4 Answers to Selected Exercises

12.5. Double the average, $2\bar{X}$. Take the maximum value of the data, $\max_{1\le i\le n} x_i$. Double the difference of the maximum and the minimum, $2(\max_{1\le i\le n} x_i - \min_{1\le i\le n} x_i)$.

12.7. The density of a gamma random variable

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Thus, for n observations

$$\begin{aligned} L(\theta|\mathbf{x}) &= f(x_1|\alpha,\beta) f(x_2|\alpha,\beta)\cdots f(x_n|\alpha,\beta) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} x_1^{\alpha-1} e^{-\beta x_1} \frac{\beta^\alpha}{\Gamma(\alpha)} x_2^{\alpha-1} e^{-\beta x_2} \cdots \frac{\beta^\alpha}{\Gamma(\alpha)} x_n^{\alpha-1} e^{-\beta x_n} \\ &= \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} (x_1 x_2 \cdots x_n)^{\alpha-1} e^{-\beta(x_1+x_2+\cdots+x_n)} \end{aligned}$$

12.9. We need to verify that the density $\mathbf{f}$ is a non-negative function and that the integral over the sample space is 1. Note that both $\mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)$ and $\pi(\psi)$ and thus their product is positive. Consequently,

$$\mathbf{f}_X(x) = \int_\Theta \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi \ge 0 \quad \text{for all } x.$$

Next, we reverse the order of the double integral,

$$\int_{\mathbb{R}^n} \mathbf{f}_X(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^n} \left(\int_\Theta \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi \right) dx = \int_\Theta \left(\int_{\mathbb{R}^n} \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\, dx \right) \pi(\psi) d\psi.$$

Because $\mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)$, is a density function, the integral inside the parentheses is 1. Now use the fact that π is a probability density,

$$\int_{\mathbb{R}^n} \mathbf{f}_X(\mathbf{x}) d\mathbf{x} = \int_\Theta \pi(\psi) d\psi = 1.$$