# Chapter 8

# Estimators

In probability theory we aim to deduce the likelihood of different outcomes based on known probability distributions. Statistics is the inverse problem. We want to infer unknown probability distributions from outcomes we observe. In this chapter we begin to explore these ideas.

## 8.1 The Estimation Problem

Let's start off by clarifying the nature of the problem we want to study.

### 8.1.1 Definitions

From a pedagogical point of view, it's best to start our study of statistics in a setting where successive observations are both independent and identically distributed. Later we'll extend to dependent data. In an IID setting, the fundamental problem of econometrics and statistics is this:

**Problem 8.1.1** We observe independent $Z$-valued draws $\mathbf{z}_1, \ldots, \mathbf{z}_N$ from a common but unknown distribution $P \in \mathscr{P}$, where $\mathscr{P}$ is a class of distributions on $Z$. We wish to infer some features of $P$ from this sample.

The set $\mathscr{P}$ is the universe of distributions we are willing to consider. It can be anything, including the set of all distributions on the outcome space. One of the main tasks of economic theory is to restrict $\mathscr{P}$ and thereby narrow down the set of distributions we have to search over.

**Example 8.1.1** Benhabib et al. (2015) study the wealth distribution in a model with idiosyncratic capital income risk. The model predicts that the wealth distribution will

have a Pareto right tail. Intuitively, this is because stochastic returns on assets are multiplicative, so positive shocks are self-reinforcing, and a sequence of positive shocks can lead to very high wealth relative to the median.

Let's set out some notation:

- $Z$ is the **outcome space** where each **observation** $\mathbf{z}_n$ takes values.
- $\mathbf{z}_\mathcal{D}$ denotes the **sample** or **data set** $(\mathbf{z}_1, \ldots, \mathbf{z}_N)$.
- $Z_\mathcal{D} := \times_{n=1}^N Z$ is the **sample space** in which $\mathbf{z}_\mathcal{D}$ takes values.
- $P_\mathcal{D} := \mathcal{L}(\mathbf{z}_\mathcal{D})$ is the **joint distribution of the sample**.

Although calling $Z_\mathcal{D}$ the sample space follows standard nomenclature, we are overloading the terminology first used on page 79. (Unfortunately, statisticians and probabilists don't always coordinate on terminology.)

As above, $P = \mathcal{L}(\mathbf{z}_n)$ is the common distribution of the observations. For as long as we continue to focus on IID data, the joint distribution $P_\mathcal{D}$ will be the $N$th product of $P$, as defined on page 130.

**Example 8.1.2** Let $x_1, \ldots, x_N$ be observations of labor income from a given population. We model $x_1, \ldots, x_N$ as IID draws from a common univariate distribution $P$. In our terminology, $x_n$ is an observation, the outcome space $Z$ is $\mathbb{R}$, the sample space $Z_\mathcal{D}$ is $\mathbb{R}^N$, and the sample $\mathbf{z}_\mathcal{D}$ is the vector $(x_1, \ldots, x_N)$. Features of $P$ that we might wish to learn about include

- the mean and higher moments of $P$,
- measures of dispersion, such as the variance, or properties of tails,
- the median and other quantiles, and
- $P$ itself, or the density of $P$ if it exists.

In example 8.1.2, the individual observations are scalar. In other scenarios, observations are vector-valued:

**Example 8.1.3** Suppose that we wish to learn about the relationship between profitability and R&D spending within a group of firms. Let $\mathbf{z}_n = (x_n, y_n)$ be an observation of these two quantities at the $n$th firm. We treat the observations as IID across firms, with common marginal distribution $P = \mathcal{L}(\mathbf{z}_n)$. The outcome space is $Z = \mathbb{R}^2$, and the sample space is

$$Z_\mathcal{D} := \mathbb{R}^2 \times \cdots \times \mathbb{R}^2 = \mathbb{R}^{2 \times N}$$

Because the marginal distribution $P$ of the observations is now multivariate, new features of the distribution come in to play, such as

- correlations across coordinates of $P$,
- the variance–covariance matrix associated with $P$, and
- parameters controlling dependence when, say, $P$ is modeled via a copula over certain marginals.

#### 8.1.1.1 Features

We referred above to "features" of $P$ that we might be interested in estimating. Let's define a **feature** of $P$ to be an object of the form

$$\gamma(P) \quad \text{for some} \quad \gamma\colon \mathscr{P} \to S \tag{8.1}$$

The set $S$ is left arbitrary to accommodate all possible features. When $P$ is understood, we'll write $\gamma(P)$ as $\gamma$. Here are some examples for univariate $P$ routinely estimated in econometric studies:

- $\gamma(P) = \int s^k P(\mathrm{d}s)$, the $k$th moment of $P$.
- $\gamma(P) = \inf\{s \in \mathbb{R} : P(-\infty, s] \geqslant 1/2\}$, the median of $P$.
- $\gamma(P) = P$, when we want to estimate $P$ itself.
- $\gamma(P) =$ the density of $P$ when $P$ is absolutely continuous.

If $P$ is multivariate over $\mathbf{z} = (\mathbf{x}, y)$, then one feature of interest is the regression function $f^*(\mathbf{x}) := \mathbb{E}[y \mid \mathbf{x}]$. This function is uniquely determined by $P$ (see §5.2.5).

#### 8.1.1.2 Parametric versus Nonparametric Classes

In the statistical problems listed above, it is assumed that the unknown distribution belongs to some class $\mathscr{P}$. We call $\mathscr{P}$ a **parametric class** if it can be expressed as

$$\mathscr{P} = \{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta} =: \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\} \quad \text{for some } \Theta \subset \mathbb{R}^K$$

In other words, a class of distributions is parametric if it can be indexed by finitely many parameters. A class of distributions is called **nonparametric** if it is not parametric.

**Example 8.1.4** Let $\mathscr{P}$ be the set of all univariate normal distributions with positive variance. Equivalently,

$$\mathscr{P} := \left\{ \text{all } p \text{ s.t. } p(s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(s-\mu)^2}{2\sigma^2} \right\} \text{ for some } \mu \in \mathbb{R},\ \sigma > 0 \right\}$$

The set $\mathscr{P}$ is an example of a parametric class. The parameters are $\boldsymbol{\theta} = (\mu, \sigma)$. A particular choice of the parameters determines (parameterizes) an element of the class.

**Example 8.1.5** If the outcome space $Z$ is finite, containing $J$ elements, then every distribution $P$ on $Z$ can be represented by $J - 1$ parameters (the probability $p_j$ of each outcome but the last) and hence any family $\mathscr{P}$ of distributions on $Z$ is necessarily a parametric class.

**Example 8.1.6** The set of all distributions on $\mathbb{R}$ cannot be parameterized by a finite vector of parameters because the space of distributions is infinite dimensional. Hence $\mathscr{P} :=$ all distributions on $\mathbb{R}$ is a nonparametric class.

**Example 8.1.7** Let $\mathscr{P}$ be the set of all absolutely continuous distributions on $\mathbb{R}$ with finite second moment:

$$\mathscr{P} := \left\{ \text{all } p \colon \mathbb{R} \to \mathbb{R} \text{ s.t. } p \geqslant 0, \int p(s)\, ds = 1, \int s^2 p(s)\, ds < \infty \right\}$$

This set is nonparametric.

Many traditional methods of inference are parametric in nature: The data are assumed to be generated by an unknown element $P_{\boldsymbol{\theta}}$ of a parametric class $\mathscr{P}$. The aim is to estimate $\boldsymbol{\theta}$ using the data. Once we have an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, we can plug this estimate back into the parametric class to obtain an estimate $P_{\hat{\boldsymbol{\theta}}}$ of $P_{\boldsymbol{\theta}}$.

The concept of a feature defined in (8.1) is a generalization of the notion of a parameter or vector of parameters. In parametric settings, the feature $\gamma$ that we are most interested in estimating is the parameter vector $\boldsymbol{\theta}$. If we have a good estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, we can estimate any feature $\gamma = \gamma(P_{\boldsymbol{\theta}})$ via $\hat{\gamma} = \gamma(P_{\hat{\boldsymbol{\theta}}})$.

Following common usage, we will sometimes refer to the $\boldsymbol{\theta}$ associated with the $P_{\boldsymbol{\theta}}$ that generates the data as the **true value** of the parameter vector. However, it's important to remember that this is an assumption, not a "truth." We *assume* that the data are generated by some member of a parametric class. This assumption can be completely false.

In our initial formulation of the estimation problem we considered estimation of features rather than just parameters because it is suboptimal to always restrict ourselves to parametric assumptions on $\mathscr{P}$. We'll discuss the relative merits of parametric and nonparametric estimation more in chapter 14.

### 8.1.2 Statistics and Estimators

A statistic is any observable function of the data. More precisely, a **statistic** is any $\mathscr{B}$-measurable function

$$T \colon Z_{\mathcal{D}} \to S$$

that can be evaluated once the data $\mathbf{z}_{\mathcal{D}}$ are observed. As in (8.1), the set $S$ is left arbitrary to accommodate all the possible features that we might wish to estimate. Sometimes we write $T$ as $T_N$ to emphasize dependence on the sample size. The assumption that a statistic is $\mathscr{B}$-measurable is just a basic regularity condition.

An **estimator** is a statistic used to infer some feature $\gamma(P)$ of an unknown distribution $P$. Thus a statistic becomes an estimator when paired with and compared to a feature of the distribution. Note that there is nothing in the definition of an estimator that implies it will be a sensible estimator of the target feature, let alone a good one.

**Example 8.1.8** If the feature $\gamma$ we wish to infer is the mean of the marginal distribution $P$ of IID data $x_1, \ldots, x_N$, then the most common estimator is the **sample mean**

$$\bar{x}_N := \frac{1}{N} \sum_{n=1}^{N} x_n$$

Formally, $\bar{x}_N$ is the mapping from $Z_{\mathcal{D}} = \mathbb{R}^N$ to $S = \mathbb{R}$ defined by

$$\mathbf{z}_{\mathcal{D}} = (x_1, \ldots, x_N) \mapsto T(x_1, \ldots, x_N) = \frac{1}{N} \sum_{n=1}^{N} x_n \in \mathbb{R}$$

This mapping is regarded as an estimator of the unknown mean $\gamma(P) = \int s P(\mathrm{d}s)$.

**Example 8.1.9** The sample mean is not the only way to estimate the mean. For example, we could also use the so-called **mid-range estimator**

$$m_N := \frac{\min_n x_n + \max_n x_n}{2}$$

Another option is a **truncated sample mean**, where values $x_n$ with $|x_n| \geqslant r$ are truncated for some specified value of $r$. The truncated sample mean is often used to estimate location parameters in heavy tailed distributions.

**Example 8.1.10** Given sample $x_1, \ldots, x_N$, let $y_n$ be the $n$th largest observation of the sample. If $N$ is the odd number $2m + 1$, the **sample median** is defined as $y_{m+1}$. If $N = 2m$, the sample median is $0.5(y_m + y_{m+1})$. For example,

```
julia> median([1, 3, 5])
3.0
julia> median([2, 4, 6, 8])
5.0
```

**Example 8.1.11** Following on from example 8.1.8, a common estimator of the $k$th moment $\int s^k P(\mathrm{d}s)$ of $P$ is the $k$th **sample moment** $\frac{1}{N}\sum_{n=1}^{N} x_n^k$.

**Example 8.1.12** A common estimator of the variance of $P$ is the **sample variance**

$$s_N^2 := \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x}_N)^2 \tag{8.2}$$

The standard deviation is usually estimated using **sample standard deviation**

$$s_N := \sqrt{s_N^2} = \left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x}_N)^2\right]^{1/2} \tag{8.3}$$

In some texts the sample variance is defined as $\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x}_N)^2$ rather than (8.2) in order to produce an unbiased estimator. However, (8.2) fits better with the theory in this text and is perhaps the more common definition.

**Example 8.1.13** Given bivariate data $\mathbf{z}_{\mathcal{D}} = ((x_1, y_1), \ldots, (x_N, y_N))$, the **sample covariance** is the statistic

$$\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x}_N)(y_n - \bar{y}_N) \tag{8.4}$$

The **sample correlation** is the sample covariance divided by the product of the two sample standard deviations. With some rearranging, this becomes

$$\frac{\sum_{n=1}^{N}(x_n - \bar{x}_N)(y_n - \bar{y}_N)}{\sqrt{\sum_{n=1}^{N}(x_n - \bar{x}_N)^2 \sum_{n=1}^{N}(y_n - \bar{y}_N)^2}} \tag{8.5}$$

**Example 8.1.14** In the case where our observations are vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in $\mathbb{R}^K$, the sample mean is the random vector defined by

$$\bar{\mathbf{x}}_N := \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

The variance–covariance matrix defined in (5.16) is most often estimated with the **sample variance–covariance matrix**

$$\hat{\boldsymbol{\Sigma}}_N := \frac{1}{N}\sum_{n=1}^{N}[(\mathbf{x}_n - \bar{\mathbf{x}}_N)(\mathbf{x}_n - \bar{\mathbf{x}}_N)^\intercal] \tag{8.6}$$

### 8.1.3 Empirical Distributions

The **empirical distribution** of a given $Z$-valued sample $\mathbf{z}_\mathcal{D} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ is the discrete distribution on $Z$ that puts equal probability $1/N$ on each sample point $\mathbf{z}_n$. Another way to state this is that $\hat{P}_N$ assigns to each Borel set $B \subset Z$ the number

$$\hat{P}_N(B) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\mathbf{z}_n \in B\} \tag{8.7}$$

This is just the fraction of the sample that falls in $B$. (Refer back to figure 4.2 on page 84 if you wish to see a visual representation.) The expectation of a function $h$ with respect to $\hat{P}_N$ is

$$\int h(\mathbf{s}) \hat{P}_N(\mathrm{d}\mathbf{s}) = \frac{1}{N} \sum_{n=1}^{N} h(\mathbf{z}_n) \tag{8.8}$$

as follows from (5.14) on page 134.

**Example 8.1.15** Let $\mathbf{z}_n$ be the scalar $x_n$. The sample mean can be expressed in terms of the empirical distribution as

$$\bar{x}_N = \frac{1}{N} \sum_{n=1}^{N} x_n = \int s \hat{P}_N(\mathrm{d}s) \tag{8.9}$$

In other words, the sample mean is the mean of the empirical distribution.

The empirical distribution is a statistic, mapping observations $\mathbf{z}_1, \ldots, \mathbf{z}_N$ into $\hat{P}_N = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\mathbf{z}_n \in \cdot\}$, a random element of the set of all distributions on $Z$. If we think of $\mathbf{z}_1, \ldots, \mathbf{z}_N$ as independent draws from common but unknown distribution $P$, then $\hat{P}_N$ becomes an estimator of $P$. In particular, if the feature of $P$ we want to infer is $P$ itself, the simplest natural estimator is the empirical distribution.[1]

With scalar data $x_1, \ldots, x_N$ we can visualize the empirical distribution by plotting its CDF. The CDF of $\hat{P}_N$ will be denoted in what follows by $\hat{F}_N$. Specializing $B$ in (8.7) to $(-\infty, s]$, we get

$$\hat{F}_N(s) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{x_n \leqslant s\} \qquad (s \in \mathbb{R})$$

The CDF $\hat{F}_N$ is called the **empirical cumulative distribution function**, or ECDF, corre-

1. Why not always try to infer $P$ itself? Although it's true that when we know $P$ we can in principle recover any feature $\gamma(P)$, we should also bear in mind the following general principle of inference with limited information: In solving a given problem, try to avoid first solving a more general problem as an intermediate step. Distributions are much more complicated objects than real numbers, so if we care only about the median of a distribution, say, it might be best to try to discover this single value directly rather than trying to infer the entire distribution first.

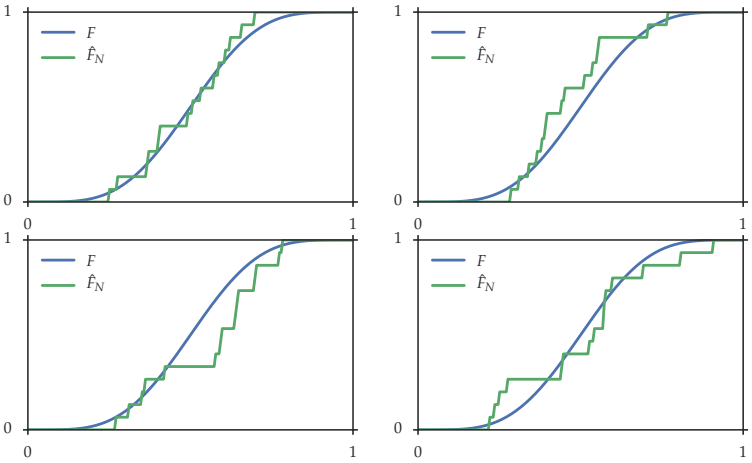


**Figure 8.1** $F$ and four observations of $\hat{F}_N$ when $N = 15$

sponding to the sample. Graphically, $\hat{F}_N$ is a step function, with an upward jump of $1/N$ at each data point. Figure 8.1 shows an example where draws are from the CDF $F$ of the Beta$(5, 5)$ distribution. Each panel shows an observation of the ECDF, which is in turn constructed from $N = 15$ observations from $F$.

#### 8.1.3.1 Convergence

The empirical distribution is asymptotically an excellent estimator of $P$. If $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are IID with common distribution $P$ and $\hat{P}_N$ is the empirical distribution, then, by the law of large numbers,

$$\hat{P}_N(B) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\mathbf{z}_n \in B\} \xrightarrow{p} \mathbb{P}\{\mathbf{z}_n \in B\} = P(B) \tag{8.10}$$

for any Borel set $B$. See, in particular, (6.8) on page 165.

Specializing to the scalar case with $B := (-\infty, s]$, we have $\hat{F}_N(s) \xrightarrow{p} F(s)$ for any $s \in \mathbb{R}$, where $F$ is the CDF of $P$. In fact a much stronger statement is also true. It is sometimes called the fundamental theorem of statistics, or the **Glivenko–Cantelli** theorem:

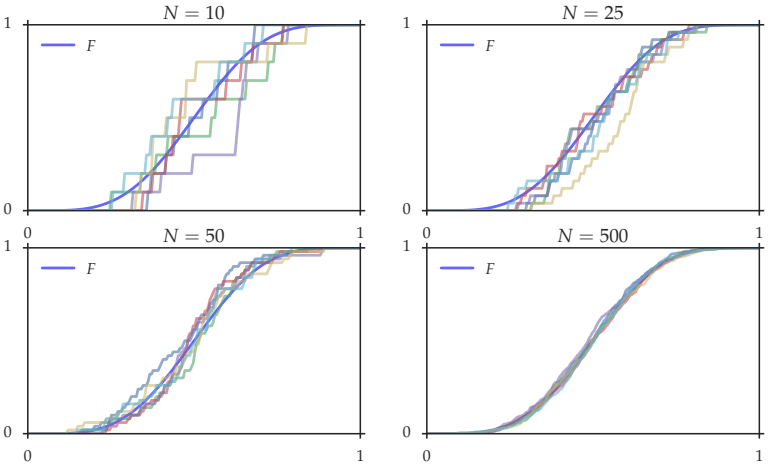


**Figure 8.2** Realizations of $\hat{F}_N$ with four different sample sizes

**Theorem 8.1.1** *Let $x_1, \ldots, x_N$ be* IID *draws from $F$. If $\hat{F}_N$ is the corresponding* ECDF*, then*

$$\|F - \hat{F}_N\|_\infty \xrightarrow{p} 0 \quad \textit{as} \quad N \to \infty \tag{8.11}$$

The norm here is the supremum norm, meaning that

$$\|F - \hat{F}_N\|_\infty := \sup_{s \in \mathbb{R}} |\hat{F}_N(s) - F(s)|$$

Roughly speaking, this is the maximal deviation over the domain. (For a definition of the supremum, see §15.4.1.) In fact the convergence occurs "almost surely," which is a stronger notion than in probability. See theorem 11.4.2 of Dudley (2002).

Figure 8.2 illustrates the convergence in (8.11). Each ECDF $\hat{F}_N$ is based on a single sample $x_1, \ldots, x_N$ of independent draws from $F$, where $F$ is Beta$(5, 5)$. Each subplot shows 8 realizations of the ECDF at a different value of $N$. Across the subplots, the sample size is stepped up from 10 to 500.

Theorem 8.1.1 tells us that, at least in the IID setting, if we have an infinite amount of data, then in the limit we can learn the underlying distribution without having to impose any assumptions at all. Moreover, once we know the distribution $P$, we know any feature $\gamma = \gamma(P)$.

Does this offer a solution to the basic problem of estimation stated on page 213? The answer is no. In practice, we only ever have a finite amount of data. Inference is

about generalization, as discussed in the introduction. There's no need to generalize if we know the full population.

Moreover, when we still have only finitely many observations, the empirical distribution puts all of its mass on these points. In other words, it treats the sample like the unknown distribution, without acknowledging the fact that we have limited information. This can be viewed as an extreme form of overfitting. We'll discuss related ideas in a number of contexts throughout the book.

#### 8.1.3.2 Identification

A class of distributions $\mathscr{P} = \{P_{\boldsymbol{\theta}}\}$ indexed by $\boldsymbol{\theta} \in \Theta$ is called **identifiable** if the map $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$ is one-to-one (see §15.2) on $\Theta$.

**Example 8.1.16** The class of normal densities in example 8.1.4 is identifiable. In particular, it can be shown that if $(\mu_a, \sigma_a)$ and $(\mu_b, \sigma_b)$ are distinct vectors, then the distributions $\mathrm{N}(\mu_a, \sigma_a^2)$ and $\mathrm{N}(\mu_b, \sigma_b^2)$ differ at at least one point.

Identifiability means that the parameter vector associated with the unknown distribution can eventually be distinguished from the data. To see this, suppose that $\{P_{\boldsymbol{\theta}}\}$ is identified on $\Theta$ and nature generates an infinite sequence of observations $\{\mathbf{z}_n\}$ from $P = P_{\boldsymbol{\theta}}$. Let $\boldsymbol{\theta}'$ be any other vector in $\Theta$ and let $P' = P_{\boldsymbol{\theta}'}$. By identifiability, there exists at least one Borel set $B$ with $P(B) \neq P'(B)$. Since the empirical distribution $\hat{P}_N(B)$ converges to $P(B)$, we can in the limit conclude that the data are not generated by $P'$.

## 8.2 Estimation Principles

We started off with the problem of estimating a feature $\gamma = \gamma(P)$ of the unknown distribution $P$ from a sample. We've introduced a number of different statistics aimed at estimating various features. Where do these estimators come from and what, if anything, do they have in common?

### 8.2.1 The Sample Analogue Principle

Most of the estimators defined above can be derived from a simple principle. In the modern statistical literature this is often called the **plug-in method**. In econometrics it goes by a variety of names, including "analogue estimation" (see, in particular, Manski 1986). We'll call it the **sample analogue principle**. The principle is:

$$\text{to estimate } \gamma(P), \quad \text{use } \gamma(\hat{P}_N)$$

Here $\hat{P}_N$ is the empirical distribution constructed from the sample. We replace the unknown distribution $P$ with the observable distribution $\hat{P}_N$ and then evaluate $\gamma(\hat{P}_N)$.

**Example 8.2.1** Let $x_1, \ldots, x_N$ be draws from unknown distribution $P$. Suppose we want to estimate the mean $\gamma(P) := \int sP(\mathrm{d}s)$. The sample analogue principle tells us to replace $P$ with $\hat{P}_N$, which gives

$$\gamma(\hat{P}_N) = \int s\hat{P}_N(\mathrm{d}s) = \bar{x}_N$$

(The last equality is just (8.9).) Thus the sample mean is the estimator of the mean produced by the sample analogue principle.

The $k$th sample moment applies the same principle to estimate the $k$th moment.

**Example 8.2.2** When it exists, the variance of $P$ can be written as

$$\sigma^2 = \gamma(P) = \int \left[ t - \int sP(\mathrm{d}s) \right]^2 P(\mathrm{d}t)$$

Applying the sample analogue principle leads to the estimator

$$\int \left[ t - \int s\hat{P}_N(\mathrm{d}s) \right]^2 \hat{P}_N(\mathrm{d}t) = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x}_N)^2$$

This is precisely the sample variance from example 8.1.12.

The estimators in these two examples are relatively straightforward, but the set of all applications of the sample analogue principle is very broad. As we'll see below, estimation methods that can be obtained as special cases include least squares regression, maximum likelihood, the method of moments, and the generalized method of moments.

#### 8.2.1.1 Best Linear Prediction

Here's an example that illustrates the generality of the sample analogue principle. Recall the best linear prediction problem from §4.1.5.1, where $\alpha$ and $\beta$ are chosen to minimize $\mathbb{E}[(y - \alpha - \beta x)^2]$. Letting $P$ be the distribution of $(x, y)$, we can write the problem as

$$(\alpha^*, \beta^*) = \gamma(P) := \operatorname*{argmin}_{\alpha, \beta \in \mathbb{R}} \int [(t - \alpha - \beta s)^2] P(\mathrm{d}s, \mathrm{d}t) \tag{8.12}$$

There is a corresponding statistical problem when the underlying distribution $P$ is unknown. The problem is to produce the best linear predictor based only on a sample

$\mathbf{z}_{\mathcal{D}} = ((x_1, y_1), \ldots, (x_N, y_N))$ of observations from $P$. We proceed as follows: Given $\mathbf{z}_{\mathcal{D}}$, we form the empirical distribution

$$\hat{P}_N(B) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{(x_n, y_n) \in B\}$$

Applying the sample analogue principle, we plug $\hat{P}_N$ into (8.12), yielding the estimator

$$(\hat{\alpha}_N, \hat{\beta}_N) = \gamma(\hat{P}_N) = \operatorname*{argmin}_{\alpha,\beta \in \mathbb{R}} \sum_{n=1}^{N} [(y_n - \alpha - \beta x_n)^2] \tag{8.13}$$

(The term $\frac{1}{N}$ has been dropped from the right-hand side of (8.13) because it doesn't affect the minimizers.) This is the simple linear bivariate least squares problem. The minimizers are

$$\hat{\beta}_N = \frac{\sum_{n=1}^{N}(x_n - \bar{x}_N)(y_n - \bar{y}_N)}{\sum_{n=1}^{N}(x_n - \bar{x}_N)^2} \quad \text{and} \quad \hat{\alpha}_N = \bar{y}_N - \hat{\beta}_N \bar{x}_N \tag{8.14}$$

(See ex. 8.5.2.) Comparing with (4.23) on page 99, we see that these estimators are themselves the sample analogues of the underlying features $\alpha^*$ and $\beta^*$.

#### 8.2.1.2 Limitations

While the sample analogue principle produces sensible estimators in many situations, there are also instances where it fails completely. Let's look at one example.

Let $\mathscr{P}$ be the set of absolutely continuous distributions on $\mathbb{R}$, so that the distribution $P$ generating the data is assumed to have a density. The density of $P$ can be regarded as the feature

$$\gamma(P) = DP \tag{8.15}$$

Here $DP :=$ the derivative of the CDF $F$ of $P$.

Let $\hat{P}_N$ be the empirical distribution from any given sample. What happens when we plug this empirical distribution into the right-hand side of (8.15)? In other words, what information do we get when we differentiate the function $\hat{F}_N$? Since $\hat{F}_N$ is a step function, the derivative is zero everywhere, except at a finite number of jump points where the derivative is undefined. Thus the density estimate produced by the sample analogue principle is useless. This statement remains true even when $N$ is enormous.

One way to understand this outcome is that when we wish to estimate complex objects with finite samples, we need to combine the data with some kind of **regularization**. Loosely speaking, this means that we penalize complexity in our search for a solution, or impose some kind of "smoothing" a priori. This idea has a long history

in the field of numerical methods, where certain inverse problems are unstable, or "ill-posed," and their solution requires regularization.

Regularization is an essential element of statistics too. It says that we should not grant the empirical distribution equal status with the unknown true distribution. That is, we should regard the empirical distribution only as partial information, and seek to combine it with some form of prior information or external theory. Regularization is one application of prior information: It applies the idea that probability mass most likely falls in places other than just the sample points observed so far. We pick up the topic of regularization again in chapter 14.

### 8.2.2 Empirical Risk Minimization

Let's now look at another inductive principle that combines the sample analogue principle with additional structure. The principle is called empirical risk minimization, or ERM. The terminology and main concepts come from the machine learning literature. Nonetheless, many standard estimators in econometrics are special cases of ERM, including maximum likelihood and least squares. The ideas discussed here help to frame these tools as part of a broader set of methods.

Consider a setting where we observe an input $\mathbf{x} \in \mathbb{R}^K$ to a system, followed by a scalar output $y$. Both are random variables and the joint distribution of $\mathbf{z} := (\mathbf{x}, y)$ is $P$. Our aim is to predict new output values from observed input values. We'll do this by choosing a function $f$ such that $f(\mathbf{x})$ is our prediction of $y$ once $\mathbf{x}$ is observed. In the machine learning literature, $f$ is called a **prediction rule**. In economics, $f$ is called a **strategy** or **policy function**.

Incorrect prediction incurs a loss. The size of this loss is written as $L(y, f(\mathbf{x}))$. The function $L$ is called the **loss function**. Common choices for the loss function include

- the **quadratic loss function** $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$,
- the **absolute deviation loss function** $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$, and
- the **discrete loss function** $L(y, f(\mathbf{x})) = \mathbb{1}\{y \neq f(\mathbf{x})\}$.

Given loss function $L$, we consider choosing $f$ so as to minimize the **prediction risk**, which is defined as the expected loss

$$R(f) := \mathbb{E}\, L(y, f(\mathbf{x})) \tag{8.16}$$

Sometimes $R(f)$ is also called the expected prediction error.[2] The expectation in (8.16) is computed using the joint distribution $P$ of $(\mathbf{x}, y)$. Sometimes we'll write $\mathbb{E}_P\, L(y, f(\mathbf{x}))$ to emphasize this.

2. See, for example, §2.4 of Friedman et al. (2009).

**Example 8.2.3** Let $L$ be the quadratic loss function. As shown in §5.2.5, the minimizer of (8.16) is the regression function, defined at $\mathbf{x}$ by $f^*(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$. In particular, from (5.39) on page 152, for any alternative policy $g$ we have

$$R(g) = R(f^*) + \mathbb{E}[(f^*(\mathbf{x}) - g(\mathbf{x}))^2] \tag{8.17}$$

In other words, with quadratic loss, good prediction equates to choosing $g$ to be close to the regression function, thereby minimizing the second term on the right-hand side of (8.17). In view of (8.17), the term $R(f^*)$ represents a lower bound for prediction risk.

In a statistical setting we are constrained in our ability to minimize risk by the fact that we cannot evaluate $\mathbb{E}_P$, as this requires knowledge of the joint distribution $P$. Suppose, however, that we have access to data $\mathbf{z}_1, \ldots, \mathbf{z}_N$, where each pair $\mathbf{z}_n = (\mathbf{x}_n, y_n)$ is an independent draw from $P$. To make use of the sample, we will apply the sample analogue principle, replacing $P$ in (8.16) with the empirical distribution $\hat{P}_N$. This produces the new objective function

$$R_{\text{emp}}(f) := \mathbb{E}_{\hat{P}_N} L(y, f(\mathbf{x})) = \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(\mathbf{x}_n)) \tag{8.18}$$

This function is called the **empirical risk**. The empirical risk is the prediction risk evaluated under the empirical distribution of the sample rather than the true distribution $P$. Minimizing the empirical risk is called **empirical risk minimization** (ERM).

There is an additional step. When we choose a decision rule the problem that we actually solve is

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}} R_{\text{emp}}(f) \tag{8.19}$$

We have restricted the domain to a class of functions $\mathcal{H}$. This set of functions is called the **hypothesis space**. Choice of $\mathcal{H}$ is structure we impose on the estimation problem.

It might seem at first pass that we should set $\mathcal{H}$ to be the set of all functions $f\colon \mathbb{R}^K \to \mathbb{R}$. After all, if the risk-minimizing function $f^* := \operatorname{argmin}_f R(f)$ is not in $\mathcal{H}$, as visualized in figure 8.3, then the solution to (8.19) is not equal to $f^*$, and we are making a suboptimal choice.

Actually this reasoning is false. In fact we usually want to be quite restrictive in our choice of $\mathcal{H}$. This is because minimizing the empirical risk is not the same as minimizing the prediction risk. Occam's razor comes in to play here: we are solving a complex problem on the basis of limited information. It would be a serious mistake to act as if we had unlimited information. These ideas are explored further in §8.2.3.

**Example 8.2.4** Specializing to scalar $x$ and quadratic loss function $L(y, f(x)) = (y -$

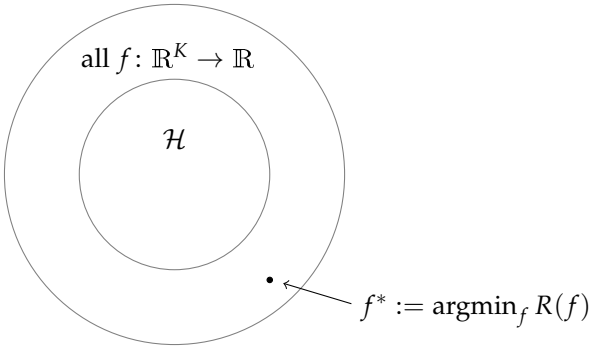


**Figure 8.3** Choosing the hypothesis space

$f(x))^2$, and observing that the term $\frac{1}{N}$ makes no difference to the solution, the ERM problem becomes

$$\min_{f\in\mathcal{H}} \sum_{n=1}^{N} (y_n - f(x_n))^2 \tag{8.20}$$

For obvious reasons, this optimization problem is called the **least squares problem**. If we specialize $\mathcal{H}$ to be the set of affine functions

$$\mathcal{H}_\ell := \{ \text{ all functions of the form } \ell(x) = \alpha + \beta x\} \tag{8.21}$$

then the problem reduces to the **simple linear least squares problem**

$$\min_{\ell\in\mathcal{H}_\ell} \sum_{n=1}^{N} (y_n - \ell(x_n))^2 = \min_{\alpha,\beta} \sum_{n=1}^{N} (y_n - \alpha - \beta x_n)^2 \tag{8.22}$$

This is exactly the same minimization problem we obtained in (8.13). The solutions are given in (8.14).

While we ended up with the same minimization problem in two different ways, the reasoning here is more natural. When we first came to this problem in (8.13) it was by applying the sample analogue principle to the problem of finding best linear predictors. Here we are simply trying to find the best predictor. At the same time we have restricted ourselves to linear approximations in recognition of the fact that we are basing our estimate on limited information. Linearity implements Occam's razor by bounding the complexity of the learning algorithm.

#### 8.2.2.1 Quantile Regression

Quantile regression is used when we want to estimate a quantile of a given distribution based on a set of predictive variables. For example, quantile regression has been used to estimate quantiles of CEO compensation as a function of the market value of their firms (Koenker and Hallock 2001). In this section we show how quantile regression can be regarded as a special case of empirical risk minimization.

To describe the idea, let $F$ be a strictly increasing CDF on $\mathbb{R}$ and let $\tau \in (0,1)$ be given. Recall from §4.2.6 that the $\tau$th quantile of $F$ is the $\xi$ that solves $F(\xi) = \tau$. Although it might not be immediately obvious, the $\tau$th quantile can also be defined as the solution to the optimization problem

$$\min_{\xi \in \mathbb{R}} \mathbb{E} L_\tau(y, \xi) \tag{8.23}$$

where $y$ is a random variable with distribution $F$ and

$$L_\tau(y, \xi) := |(y - \xi)(\tau - \mathbb{1}\{y < \xi\})|$$

Exercise 8.5.6 asks you to confirm that the solution to (8.23) is precisely the $\xi$ that solves $F(\xi) = \tau$.

Suppose now that we want to estimate the $\tau$th quantile of $F$ using some input variable $x$ (e.g., a CEO compensation quantile as a function of firm size). Motivated by (8.23), we can frame the search for a suitable function as a problem of minimizing the prediction risk

$$R(f) := \mathbb{E} L_\tau(y, f(x)) \tag{8.24}$$

where $L_\tau$ is as defined above. If we employ the principle of empirical risk minimization with data set $(x_1, y_1), \ldots, (x_N, y_N)$, we get $\min_{f \in \mathcal{H}} \sum_{n=1}^N L_\tau(y_n, f(x_n))$, where $\mathcal{H}$ is the hypothesis space. When $\mathcal{H} = \mathcal{H}_\ell$ as defined in (8.21), the ERM problem is

$$\min_{\alpha, \beta} \sum_{n=1}^N |(y_n - \alpha - \beta x_n)(\tau - \mathbb{1}\{y_n < \alpha + \beta x_n\})|$$

This is the standard expression for the quantile regression problem.

If $\tau = 0.5$, then the objective function is proportional to $\sum_{n=1}^N |y_n - \alpha - \beta x_n|$. This case is called **median regression** or **least absolute deviation regression**.

### 8.2.3 The Choice of Hypothesis Space

Let's return to the issue of choosing the hypothesis space. We begin with a straightforward result:

**Fact 8.2.1** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hypothesis spaces. If $\hat{f}_i$ is the empirical risk minimizer over $\mathcal{H}_i$ as defined in (8.19), then

$$\mathcal{H}_1 \subset \mathcal{H}_2 \implies R_{\text{emp}}(\hat{f}_1) \geqslant R_{\text{emp}}(\hat{f}_2)$$

In other words, we can always decrease empirical risk by increasing the hypothesis space. The reason that fact 8.2.1 holds is that we are minimizing over a larger set.

At the same time, our actual objective is to minimize the prediction risk. The prediction risk of a given predictor $f$ measures the **out-of-sample fit** of $f$, which is the expected performance of $f$ when confronted with new data. A statistical procedure that takes a given data set and produces a predictor $f$ with low prediction risk (relative to the benchmark $f^*$—see (8.17)) means that we have succeeded in meeting the goal we set at the start of the book: we have *generalized* from data.

Of course, prediction risk is unobservable. Sometimes it is tempting to use empirical risk $R_{\text{emp}}(f)$ as an estimator of $R(f)$. However, empirical risk, which measures **in-sample fit** of $f$, is a downward biased estimator of risk. In particular, as we vary the hypothesis space, prediction risk can rise even when empirical risk is falling.

Let's illustrate this by way of an example, where empirical risk is minimized over progressively larger hypothesis spaces. The model we will consider is one that generates input-output pairs via

$$x \sim U[-1,1] \quad \text{and then} \quad y = \cos(\pi x) + u \quad \text{where} \quad u \sim N(0,1) \tag{8.25}$$

Here $U[-1,1]$ is the uniform distribution on the interval $[-1,1]$. Our hypothesis spaces for predicting $y$ from $x$ will be sets of polynomial functions. To fix notation, let $\mathscr{P}_d$ be the set of all polynomials of degree $d$. That is,

$$\mathscr{P}_d := \{ \text{ all functions } f_d(x) = c_0 x^0 + c_1 x^1 + \cdots c_d x^d \text{ where each } c_i \in \mathbb{R} \}$$

This sequence of hypothesis spaces is increasing, in the sense that

$$\mathscr{P}_1 \subset \mathscr{P}_2 \subset \mathscr{P}_3 \subset \cdots$$

Indeed, if $f$ is a polynomial of degree $d$, then $f$ can be represented as a polynomial of degree $d+1$ just by setting the last coefficient $c_{d+1}$ to zero. The set of linear functions $\mathcal{H}_\ell$ defined in (8.21) is equal to $\mathscr{P}_1$.

If we seek to predict $y$ from $x$ using quadratic loss and the set $\mathscr{P}_d$ as our candidate functions, the risk minimization problem is

$$\min_{f \in \mathscr{P}_d} R(f) \quad \text{where} \quad R(f) = \mathbb{E}[(y - f(x))^2] \tag{8.26}$$

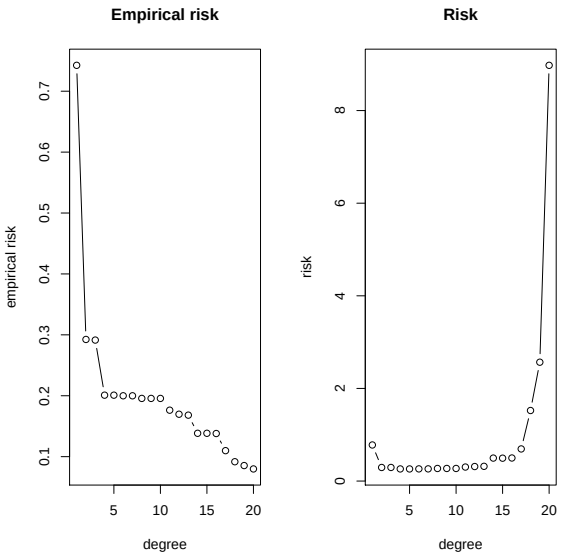


**Figure 8.4** Risk and empirical risk as a function of $d$

while the empirical risk minimization problem is

$$\min_{f \in \mathscr{P}_d} R_{\text{emp}}(f) \quad \text{where} \quad R_{\text{emp}}(f) = \frac{1}{N} \sum_{n=1}^{N} (y_n - f(x_n))^2 \tag{8.27}$$

To illustrate the difference between risk and empirical risk, we first generate $N = 25$ data points from the model (8.25). Taking this as our data set, we then solve (8.27) repeatedly, once for each $d$ in $1, 2, \ldots, 20$. The solution to the $d$th minimization problem is denoted $\hat{f}_d$, and is, by construction, a polynomial of degree $d$. Finally, we compare the risk $R(\hat{f}_d)$ and empirical risk $R_{\text{emp}}(\hat{f}_d)$.[3] The results are in figure 8.4.

As expected, empirical risk falls monotonically with $d$. But the risk decreases slightly and then increases rapidly. For large $d$, the minimizer $\hat{f}_d$ of the empirical risk is associated with very high risk in the sense of large expected loss.

We can get a sense for what is happening by plotting the data and the functions. In figures 8.5–8.8, the $N$ data points are plotted alongside the function $y = \cos(\pi x)$ from the true model (8.25) in black, and fitted polynomial $\hat{f}_d$ in red. The function

3. The risk $R(\hat{f}_d)$ is evaluated by substituting $\hat{f}_d$ into the expression for $R$ in (8.26). You can find the code at `johnstachurski.net/emet.html`.

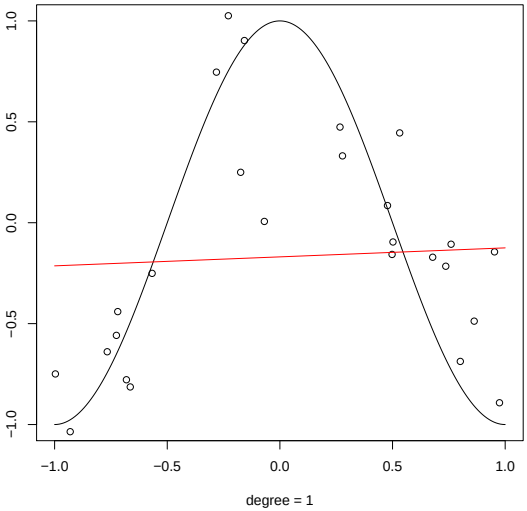


**Figure 8.5** Fitted polynomial, $d = 1$

$y = \cos(\pi x)$ is the risk minimizer, and represents the ideal prediction function. In figure 8.5 we have $d = 1$, and the fitted polynomial $\hat{f}_1$ is the linear regression line. In figures 8.6, 8.7 and 8.8 we have $d = 3$, $d = 11$ and $d = 14$ respectively. The fitted polynomials are $\hat{f}_3$, $\hat{f}_{11}$ and $\hat{f}_{14}$.

On the one hand, when $d = 1$, the hypothesis space $\mathscr{P}_d = \mathscr{P}_1$ is small and no function in this class can effectively fit the underlying model. This is called **underfitting**, and is reflected in the poor fit of the red line to the black line in figure 8.5.

When $d = 3$, the class of functions $\mathscr{P}_d = \mathscr{P}_3$ is considerably larger. Given that the data are relatively noisy, and that we only have 25 observations, the fit of the function is good (figure 8.6). If we look at the risk for $d = 3$ on the right-hand side of figure 8.4, we see that it is lower than for $d = 1$.

On the other hand, if we take $d$ larger, the fit to the underlying model becomes poor and the risk is high. Examining figures 8.7 and 8.8, which correspond to $d = 11$ and $d = 14$, we see that the fitted polynomial has been able to fit the observed data closely, passing near many of the data points. Too much emphasis has been given to this particular realization of the data. When a new input $x$ is drawn, the prediction $\hat{f}_d(x)$ is likely to be a poor predictor of $y$. This situation is called **overfitting**.

To summarize, the choice of $\mathcal{H}$ is central to our ability to generalize from the data. When $\mathcal{H}$ is too small, no function in $\mathcal{H}$ provides a good fit to the regression function.

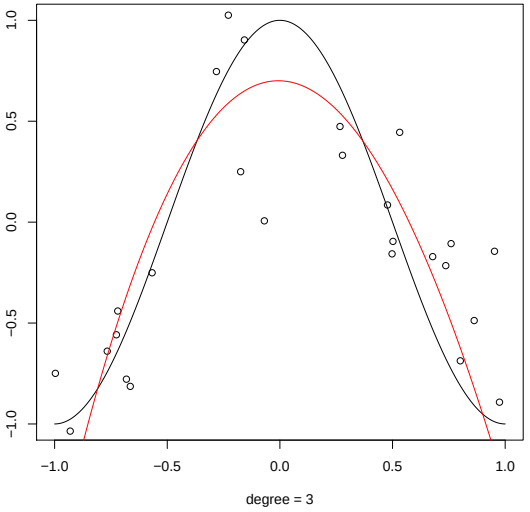


**Figure 8.6** Fitted polynomial, $d = 3$

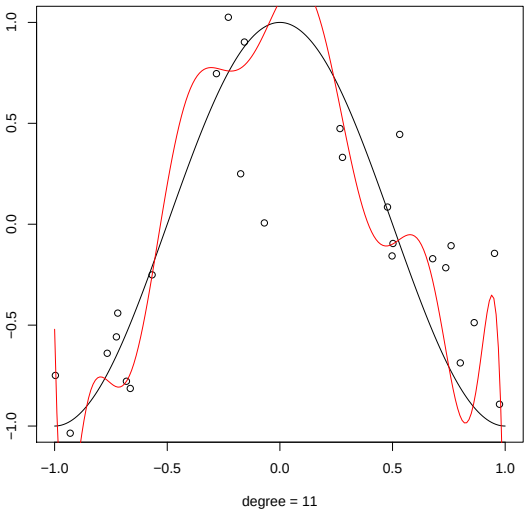


**Figure 8.7** Fitted polynomial, $d = 11$

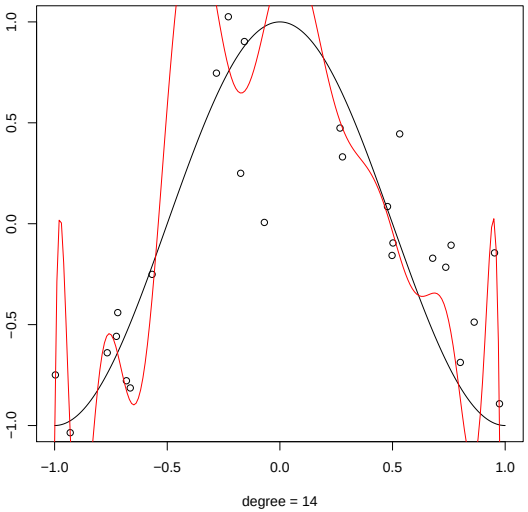


**Figure 8.8** Fitted polynomial, $d = 14$

If this is so, the prediction risk cannot be made small regardless of the data. This can be seen directly from (8.17). Any estimate $\hat{f}$ in $\mathcal{H}$ must satisfy

$$R(\hat{f}) \geqslant R(f^*) + \min_{f \in \mathcal{H}} \mathbb{E}\left[(f^*(\mathbf{x}) - f(\mathbf{x}))^2\right]$$

Conversely, if $\mathcal{H}$ is too large, then we can attain low empirical risk but the prediction risk itself is large.

Of course, in statistical applications we do not have the luxury of knowing the true data-generating process when we choose $\mathcal{H}$. The best scenario is that we have firm theory that guides us to a suitable hypothesis space. The worst scenario is that we have no idea and choose blindly. Once again, the message is that statistical learning equals prior knowledge plus data.

## 8.3 Some Parametric Methods

In this section we review some standard parametric estimation methods, including maximum likelihood, Bayesian estimation, and the generalized method of moments.

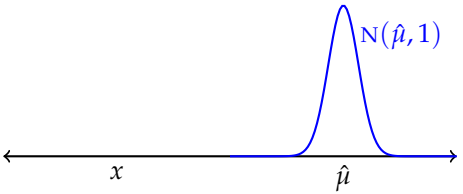


**Figure 8.9** Maximizing the likelihood

### 8.3.1 Maximum Likelihood

One standard approach to deriving estimators in a parametric setting is the **principle of maximum likelihood**. Let's begin with a simple case and then progress to more complex applications.

Suppose that we observe a single draw $x$ from distribution $\mathrm{N}(\mu, 1)$ where $\mu$ is unknown. The variance is known and set to 1 for simplicity. Our task is to guess the value of $\mu$ given the observation $x$. A guess $\hat{\mu}$ of $\mu$ also pins down the distribution $\mathrm{N}(\hat{\mu}, 1)$, so we could equivalently state that our job is to guess the distribution that generated $x$. In guessing this distribution, if we centered it around some number $\hat{\mu}$ much larger than $x$, say, then our observed data point $x$ would be an "unlikely" outcome for this distribution. See figure 8.9. The same logic would apply if we centered the density at a point much smaller than $x$.[4]

In fact, in the absence of any additional information, the most obvious guess would be that the normal density is centered on $x$. To center the density on $x$, we just set $\hat{\mu} = x$.

Maximum likelihood formalizes these steps: The density of $x$ is the density $p(s; \mu)$ of the distribution $\mathrm{N}(\mu, 1)$. Plugging the observed value $x$ into this density gives

$$p(x; \mu) = (2\pi)^{-1/2} \exp\left\{-\frac{(x-\mu)^2}{2}\right\}$$

Think of $p(x; \mu)$ as representing the probability of realizing our sample point $x$. The principle of maximum likelihood suggests that we take as our guess $\hat{\mu}$ of $\mu$ the value that maximizes this probability. It's not hard to show that $\hat{\mu} = x$ is the maximizer. This coincides with our intuition from figure 8.9.

The same principle applies when the data are independent draws $x_1, \ldots, x_N$ from $\mathrm{N}(\mu, 1)$, where $\mu$ is unknown. The joint density of the sample is the product of the

4. The scare quotes on "unlikely" are just a reminder that our distribution is absolutely continuous and hence all individual outcomes $s \in \mathbb{R}$ have probability zero. When we say that an outcome $x$ is unlikely, we mean that there is little probability mass in the neighborhood of $x$.

marginals. Plugging the sample values into the joint density, one then maximizes the joint density with respect to $\mu$:

$$\hat{\mu} := \operatorname*{argmax}_{\mu \in \mathbb{R}} \frac{1}{(2\pi)^{N/2}} \prod_{n=1}^{N} \exp\left\{ -\frac{(x_n - \mu)^2}{2} \right\} \tag{8.28}$$

The maximizer $\hat{\mu}$ is precisely the sample mean of $x_1, \ldots, x_N$ (see ex. 8.5.4).

We can generalize these ideas in several ways. Let's suppose now that the data $x_1, \ldots, x_N$ has joint density $p$ in the sense of (5.8) on page 132. We will assume that $p = p(\cdot\,; \boldsymbol{\theta})$ is a member of a parametric class $\mathscr{P}$ indexed by parameter vector $\boldsymbol{\theta} \in \Theta$. Each choice of $\boldsymbol{\theta}$ pins down a particular density $p = p(\cdot\,; \boldsymbol{\theta})$, but the value of $\boldsymbol{\theta}$ that generated the data is unknown.

In this setting, the **likelihood function** is $p$ evaluated at the sample $x_1, \ldots, x_N$, and regarded as a function of $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) := p(x_1, \ldots, x_N; \boldsymbol{\theta}) \qquad (\boldsymbol{\theta} \in \Theta) \tag{8.29}$$

The principle of maximum likelihood tells us to estimate $\boldsymbol{\theta}$ by maximizing $L(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$. A statistic $\hat{\boldsymbol{\theta}}$ is called a **maximum likelihood estimate** (MLE) of $\boldsymbol{\theta}$ if

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) \tag{8.30}$$

This is equivalent to maximizing the **log likelihood function**

$$\ell(\boldsymbol{\theta}) := \ln(L(\boldsymbol{\theta})) \qquad (\boldsymbol{\theta} \in \Theta)$$

The set of MLEs can in general be a singleton, contain multiple elements or be empty.

In the preceding discussion, $p$ was a density function, but it can be a probability mass function as well. Examples are given below.

To implement maximum likelihood estimation, we need to compute the joint distribution of the data. As we saw in (8.28), when the data points are independent this is easy because the joint density $p$ is the product of the marginals. More generally, if each $x_n$ is drawn independently from fixed arbitrary (marginal) density $p_n(\cdot\,; \boldsymbol{\theta})$ on $\mathbb{R}$, then

$$L(\boldsymbol{\theta}) = \prod_{n=1}^{N} p_n(x_n; \boldsymbol{\theta}) \quad \text{and} \quad \ell(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p_n(x_n; \boldsymbol{\theta}) \tag{8.31}$$

If each data point is multivariate, just replace $x_n$ with $\mathbf{x}_n$.

**Example 8.3.1** Suppose that $x_1, \ldots, x_N$ are IID draws from a normal distribution $\mathrm{N}(\mu, v)$

with $\boldsymbol{\theta} = (\mu, v)$ unknown. The log likelihood function is

$$\ell(\mu, v) = -\frac{N}{2}\ln(2\pi v) - \frac{1}{2}\sum_{n=1}^{N}\frac{(x_n - \mu)^2}{v} \tag{8.32}$$

Joint maximization over $(\mu, v)$ gives the maximum likelihood estimators

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{and} \quad \hat{v} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x}_N)^2 \tag{8.33}$$

Thus, the MLEs of $\mu$ and $v$ are the sample mean and sample variance respectively.

Maximum likelihood estimation is a much celebrated theory of estimation. MLE estimators typically have excellent asymptotic properties. But good finite sample properties are not guaranteed (more about this soon) and the attractive asymptotic theory is dependent on correct specification of the underlying parametric class. Indeed we need to bring a lot of knowledge to the table just to form the estimator. (To specify the likelihood function, we must specify the entire joint distribution of the sample.)

**Example 8.3.2** Let $x_1, \ldots, x_N$ be IID observations from some unknown density on $\mathbb{R}$. If the class of candidate distributions is too large, maximum likelihood estimation will fail. For example, there is no solution to the problem

$$\max_p \prod_{n=1}^{N} p(x_n)$$

where the maximization is over all densities. We can make this expression arbitrarily large by choosing $p$ to concentrate its mass in small neighborhoods around each sample point. The limit of this process is not a density.

#### 8.3.1.1 Conditional Maximum Likelihood

As in §8.2.2, suppose that we observe inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to some system and corresponding outputs $y_1, \ldots, y_N$. The pairs $(\mathbf{x}_n, y_n)$ are assumed to be IID. Our aim is to estimate $\boldsymbol{\theta}$ in $p(y \mid \mathbf{x}; \boldsymbol{\theta})$ in order to pin down the conditional density of $y$ given $\mathbf{x}$.

The principle of maximum likelihood tells us to maximize

$$\ell(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n, y_n; \boldsymbol{\theta}) \quad \text{where} \quad p := \text{ the joint density of } (\mathbf{x}_n, y_n)$$

Letting $\pi$ be the marginal density of $\mathbf{x}$, we can use the decomposition (5.27) on page 142 to write

$$p(\mathbf{x}, y; \boldsymbol{\theta}) = p(y \mid \mathbf{x}; \boldsymbol{\theta})\, \pi(\mathbf{x})$$

The density $\pi$ is unknown but we have not parameterized it because we aren't trying to estimate it. We can now rewrite the log likelihood as

$$\ell(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p(y_n \mid \mathbf{x}_n; \boldsymbol{\theta}) + \sum_{n=1}^{N} \ln \pi(\mathbf{x}_n)$$

The second term on the right-hand side is independent of $\boldsymbol{\theta}$ and as such it does not affect the maximizer. Hence the MLE is

$$\operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{n=1}^{N} \ln p(y_n \mid \mathbf{x}_n; \boldsymbol{\theta})$$

The objective function here is called the **conditional log likelihood**. The preceding argument tells us that, when we want to estimate parameters in the conditional density of $y$ given $\mathbf{x}$, we can maximize the conditional log likelihood directly, which is simpler and more direct than maximizing the full log likelihood.

**Example 8.3.3** Consider a discrete response model with binary output $y_n$, where $y_n = 1$ indicates that the $n$th individual in a sample of women participates in the labor force. This decision is influenced by a vector $\mathbf{x}_n$ measuring characteristics such as income from the rest of the household. Let

$$q(\mathbf{s}) := \mathbb{P}\{y = 1 \mid \mathbf{x} = \mathbf{s}\} \qquad \left(\mathbf{s} \in \mathbb{R}^K\right)$$

One modeling approach is to take $q(\mathbf{s}) = F(\boldsymbol{\beta}^\mathsf{T}\mathbf{s})$, where $\boldsymbol{\beta}$ is a vector of parameters and $F$ is a specified CDF. We can then write

$$\mathbb{P}\{y = i \mid \mathbf{x} = \mathbf{s}\} = F(\boldsymbol{\beta}^\mathsf{T}\mathbf{s})^i (1 - F(\boldsymbol{\beta}^\mathsf{T}\mathbf{s}))^{1-i} \quad \text{for } \mathbf{s} \in \mathbb{R}^K \text{ and } i \in \{0, 1\}$$

This is the conditional PMF of $y$ given $\mathbf{x}$, so the conditional log likelihood of the sample is

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{n=1}^{N} \ln[F(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_n)^{y_n} (1 - F(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_n))^{1-y_n}] \\ &= \sum_{n=1}^{N} y_n \ln F(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_n) + \sum_{n=1}^{N} (1 - y_n) \ln(1 - F(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_n)) \end{aligned}$$

If $F$ is the standard normal CDF $\Phi$, then this model is called the **probit** model. If $F$ is the logistic CDF $F(\mathbf{s}) = 1/(1+e^{-s})$, then it is called the **logit** model.

To find the MLE in this example, we can differentiate $\ell$ to obtain first order conditions, but this does not in general lead to an analytical solution. Instead, numerical optimization of the likelihood is required. Discussion of numerical optimization is given in §13.2.2.

### 8.3.2 Maximum Likelihood via ERM

Maximum likelihood is a special case of empirical risk minimization. To see this, suppose that we wish to learn unknown density $q$ on the basis of observations from $q$. We take our loss function to be

$$L(p, x) := -\ln p(x)$$

The meaning is this: if our guess of $q$ is $p$ and the value $x$ is realized, then our loss is $-\ln p(x)$. Thus we suffer large loss when $p$ puts small probabilities near where $x$ is realized. The corresponding risk function is

$$R(p) = \mathbb{E}_q[L(p, x)] = -\int \ln[p(s)]q(s)\,\mathrm{d}s$$

Now suppose that we observe IID draws $x_1, \ldots, x_N$ from $q$. To estimate $q$, the ERM principle indicates we should solve for

$$\hat{p} := \operatorname*{argmin}_{p \in \mathscr{P}} \left\{ \frac{1}{N} \sum_{n=1}^{N} -\ln p(x_n) \right\} = \operatorname*{argmax}_{p \in \mathscr{P}} \left\{ \sum_{n=1}^{N} \ln p(x_n) \right\}$$

where $\mathscr{P}$ is a hypothesis space of densities. To clarify the connection with maximum likelihood, take $\mathscr{P}$ to be a parametric class $\{p(\cdot\,; \boldsymbol{\theta})\}_{\theta \in \Theta}$. Choosing our estimate $\hat{p}$ of $q$ now reduces to choosing an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Rewriting our optimization problem for this case, we obtain

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \left\{ \sum_{n=1}^{N} \ln p(x_n; \boldsymbol{\theta}) \right\} = \operatorname*{argmax}_{\boldsymbol{\theta}} \; \ell(\boldsymbol{\theta})$$

Here $\ell$ is the log-likelihood. It follows from this expression that the ERM estimator is precisely the maximum likelihood estimator.

Incidentally, the minimizer of the true risk $R(p)$ is the unknown density $q$. To see

this, let us first transform our expression for the risk function to

$$R(p) = \int \ln \left[ \frac{q(s)}{p(s)} \right] q(s)\, \mathrm{d}s - \int \ln[q(s)] q(s)\, \mathrm{d}s$$

The term on the far right is called the **entropy** of $q$, and does not involve $p$. Hence minimization of the risk comes down to minimization of

$$D(q, p) := \int \ln \left[ \frac{q(s)}{p(s)} \right] q(s)\, \mathrm{d}s \tag{8.34}$$

This quantity is called the **Kullback–Leibler (KL) deviation** between $q$ and $p$. The KL deviation is possibly infinite, always nonnegative, and zero if and only if $p = q$.[5] It follows that the unique minimizer of the risk is the true density $q$.

### 8.3.3 The Method of Moments and GMM

Suppose that we wish to estimate a vector $\boldsymbol{\theta}$ that solves an equation of the form

$$g(\boldsymbol{\theta}) = \mathbb{E} h(\mathbf{x}) \tag{8.35}$$

In this expression, $g$ and $h$ are observable vector-valued functions (taking values in $\mathbb{R}^K$, say). We cannot solve this expression because the distribution $P$ of $\mathbf{x}$ is unknown, and hence the expectation cannot be evaluated. If, however, we have observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from $P$ then we can apply the sample analogue principle and replace the population expectation with the expectation under the empirical distribution. This yields the **method of moments estimator**, which is the solution $\hat{\boldsymbol{\theta}}$, if it exists, to the equation

$$g(\hat{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{n=1}^{N} h(\mathbf{x}_n) \tag{8.36}$$

**Example 8.3.4** The mean of the Pareto distribution with scale parameter $s_0 = 1$ and shape parameter $\alpha$ is $\alpha/(\alpha - 1)$. (The condition for existence of the mean is $\alpha > 1$.) Letting $g(\alpha) := \alpha/(\alpha - 1)$, we can write the same statement as

$$g(\alpha) = \mathbb{E} x \quad \text{where} \quad \mathcal{L}(x) = \text{Pareto}(\alpha, 1)$$

The first equation is a version of (8.35). To estimate $\alpha$ with observations $x_1, \ldots, x_N$ from a Pareto$(\alpha, 1)$ distribution, we can apply the method of moments, which tells us to solve $g(\hat{\alpha}) = \frac{1}{N} \sum_{n=1}^{N} x_n$ for $\hat{\alpha}$. The result is $\hat{\alpha} := \bar{x}_N/(\bar{x}_N - 1)$.

5. More precisely, $D(q, p) = 0$ if and only if $p = q$ almost everywhere. Densities are equal almost everywhere when the set of points at which they fail to be equal has Lebesgue measure zero.

**Generalized method of moments** (GMM) is a small step from method of moments. If we express (8.35) as $\mathbb{E}[g(\boldsymbol{\theta}) - h(\mathbf{x})] = \mathbf{0}$, then it becomes natural to consider the more general expression

$$\mathbb{E}G(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{0} \tag{8.37}$$

This expression is called the **orthogonality condition**. The generalized method of moments estimator of $\boldsymbol{\theta}$ is the solution $\hat{\boldsymbol{\theta}}$ to the empirical counterpart, which is

$$\frac{1}{N}\sum_{n=1}^{N} G(\hat{\boldsymbol{\theta}}, \mathbf{x}_n) = \mathbf{0} \tag{8.38}$$

Of course, there is no guarantee that a solution will exist here, partly because the function $G$ can be nonlinear and partly because the number of equations can be greater than the number of unknowns. If the number of equations is greater, then the estimation problem is said to be **overidentified**.

Our study of overdetermined systems of equations in §3.3.2 suggests a logical way to handle the overidentified case: Minimize the norm of the left-hand side of (8.38). This leads to the expression

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \left\| \frac{1}{N}\sum_{n=1}^{N} G(\boldsymbol{\theta}, \mathbf{x}_n) \right\| \tag{8.39}$$

In practice, we usually make two adjustments to this expression. The first is trivial: we minimize the squared norm, instead of the norm, leaving the minimizer unchanged. The second is to replace the Euclidean norm $\|\cdot\|$ with a weighted norm $\|\cdot\|_W$ defined by $\|\mathbf{x}\|_W^2 = \mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is a positive definite **weighting matrix**. Positive definiteness is desired here because it shares with the Euclidean norm the property that led us to (8.39). That is, $\|\mathbf{x}\|_W = 0$ if and only if $\mathbf{x} = \mathbf{0}$. The estimation problem is then

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \left[\frac{1}{N}\sum_{n=1}^{N} G(\boldsymbol{\theta}, \mathbf{x}_n)\right]^\mathsf{T} \hat{\mathbf{W}} \left[\frac{1}{N}\sum_{n=1}^{N} G(\boldsymbol{\theta}, \mathbf{x}_n)\right] \tag{8.40}$$

The weighting matrix has been written as $\hat{\mathbf{W}}$ because it is allowed to depend on the sample. Evidently the choice of $\hat{\mathbf{W}}$ affects the minimizer, and the objective in choosing this matrix is to produce an estimator that has small variance asymptotically.

**Example 8.3.5** GMM is often used to estimate and test asset pricing models. Many asset pricing models lead to equations of the form

$$\mathbf{p}_t = \mathbb{E}\left[M_{t+1}\mathbf{x}_{t+1} \,|\, \mathcal{G}_t\right]$$

where $\mathbf{x}_{t+1}$ is a vector of payoffs on $K$ assets at $t+1$, $\mathbf{p}_t$ is a corresponding vector of time $t$ asset prices, $M_{t+1}$ is a stochastic discount factor and $\mathcal{G}_t$ is the time $t$ information set. See, for example, Hansen (2014b). If $\mathbf{Z}_t$ is a conformable matrix of observable variables adapted to the filtration $\mathcal{G}_t$, then we can postmultiply by $\mathbf{Z}_t$, pass through the conditional expectation and rearrange to obtain

$$\mathbb{E}\left[M_{t+1}\mathbf{x}_{t+1}\mathbf{Z}_t - \mathbf{p}_t\mathbf{Z}_t \mid \mathcal{G}_t\right] = \mathbf{0}$$

Taking the unconditional expectation and using the law of iterated expectations gives an expression in the form of (8.37). In Hansen and Singleton (1982), the variables in $\mathbf{Z}_t$ include lagged values of asset returns and aggregate consumption growth.

### 8.3.4 Bayesian Estimation

Bayesian inference adopts a rather different strategy from the methods discussed so far. The main idea is to treat parameters as unknown quantities for which we hold subjective beliefs regarding their values. These subjective beliefs are called **priors**. The Bayesian approach to estimation suggests that we take both data and prior knowledge into account when forming an estimate or prediction.

**Example 8.3.6** Consider the expression "when you hear hooves, think horses, not zebras." Prior knowledge should be given some weighting when assessing evidence.

For the purposes of estimation, a prior can be thought of as a distribution over $\mathscr{P}$, the set of distributions in play. The standard Bayesian approach is parametric, so we can specialize this further to a density over parameter space. Thus the primitives in our analysis are:

- $\boldsymbol{\theta}$, the parameter vector, which takes values in $\Theta \subset \mathbb{R}^J$,
- $\pi$, the **prior distribution**, a density over $\Theta$,
- $\mathbf{x}$, the data, and
- $p(\cdot \mid \boldsymbol{\theta})$, the joint density of the data given $\boldsymbol{\theta}$.

Note that $L(\boldsymbol{\theta}) := p(\mathbf{x} \mid \boldsymbol{\theta})$ is the likelihood function.

Priors are reassessed based on evidence in the data. This process leads to an updated density over parameter space called the **posterior distribution**, which we represent by $\pi(\boldsymbol{\theta} \mid \mathbf{x})$. The posterior is obtained via an application of Bayes' law (see page 142), which leads us to

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{x} \mid \boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}'} \tag{8.41}$$

Here $p(\mathbf{x})$ represents the unconditional density of $\mathbf{x}$ evaluated at the outcome. The term on the far right shows why this unconditional density is not listed as a primitive: we can recover it from the other primitives using the law of total probability.

The same method can be applied when the densities are replaced with probability mass functions and the integral is replaced with a sum. Here's a standard example that mixes a density over priors with a binomial PMF for the likelihood:

**Example 8.3.7** Consider a one-armed bandit (slot machine) with binary response $v$ indicating a fixed payout ($v = 1$) or nothing ($v = 0$). We would like to know the probability $\theta$ of $v = 1$. Let $v_1, \ldots, v_N$ be a sequence of independent outcomes and let $x := \sum_{n=1}^{N} v_n$ be the total number of payouts. Recalling example 4.2.6 on page 104, the likelihood for $x$ conditional on $\theta$ is

$$p(x \mid \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$$

For our prior we take a Beta$(\alpha, \beta)$ distribution, so

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \tag{8.42}$$

for $0 < \theta < 1$. (Parameters like $\alpha$ and $\beta$ that are used to define the prior density are called **hyperparameters**.) Applying (8.41) gives

$$\pi(\theta \mid x) = \frac{\theta^{x+\alpha-1}(1-\theta)^{N-x+\beta-1}}{c(x)} \tag{8.43}$$

where $c(x) := p(x)B(\alpha, \beta)/\binom{N}{x}$. We could try to calculate $c(x)$ directly but there's an easier way. We know that (8.43) is a density in $\theta$ given $x$. Hence $c(x)$ must be the normalizing constant at $x$. Moreover, in comparing (8.42) with (8.43), it's clear that $\pi(\theta \mid x)$ is a beta density. This leads us to the full form of the posterior, which is

$$\pi(\theta \mid x) = \frac{\theta^{\alpha+x-1}(1-\theta)^{N-x+\beta-1}}{B(x+\alpha, N-x+\beta)} \tag{8.44}$$

Figure 8.10 shows evolution of the posterior density (8.44) in a simulation. The prior is set to Beta$(3, 5)$. The true payout probability is $\theta_0 = 0.7$. Despite the poor prior, the data shift probability mass towards $\theta_0$. See `johnstachurski.net/emet.html` for code.

Point estimates are extracted from the posterior distribution based on some measure of central tendency, such as the mean, the median, or the **mode** of the posterior (i.e., the maximizer in the unimodal case).

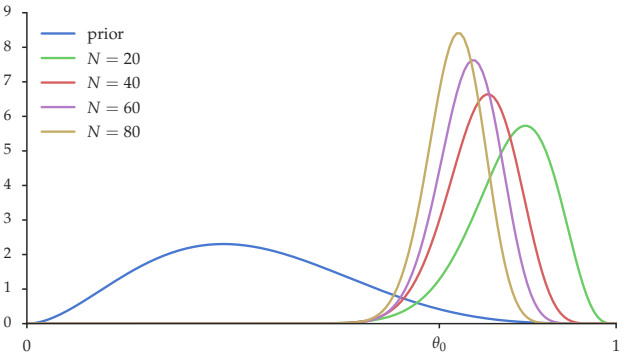


**Figure 8.10** Evolution of the posterior from Beta$(3, 5)$ prior

**Example 8.3.8** The mean of the posterior in (8.44) yields the estimator

$$\hat{\theta} := \frac{\alpha + x}{\alpha + \beta + N}$$

More payouts shift our estimator upwards. In the limit, $\hat{\theta}$ is near $\frac{x}{N}$, which is the MLE of $\theta$. This illustrates a common theme: The difference between maximum likelihood and Bayesian estimates typically concerns finite sample properties.

The posterior we came up with in (8.44) belongs to the same parametric class as the prior. Priors where the parametric class is preserved under Bayesian updating for a specific likelihood function are called **conjugate**. Conjugacy makes application of Bayes' law particularly simple.

In applications, the existence of a closed-form solution for the posterior is rare, and the integration over the parameter space in (8.41) has to be carried out numerically. The standard technique is Markov chain Monte Carlo. An exposition of the Metropolis–Hastings algorithm was given in §7.4.2. If you return to (7.38), you will notice how implementation of this algorithm only requires us to evaluate ratios of of the posterior, which means that the integral term on the right-hand side of (8.41) cancels out. Because the integral can be very high dimensional, this often eliminates a huge amount of complexity.

Bayesian inference has experienced a surge of popularity in recent years. One reason is that in high dimensions, exploring the posterior via MCMC has proved to be more successful in practice than the numerical optimization required to obtain maximum likelihood estimates. Another is that Bayesian estimation provides a form of reg-

ularization that stabilizes and typically improves estimation of complex models. See §14.2.3. A third is that Bayesian estimation comes with an elegant, unified decision-theoretic approach to inference. Some discussion is given in §9.2.5.

## 8.4 Further Reading

The ERM principle is a very general principle for solving statistical problems and producing estimators. For such a general method it is difficult to give a set of strong results showing that ERM produces good estimators. Indeed, there will be instances when ERM produces poor estimators, as discussed in 8.2.3. Having said that, some rather general consistency results have been obtained. Additional discussion can be found in Vapnik (2000).

There are many good treatments of Bayesian estimation available in the literature, including Geweke (2005), Geweke et al. (2011), and Kroese and Chan (2014).

Chernozhukov and Fernández-Val (2011) discuss quantile regression of the $\tau$th quantile in cases where $\tau$ is close to zero or one.

## 8.5 Exercises

**Ex. 8.5.1** Let $x$ be a random variable with $\mu := \mathbb{E}[x]$ and finite second moment. Consider the risk function given by $R(\theta) = \mathbb{E}[(\theta - x)^2]$. Show that $\mu$ is the minimizer of $R(\theta)$ over all $\theta \in \mathbb{R}$.

**Ex. 8.5.2** Confirm the solutions in (8.14) by differentiating (8.13).

**Ex. 8.5.3** Show that the method of moments estimator $\hat{\alpha}$ in example 8.3.4 converges in probability to $\alpha$ whenever $\alpha > 1$. (Later we will call this property consistency.)

**Ex. 8.5.4** Confirm that the maximizer of (8.28) is the sample mean of $x_1, \ldots, x_N$.

**Ex. 8.5.5** Confirm the results in (8.33) by differentiating (8.32).

**Ex. 8.5.6** Let $F$ be a strictly increasing CDF on $\mathbb{R}$ and let $\tau \in (0, 1)$ be given. Let $y$ be a random variable with $\mathcal{L}(y) = F$. Adopting the notation of §8.2.2.1, show that the solution to the optimization problem $\min_{\xi \in \mathbb{R}} \mathbb{E} L_\tau(y, \xi)$ is the $\xi$ that solves $F(\xi) = \tau$.

**Ex. 8.5.7** Consider the one-armed bandit problem in example 8.3.7. Directly obtain the log likelihood function of $\theta$ given $v_1, \ldots, v_N$. Use the principle of maximum likelihood to show that the sample mean $\bar{v}_N$ is the MLE for $\theta$.

**Ex. 8.5.8** Let $f$ and $g$ be two fixed densities, let $x_1, \ldots, x_N$ be IID and consider the **likelihood ratio statistic**

$$y_n := \prod_{i=1}^{n} \frac{g(x_i)}{f(x_i)}$$

Show that if $\mathcal{L}(x_n) = f$ for all $n$, then $\{y_n\}$ is a martingale with respect to the filtration $\{\mathscr{F}_n\}$ defined by $\mathscr{F}_n := \{x_1, \ldots, x_n\}$.

**Ex. 8.5.9** Let $D(p_1, p_2)$ be the KL deviation between normal densities $p_1$ and $p_2$ where $p_i = \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2$. Show that

$$D(p_1, p_2) = \ln \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

### 8.5.1 Solutions to Selected Exercises

**Solution to Ex. 8.5.1.** Adding and subtracting $\mu$, we can express $R(\theta)$ as

$$R(\theta) = \mathbb{E}\{[(\theta - \mu) + (\mu - x)]^2\}$$

Expanding this out and using $\mathbb{E}[x] = \mu$, we obtain $R(\theta) = (\theta - \mu)^2 + \operatorname{var} x$. Evidently a minimum is obtained when $\theta = \mu$. □

**Solution to Ex. 8.5.6.** We can write the minimization problem as

$$\min_{\xi \in \mathbb{R}} \left\{ (\tau - 1) \int_{-\infty}^{\xi} (t - \xi) \, F(\mathrm{d}t) + \tau \int_{\xi}^{\infty} (t - \xi) \, F(\mathrm{d}t) \right\}$$

The first-order condition is $(1 - \tau)F(\xi) + \tau(F(\xi) - 1) = 0$. Simplifying gives $F(\xi) = \tau$, as was to be shown. □

**Solution to Ex. 8.5.7.** Each $v_n$ is binary, with PMF given by $p(s; \theta) := \theta^s (1 - \theta)^{1-s}$ for $s \in \{0, 1\}$. By independence, the joint distribution is the product of the marginals, and hence the log likelihood is

$$\ell(\theta) = \sum_{n=1}^{N} \log p(v_n; \theta) = \sum_{n=1}^{N} [v_n \log \theta + (1 - v_n) \log(1 - \theta)]$$

Differentiating with respect to $\theta$ and setting the result equal to zero yields $\hat{\theta} = \bar{v}_N$ as claimed. □

**Solution to Ex. 8.5.8.** It is clear that $y_n$ is $\mathscr{F}_n$-measurable, and hence $\{y_n\}$ is adapted to $\{\mathscr{F}_n\}$. In addition, we have

$$\mathbb{E}[y_{n+1} \mid \mathscr{F}_n] = \mathbb{E}\left[\prod_{i=1}^{n+1} \frac{g(x_i)}{f(x_i)} \mid \mathscr{F}_n\right] = \prod_{i=1}^{n} \frac{g(x_i)}{f(x_i)} \mathbb{E}\left[\frac{g(x_{n+1})}{f(x_{n+1})} \mid \mathscr{F}_n\right]$$

Since $\{x_n\}$ is IID, $\mathcal{L}(x_{n+1}) = f$ and $g$ is a density, we have

$$\mathbb{E}\left[\frac{g(x_{n+1})}{f(x_{n+1})} \mid \mathscr{F}_n\right] = \mathbb{E}\left[\frac{g(x_{n+1})}{f(x_{n+1})}\right] = \int \frac{g(s)}{f(s)} f(s) \, \mathrm{d}s = \int g(s) \, \mathrm{d}s = 1$$

$$\therefore \quad \mathbb{E}[y_{n+1} \mid \mathscr{F}_n] = \prod_{i=1}^{n} \frac{g(x_i)}{f(x_i)} = y_n \qquad \square$$