

Chapter 9

Properties of Estimators

When the concept of estimators was first introduced in §8.1.2 we mentioned that there is nothing in the definition that guarantees—or even alludes to—good performance. Let’s now start to think about evaluating estimators.

9.1 Sampling Distributions

Our first step is to view estimators as random entities dependent on the data, and to consider their distributions under different specifications of the data-generating process. The notation below follows that set out in §8.1.1.

9.1.1 Estimators as Random Elements

By definition, a statistic T is a $\mathcal{B}$ -measurable transformation of the sample space $Z_{\mathcal{D}}$ into some feature space S . As a function of the sample $\mathbf{z}_{\mathcal{D}} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$, it can be viewed as a random element mapping outcomes in the underlying space Ω to S via

$$\omega \mapsto T[\mathbf{z}_{\mathcal{D}}(\omega)] \in S$$

Here “random element” stands for either a random vector (or scalar), or something more general, such as the random distribution $\hat{P}_N$, or a density estimate, which is a random function of the sample.

Example 9.1.1 Consider a sample mean $\bar{x}_N := \frac{1}{N} \sum_{n=1}^N x_n$. The observations $x_1, \dots, x_N$ are understood to be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sam-

ple mean inherits this property. Formally, it is the random variable

$$\bar{x}_N(\omega) := \frac{1}{N} \sum_{n=1}^N x_n(\omega) \quad (\omega \in \Omega) \quad (9.1)$$

Example 9.1.2 A histogram is a random estimate of the density of the unknown distribution P , mapping sample realizations into step functions on the outcome space.

The idea that estimators are random can be counterintuitive. When we begin to study econometrics, we tend to think of data as fixed numbers determined by some prior historical outcome and statistics as deterministic functions of these numbers. Hence we only observe one value of any particular statistic—one sample mean, one sample variance, and so on. However, if we want to build a theory of estimators, then we need to consider their performance prior to observing the data. Each data set is viewed as a random draw from some unknown distribution. We seek estimators that perform well in the face of this uncertainty.

Once we agree that estimators are random, we know they must have distributions. These distributions determine how effective an estimator is at estimating the intended feature. Let's now turn to investigating these distributions.

9.1.2 Sampling Distributions

Suppose that T is a statistic mapping data $\mathbf{z}_D$ into S . As just discussed, $T(\mathbf{z}_D)$ is a random element taking values in S . Its distribution $\mathcal{L}(T)$ on S can be expressed as

$$\mathcal{L}(T)(B) = \mathbb{P} \{ T(\mathbf{z}_D) \in B \} \quad (9.2)$$

$\mathcal{L}(T)$ is called the **sampling distribution** of T . Continuing to write P_D for the joint distribution of the sample $\mathbf{z}_D$, we can also define the sampling distribution as

$$\mathcal{L}(T)(B) = P_D \{ \mathbf{s} \in Z_D : T(\mathbf{s}) \in B \}$$

or even $P_D \circ T^{-1}$ if you like that kind of notation. Note that the sampling distribution of T is fully determined by

- (i) the statistic T that maps data to feature space,
- (ii) the joint distribution P_D of the data.

Since P_D is unknown, the sampling distribution is likewise unknown. It can, however, be estimated, as we'll see below.

Mathematically, there is no difference between the "sampling distribution of T " and the "distribution of T ." Nonetheless, the terminology is standard.

Example 9.1.3 Let $x_1, \dots, x_N$ be IID on $\mathbb{R}$ with $P = \mathcal{L}(x_n) = N(\mu, \sigma^2)$ for all n . Let T be the sample mean $\bar{x}_N$. By independence, the joint distribution $P_{\mathcal{D}}$ of $\mathbf{z}_{\mathcal{D}} = \mathbf{x} := (x_1, \dots, x_N)$ is $N(\mu \mathbf{1}, \sigma^2 \mathbf{I})$. Applying fact 5.1.5 on page 133 gives

$$\mathcal{L}(\bar{x}_N) = N\left(\mu, \frac{\sigma^2}{N}\right) \quad (9.3)$$

This is the sampling distribution of $\bar{x}_N$.

Sampling distributions are often complicated. Sometimes the best we can do is obtain the sampling distribution of some transformation of an estimator.

Example 9.1.4 Let $\mathbf{x} := (x_1, \dots, x_N)$ have the same distribution as example 9.1.3. Let s_N^2 be the sample variance. If $\sigma > 0$, then

$$\mathcal{L}(q_N) = \chi^2(N-1) \quad \text{where} \quad q_N := \frac{N s_N^2}{\sigma^2} \quad (9.4)$$

To see why (9.4) holds, recall from example 2.2.3 on page 33 that if $S := \text{span}\{\mathbf{1}\}$ and $\mathbf{P} = \text{proj } S$, then the residual projection satisfies $\mathbf{M}\mathbf{x} = \mathbf{x} - \bar{x}_N \mathbf{1}$. Hence

$$N s_N^2 = \sum_{n=1}^N (x_n - \bar{x}_N)^2 = \|\mathbf{M}\mathbf{x}\|^2$$

Since $\mathbf{M}\mathbf{1} = \mathbf{0}$ (see fact 2.2.8 on page 33) and $\mathbf{M}$ is idempotent and symmetric (fact 3.3.2), setting $\boldsymbol{\xi} := \sigma^{-1}(\mathbf{x} - \mu \mathbf{1})$ gives

$$q_N = \sigma^{-2} \|\mathbf{M}\mathbf{x}\|^2 = \left\| \mathbf{M} \sigma^{-1} (\mathbf{x} - \mu \mathbf{1}) \right\|^2 = \|\mathbf{M}\boldsymbol{\xi}\|^2 = \boldsymbol{\xi}^\top \mathbf{M} \boldsymbol{\xi} \quad (9.5)$$

To see that the right-hand side has distribution $\chi^2(N-1)$, apply fact 3.3.4 on page 61 to obtain $\text{trace } \mathbf{M} = N-1$, followed by fact 5.1.19 on page 141.

Fact 9.1.1 For an IID sample $x_1, \dots, x_N$ with common distribution $N(\mu, \sigma^2)$, the sample mean and sample variance are independent random variables.

Many long and involved proofs of fact 9.1.1 can be found in the literature, but a simple and informative proof can be obtained from fact 5.1.13 on page 137. See exercise 9.4.2 and its solution.

In the preceding examples we were able to obtain an analytical expression for the sampling distribution of our estimator (or a simple transformation of the estimator) from the underlying distribution of the data. In reality this case is the exception.

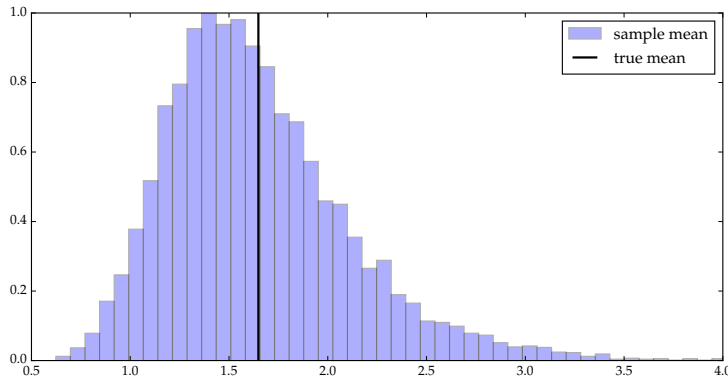


Figure 9.1 Sampling distribution of $\bar{x}_N$ when $N = 20$ and data are IID $\text{LN}(0, 1)$

Example 9.1.5 A random variable x is said to be **lognormally distributed** with parameters μ, σ if $\mathcal{L}(\ln x) = \mathcal{N}(\mu, \sigma^2)$. We write $\mathcal{L}(x) = \text{LN}(\mu, \sigma^2)$. No neat expression exists for $\mathcal{L}(\bar{x}_N)$ when the data are IID lognormal.

When no closed-form exists, we can still approximate the sampling distribution using simulation, conditional on a specification of the joint distribution of the data. For example, figure 9.1 shows a histogram approximation to $\mathcal{L}(\bar{x}_N)$ when $N = 20$ and $x_1, \dots, x_N$ is IID $\text{LN}(0, 1)$. The histogram is produced by generating 10,000 observations of $\bar{x}_N$. The loop that produced the observations is given in listing 5.

Listing 5 Generates 10,000 observations of the sample mean (Python)

```
import numpy as np

N = 20
num_reps = 10000
xbar_outcomes = np.empty(num_reps) # Allocate memory

for i in range(num_reps):
    x = np.exp(np.random.randn(N)) # Generate N iid lognormal RVs
    xbar_outcomes[i] = x.mean()
```

9.1.2.1 Desirable Properties

The quality or performance of a given estimator is characterized by its sampling distribution. Let $\hat{\gamma}$ be an estimator of some feature γ based on a sample $\mathbf{z}_{\mathcal{D}}$. What would be an ideal sampling distribution for $\hat{\gamma}$? The answer is obvious: a point mass that puts probability one on the single point γ .

This scenario rarely arises because our sample is *random*. So here's a more realistic objective:

It is desirable that the sampling distribution of an estimator of γ concentrate most of its probability mass in a small neighborhood around γ , regardless of the joint distribution $P_{\mathcal{D}}$ of the sample.

This statement considers only what is desirable, not what is possible. For example, if we look back at figure 9.1, we see that the sampling distribution of $\bar{x}_N$ does not cluster tightly around the mean. (The feature γ that we are estimating is the mean of the $\text{LN}(0, 1)$ distribution, which is represented by the vertical black line.) But this is not unexpected, since the sample size is only 20. We simply don't have enough information to estimate the feature precisely.

Nonetheless, these ideas give us a way to start comparing different estimators of a given feature from a given amount of data. For example, suppose that we are estimating the mean of a distribution from a random sample of size 20. Suppose that each is drawn from the $\text{LN}(0, 1)$ distribution, the mean of which is about 1.65. The sampling distributions of three estimators of the mean under the associated joint distribution have been approximated by simulation and represented as box plots in figure 9.2. The top sampling distribution is that of the mid-range estimator (see page 217), the second is that of a maximum likelihood estimator to be discussed later, and the third is that of the sample mean. The last two put more probability mass in the region around 1.65. Hence they outperform the mid-range estimator in this setting.

In what follows we will formalize the ideas presented above, including both finite sample and asymptotic theory. But first let's discuss a general way of estimating sampling distributions from the data.

9.1.3 The Bootstrap

While understanding sampling distributions is central to evaluating the precision of estimators, sampling distributions cannot be observed in the wild because they depend on $P_{\mathcal{D}}$. For example, suppose $x_1, \dots, x_N$ are assumed to be IID $\text{N}(\mu, \sigma^2)$ with μ and σ unknown. By (9.3), the distribution of $\bar{x}_N$ is $\text{N}(\mu, \sigma^2/N)$. While this provides valuable information, it doesn't fully pin down the distribution because μ and σ are unknown.

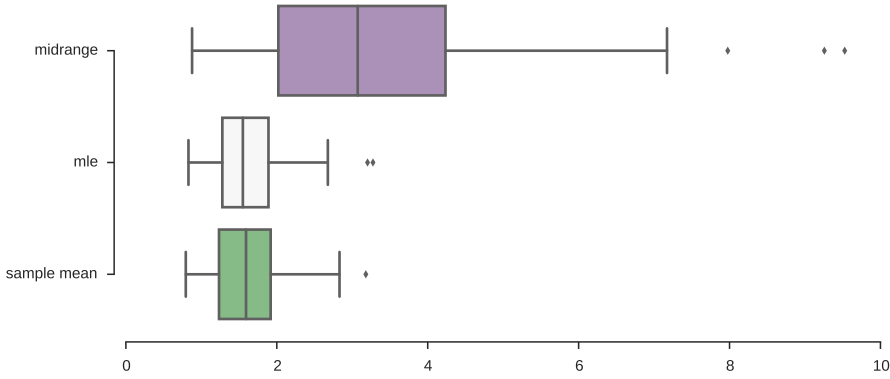


Figure 9.2 Sampling distributions of three estimators of a lognormal mean

Things are even more complicated for the sample mean of IID lognormals in example 9.1.5. There we can't easily write down the sampling distribution as a function of the parameters. Problems multiply further if we are not prepared to assume a parametric form for the distribution of the observations.

There is, however, a very general technique that allows us to approximate sampling distributions given data. We'll discuss the case of scalar data, although the vector case is similar. To begin, let T be any statistic and consider its sampling distribution

$$\mathcal{L}_P(T) := \text{the distribution of } T(x_1, \dots, x_N) \text{ when } x_1, \dots, x_N \stackrel{\text{iid}}{\sim} P$$

The sampling distribution is unknown to us because P is unknown. Suppose, however, that we have access to a sample $x_1^o, \dots, x_N^o$ of IID draws from P . We then have a good estimator of P given our sample: the empirical distribution $\hat{P}_N$ generated by $x_1^o, \dots, x_N^o$. Given this empirical distribution, the **bootstrap distribution** of T is defined as

$$\mathcal{L}_{\hat{P}_N}(T) := \text{the distribution of } T(x_1, \dots, x_N) \text{ when } x_1, \dots, x_N \stackrel{\text{iid}}{\sim} \hat{P}_N$$

Thus the bootstrap distribution is just the plug-in estimator of the sampling distribution.

Is $\mathcal{L}_{\hat{P}_N}(T)$ a good approximation to $\mathcal{L}_P(T)$? By the Glivenko–Cantelli theorem, $\hat{P}_N$ will be close to P when N is large. If $Q \mapsto \mathcal{L}_Q(T)$ is suitably continuous, then $\mathcal{L}_{\hat{P}_N}(T)$ will also be close to $\mathcal{L}_P(T)$. For details, see §29.2 of DasGupta (2008).

9.1.3.1 Simulation

One remaining problem with $\mathcal{L}_{\hat{P}_N}(T)$ is that it's not particularly tractable. We can, however, generate draws from this distribution easily on a computer. Here's an algorithm that returns M draws from $\mathcal{L}_{\hat{P}_N}(T)$:

- 1: set $\hat{P}_N$ = the empirical distribution of observed data $x_1^o, \dots, x_N^o$
- 2: **for** m in $1, \dots, M$ **do**
- 3: draw $x_1^b, \dots, x_N^b$ independently from $\hat{P}_N$
- 4: set $T_m^b = T(x_1^b, \dots, x_N^b)$
- 5: **end for**
- 6: return the sample $T_1^b, \dots, T_M^b$

Using the definition of the empirical distribution, drawing samples from $\hat{P}_N$ is easy: we just make repeated draws from the set $x_1^o, \dots, x_N^o$ with equal probability on each element. Listing 6 contains some Julia code implementing the whole algorithm as a function that returns M bootstrap samples. The other arguments to the function are `xo`, the array of observed data, and `stat`, which is a function representing T .

Listing 6 Function to generate bootstrap samples (Julia)

```
function bootstrap(xo, stat, M)
    N = length(xo)
    T_b = Array{Float64, M}
    x_b = Array{Float64, N}
    for m in 1:M
        for i in 1:N
            x_b[i] = xo[rand(1:N)]
        end
        T_b[m] = stat(x_b)
    end
    return T_b
end
```

Here's an example function call:

```
julia> bootstrap([1, 2, 3], mean, 3)
3-element Array{Float64,1}:
1.66667
1.0
1.66667
```

Here is the same routine in Python:

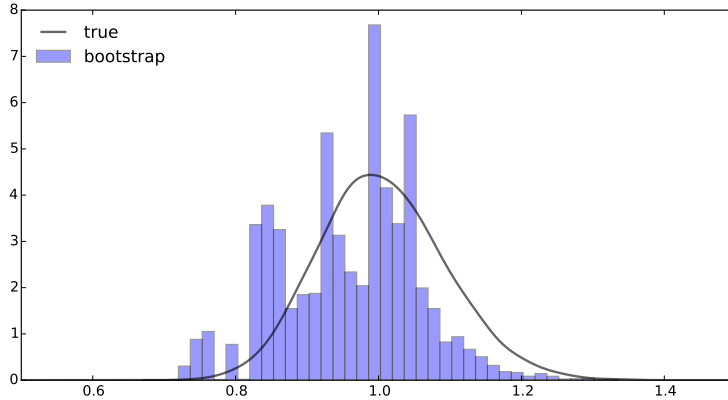


Figure 9.3 Bootstrap draws and sampling distribution for the median

```
def bootstrap(xo, stat, M):
    N = len(xo)
    T_b = np.empty(M)
    for m in range(M):
        x_b = np.random.choice(xo, size=N)
        T_b[m] = stat(x_b)
    return T_b
```

Figure 9.3 shows a histogram of observations from a bootstrap distribution generated using the Julia code above. The underlying sample $x_1^o, \dots, x_N^o$ was created using $N = 200$ IID draws from the $\text{LN}(0, 1)$ distribution. The statistic used was the sample median. The density shown in the same figure is a numerical approximation of the true sampling distribution of the median over 200 IID draws from the $\text{LN}(0, 1)$ distribution. The (population) median of the $\text{LN}(0, 1)$ distribution is 1. See johnstachurski.net/emet.html for the code.

9.1.3.2 Standard Errors

As discussed in 9.1.2.1, a good estimator is one that concentrates its probability mass close to the target feature. The extent to which probability mass concentrates is usually measured by the variance or standard deviation of the sampling distribution. The standard deviation is preferable because it comes in units that are easier to interpret.

Because the sampling distribution is typically unknown, so is its standard deviation. Hence we replace the true standard deviation with an estimate. An estimate of

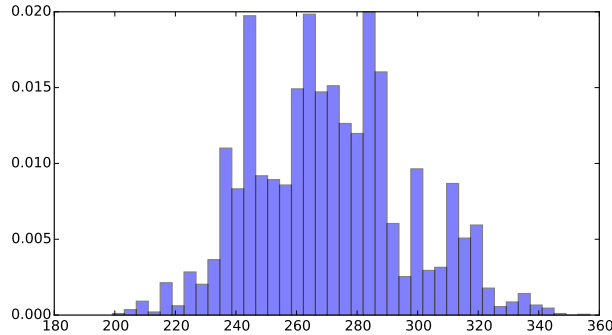


Figure 9.4 Bootstrapped draws for the median of firm sizes by sales

the standard deviation of a sampling distribution is usually referred to as the **standard error**. More generally,

$$\text{se}(\hat{\gamma}) := \text{an estimate of the standard deviation of } \hat{\gamma}$$

This isn't a proper definition because we haven't specified *which* estimate of the standard deviation we are talking about, but nevertheless the terminology is common. One way to compute a standard error of an estimator is to take the sample standard deviation of the bootstrap draws.

Example 9.1.6 Figure 9.4 shows a histogram of draws from the bootstrap distribution of the median applied to 2,118 observations of US firm sizes by sales, as measured by 2013 Compustat data. Values are in millions of US dollars. The median of the sample itself is 269.9. The sample standard deviation of this particular set of bootstrap draws is 25.4. See johnstachurski.net/emet.html for the code.

9.2 Evaluating Estimators

As mentioned above, the standard definition of an estimator does not include any restrictions that imply sensible behavior (as opposed to, say, the definition of a probability on page 83). We now turn again to the problem of assessing estimators, following our informal discussion in §9.1.2.1. We'll begin with finite sample properties.

9.2.1 Bias

One common criterion for assessing estimators is bias, or lack thereof. Let $\hat{\gamma}$ be an estimator of a given feature $\gamma = \gamma(P)$. The **bias** of $\hat{\gamma}$ is defined as

$$\text{bias}_P(\hat{\gamma}, \gamma) = \mathbb{E}_P \hat{\gamma} - \gamma(P) \quad (9.6)$$

The estimator $\hat{\gamma}$ is called **unbiased** for γ over the class of distributions $\mathcal{P}$ if

$$\text{bias}_P(\hat{\gamma}, \gamma) = 0 \quad \text{for all } P \in \mathcal{P}$$

The definition is the same when the feature γ and estimator $\hat{\gamma}$ are vector-valued.

The notation $\mathbb{E}_P \hat{\gamma}$ indicates that expectation is taken under the assumption that the data points are IID with common distribution P . In general, the bias depends on P . For example, in the IID setting, the mid-range estimator (page 217) is unbiased as an estimator of the mean of a uniform distribution on $\mathbb{R}$ with unknown end points, but biased for many other distributions, including the lognormal.

Figure 9.2 already suggests bias of the mid-range estimator in the lognormal case. Here's a simulation in R that provides further confirmation.

```
> mr <- function(x) return((min(x) + max(x)) / 2)
> observations <- replicate(5000, mr(rlnorm(20)))
> mean(observations) # Approximate sample mean of mid-range
[1] 3.800108
> exp(1/2) # Mean of lognormal
[1] 1.648721
```

Fact 9.2.1 Let P be a distribution on $\mathbb{R}^K$, let $\{\mathbf{x}_n\}$ be a sample with $\mathcal{L}(\mathbf{x}_n) = P$ for all n , and let $\hat{P}_N$ be the empirical distribution. If

- (i) $\gamma(P) = \int h(\mathbf{s})P(d\mathbf{s})$ for some integrable function h , and
- (ii) $\hat{\gamma}$ is the plug-in estimator $\hat{\gamma} = \int h(\mathbf{s})\hat{P}_N(d\mathbf{s}) = \frac{1}{N} \sum_{n=1}^N h(\mathbf{x}_n)$,

then $\hat{\gamma}$ is unbiased for γ over the set of distributions such that $\int h(\mathbf{s})P(d\mathbf{s})$ exists.

Fact 9.2.1 is immediate from linearity of expectations, since

$$\mathbb{E} \hat{\gamma} = \mathbb{E} \left[\frac{1}{N} \sum_{n=1}^N h(\mathbf{x}_n) \right] = \frac{1}{N} \sum_{n=1}^N \mathbb{E} h(\mathbf{x}_n) = \int h(\mathbf{s})P(d\mathbf{s}) = \gamma(P)$$

Note that independence isn't used here. Fact 9.2.1 depends only on the existence of a common marginal P .

Example 9.2.1 For any identically distributed sample, the k th sample moment is an unbiased estimator of the k th moment, whenever the latter exists.

Example 9.2.2 Let $x_1, \dots, x_N$ be an IID sample with finite variance σ^2 . The sample variance s_N^2 is, in general, a biased estimator of σ^2 , but not by much. In particular,

$$\mathbb{E}s_N^2 = \sigma^2 \frac{N-1}{N} \quad (9.7)$$

To obtain this expression, we can combine the definition of q_N in (9.4) and the expression in (9.5) to obtain

$$\mathbb{E}s_N^2 = \frac{1}{N}\sigma^2\mathbb{E}[\xi^\top \mathbf{M}\xi] = \frac{1}{N}\sigma^2 \text{trace } \mathbf{M}$$

Here $\mathbf{M}$ is the residual projection for the subspace $\text{span}\{\mathbf{1}\}$ and the second equality is due to fact 5.1.3 (page 127). Now apply fact 3.3.4 on page 61.

9.2.2 Variance

The variance of a given estimator gives a measure of its dispersion. Sensible estimators almost always have the property that the variance goes to zero as $N \rightarrow \infty$.

Example 9.2.3 If $x_1, \dots, x_N$ are uncorrelated with common finite variance σ^2 , then

$$\text{var } \bar{x}_N = \text{var} \left[\frac{1}{N} \sum_{n=1}^N x_n \right] = \frac{\sigma^2}{N} \quad (9.8)$$

We saw this already in §6.1.2.

Example 9.2.4 Assume the IID Gaussian setting of example 9.1.4, where it was shown that the sample variance satisfies $\mathcal{L}(\sigma^{-2}Ns_N^2) = \chi^2(N-1)$. The variance of a $\chi^2(k)$ distribution is $2k$. Using this fact and some algebra gives

$$\text{var } s_N^2 = \frac{2\sigma^4}{N^2}(N-1)$$

Is small variance always desirable for an estimator? If the estimator is biased, then perhaps not, since probability mass might be concentrated in the wrong place. But for an unbiased estimator, low variance is good. It means that probability mass is concentrated around the feature that we wish to estimate.

Low variance is, of course, relative. For example, is the variance in (9.8) low or high? One way to approach this kind of question is to take the class of unbiased

estimators of a given feature γ , and find the estimator in the class with the lowest variance. For given γ and given data $\mathbf{x}_1, \dots, \mathbf{x}_N$, the estimator in the set of unbiased estimators

$$U_\gamma := \{\text{all statistics } \hat{\gamma} \text{ with } \mathbb{E} \hat{\gamma} = \gamma\}$$

that has the lowest variance within this class is called the **minimum variance unbiased estimator**.

Of course, it's quite possible that no minimizer exists. Even if it does, it might be hard to determine in practice, or require strong assumptions on the unknown distribution of the data. For these reasons it's common to focus on smaller classes than U_γ . For example, the estimator in the set of *linear* unbiased estimators

$$U_\gamma^\ell := \{\text{all linear statistics } \hat{\gamma} \text{ with } \mathbb{E} \hat{\gamma} = \gamma\}$$

with the lowest variance—if it exists—is called the **best linear unbiased estimator**.

Example 9.2.5 Let $x_1, \dots, x_N$ be IID with common distribution P , where P has finite mean $\mu \neq 0$ and variance σ^2 . The set of linear estimators of μ is given by

$$\left\{ \text{all statistics of the form } \hat{\mu} = \sum_{n=1}^N \alpha_n x_n, \text{ where } \alpha_n \in \mathbb{R} \text{ for } n = 1, \dots, N \right\}$$

Hence the set of linear unbiased estimators of μ is

$$U_\mu^\ell := \left\{ \text{all } \hat{\mu} = \sum_{n=1}^N \alpha_n x_n \text{ with } \alpha_n \in \mathbb{R}, \ n = 1, \dots, N \text{ and } \mathbb{E} \left[\sum_{n=1}^N \alpha_n x_n \right] = \mu \right\}$$

Using linearity of expectations, this set can be rewritten as

$$U_\mu^\ell := \left\{ \text{all } \hat{\mu} = \sum_{n=1}^N \alpha_n x_n \text{ with } \sum_{n=1}^N \alpha_n = 1 \right\}$$

By independence and our rules for variance of sums (page 98), the variance of an element of this class is given by

$$\text{var} \left[\sum_{n=1}^N \alpha_n x_n \right] = \sum_{n=1}^N \alpha_n^2 \text{var } x_n = \sigma^2 \sum_{n=1}^N \alpha_n^2$$

To find the best linear unbiased estimator, we need to solve

$$\text{minimize } \sigma^2 \sum_{n=1}^N \alpha_n^2 \text{ over all } \alpha_1, \dots, \alpha_N \text{ with } \sum_{n=1}^N \alpha_n = 1$$

To solve this constrained optimization problem, we can use the Lagrangian, setting

$$L(\alpha_1, \dots, \alpha_N; \lambda) := \sigma^2 \sum_{n=1}^N \alpha_n^2 - \lambda \left[\sum_{n=1}^N \alpha_n - 1 \right]$$

where λ is the Lagrange multiplier. Differentiating with respect to α_n and setting the result equal to zero, the minimizer α_n^* satisfies $\alpha_n^* = \lambda(2\sigma^2)^{-1}$ for each n . In particular, each α_n^* takes the same value, and hence, from the constraint $\sum_n \alpha_n^* = 1$, we have $\alpha_n^* = 1/N$. Using these values, our estimator becomes

$$\sum_{n=1}^N \alpha_n^* x_n = \bar{x}_N$$

Thus, for IID data with finite variance, the sample mean is the best linear unbiased estimator of μ .

The theory of minimum variance unbiased estimators and best linear unbiased estimators is traditionally regarded as important. However, there's no convincing reason to restrict ourselves to unbiased estimators. In fact, as we'll see, there are good reasons not to.

9.2.3 Variance versus Bias

The mean squared error (MSE) of a given estimator $\hat{\gamma}$ of some feature $\gamma \in \mathbb{R}^K$ is

$$\text{mse}(\hat{\gamma}, \gamma) := \mathbb{E} \left\{ \|\hat{\gamma} - \gamma\|^2 \right\} \quad (9.9)$$

As for bias, $\text{mse}(\hat{\gamma}, \gamma)$ depends on the joint distribution of the data as well as the specification of $\hat{\gamma}$, although we haven't made that dependence explicit in the notation.

Mean squared error is an overall measure of the performance of an estimator that takes into account both bias and variance. In fact in the scalar case we can neatly decompose mean squared error into the sum of variance and squared bias:

$$\text{mse}(\hat{\gamma}, \gamma) = \text{var } \hat{\gamma} + \text{bias}(\hat{\gamma}, \gamma)^2 \quad (9.10)$$

This is just (6.3) on page 163 specialized to the current setting.

Example 9.2.6 Consider again the sample variance s_N^2 under the IID Gaussian setting of example 9.2.4. We saw there that $\text{var } s_N^2 = 2\sigma^4 N^{-2}(N-1)$. Combining this result with (9.10) and (9.7), we have

$$\text{mse}(s_N^2, \sigma^2) = \frac{2\sigma^4}{N^2}(N-1) + \left[\sigma^2 \frac{N-1}{N} - \sigma^2 \right]^2 = \frac{\sigma^4}{N^2}(2N-1)$$

As can be seen from both (9.9) and (9.10), when $\text{mse}(\hat{\gamma}, \gamma)$ is small for a given estimator $\hat{\gamma}$, the estimator concentrates probability mass around the target feature γ . Hence minimizing MSE is a natural and reasonable criterion. In doing so, we are making our subjective loss function explicit, while choosing a loss function that is simple and highly tractable.

When minimizing MSE, we typically find a trade-off between bias and variance: lower variance costs more bias and vice-versa. Moreover solutions usually turn out to be interior. In particular, we can often reduce variance significantly by accepting a small amount of bias.

Example 9.2.7 Let $\hat{\gamma}$ be any unbiased estimator of a feature $\gamma \in \mathbb{R}$. Let $v := \text{var } \hat{\gamma}$. Consider the class of estimators $\{\lambda \hat{\gamma} : \lambda \in \mathbb{R}\}$. As exercise 9.4.3 asks you to verify, the value of λ that minimizes $\text{mse}(\lambda \hat{\gamma}, \gamma)$ is

$$\lambda^* := \frac{\gamma^2}{\gamma^2 + v} \quad (9.11)$$

In other words, the unbiased estimator does not minimize MSE unless $v = 0$.

For most sensible estimators, v converges to zero as the sample size $\rightarrow \infty$. The message is that unbiased estimators only approach optimality when the data size is large. Here “large” should be understood as relative to model complexity. More complex models need more data to be estimated effectively.

9.2.3.1 Stein’s Example

There is a famous example in the theory of statistics due to Stein (1956) that pertains to so-called admissible estimators. Here we’ll borrow Stein’s problem and proposed solution to illustrate several ideas, including the relationship between bias and mean squared error.

To make things concrete, consider the following application: In the field of public finance, one topic of interest is ability-based taxation, which has optimality properties in terms of welfare in certain models. Ability is not observable. Suppose, in particular, that we have a population of K individuals, where the k th individual has idiosyncratic ability $\mu_k \in [0, 1]$. For each agent, we get to observe N noisy signals of ability, each

of the form $x_{kn} = \mu_k + \epsilon_{kn}$, $n = 1, \dots, N$. The noise terms ϵ_{kn} are all independent $N(0, \sigma^2)$. Our task is to estimate the parameter vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ given the data from the observations x_{nk} . To make things simpler we'll assume that only $\boldsymbol{\mu}$ is unknown.

A natural estimator of $\boldsymbol{\mu}$ is the vector sample mean

$$\bar{\mathbf{x}}_N = \begin{pmatrix} \frac{1}{N} \sum_{n=1}^N x_{1n} \\ \vdots \\ \frac{1}{N} \sum_{n=1}^N x_{Kn} \end{pmatrix}$$

This is the estimator produced by applying the sample analogue principle to our problem. It's also the maximum likelihood estimator and best linear unbiased.

What Stein showed is that, whenever $K > 2$, there exists an alternative estimator $\hat{\boldsymbol{\mu}}_N$ such that

$$\mathbb{E} \left\{ \|\hat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}\|^2 \right\} < \mathbb{E} \left\{ \|\bar{\mathbf{x}}_N - \boldsymbol{\mu}\|^2 \right\} \quad \text{for all } \boldsymbol{\mu} \in \mathbb{R}^K \quad (9.12)$$

In other words, $\hat{\boldsymbol{\mu}}_N$ uniformly dominates $\bar{\mathbf{x}}_N$ in terms of MSE, for all possible values of $\boldsymbol{\mu}$! The estimator in question is what is now known as the **James–Stein estimator**. It is defined as

$$\hat{\boldsymbol{\mu}}_N = \kappa \bar{\mathbf{x}}_N \quad \text{where} \quad \kappa := 1 - \frac{K-2}{N} \frac{\sigma^2}{\|\bar{\mathbf{x}}_N\|^2}$$

This estimator is biased. It achieves lower mean squared error by reducing variance at the expense of a small amount of bias. A relatively simple proof of (9.12) can be found on p. 34 of Young and Smith (2005).

We can get some of the intuition from simulation. In our simulation we set $K = 200$, $N = 5$ and $\sigma = 1$. Thus we have 5 noisy observations on the ability of each of 200 agents. For each agent we choose their ability μ_k independently from a uniform distribution on $[0, 1]$ and next generate the disturbances ϵ_{kn} as independent standard normals. We then compute the estimators $\bar{\mathbf{x}}_N$ and $\hat{\boldsymbol{\mu}}_N$ and the corresponding error vectors.

Figure 9.5 shows a typical realization. The bars in the top panel are the elements of the vector $\bar{\mathbf{x}}_N - \boldsymbol{\mu}$, which records the errors for the sample mean estimator $\bar{\mathbf{x}}_N$. Those in the lower panel correspond to the errors of the James–Stein estimator. These are noticeably downward biased and also noticeably smaller than the errors of the sample mean estimator. This is reflected in the SSE values in the top left of each panel, which give the sum of squared errors (i.e., the squared norms of the error vectors). The SSE for the sample mean estimator is nearly twice as large.¹

1. See johnstachurski.net/emet.html for the code that generates these results.

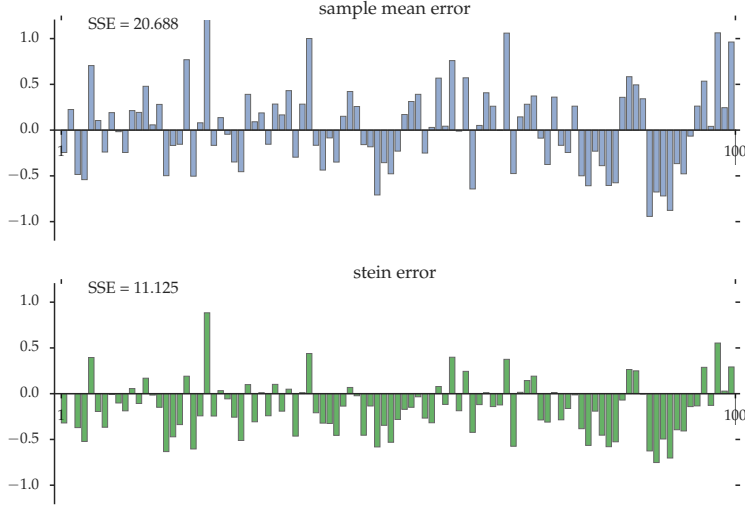


Figure 9.5 Error vectors for the two estimators

9.2.4 Asymptotic Properties

Let's now turn to asymptotic properties of estimators. Asymptotic theory concerns whether or not the sampling distribution of a given estimator increasingly concentrates around the feature we wish to estimate, the rate at which this occurs, and the shape of the sampling distribution in large samples.

A fundamental concept in asymptotic statistics is consistency. Let $\{\hat{\gamma}_N\}$ be a sequence of estimators of a given feature γ , based on a sample of size N . We say that $\hat{\gamma}_N$ is **consistent** for γ if

$$\hat{\gamma}_N \xrightarrow{P} \gamma \quad \text{as } N \rightarrow \infty \quad (9.13)$$

The definition is the same whether the estimators are vector or scalar-valued. The statement that $\hat{\gamma}_N$ is consistent means more correctly that the sequence $\{\hat{\gamma}_N\}$ is consistent.

As for bias, whether consistency holds depends not just on $\hat{\gamma}_N$ but also on the joint distribution of the data. If we say that $\hat{\gamma}_N$ is consistent over a class of distributions $\mathcal{P}$, we mean that (9.13) holds true for any distribution in $\mathcal{P}$.

Example 9.2.8 The sample mean $\bar{x}_N$ of any IID sample is consistent for the mean over the class of distributions on $\mathbb{R}$ with finite first moment. Indeed, if P is such a

distribution and $x_1, \dots, x_N$ are IID draws from P , then the law of large numbers gives

$$\frac{1}{N} \sum_{n=1}^N x_n \xrightarrow{P} \int sP(ds) \quad \text{as } N \rightarrow \infty$$

More generally, $\hat{\gamma}_N = \frac{1}{N} \sum_{n=1}^N h(\mathbf{x}_n)$ is consistent for $\gamma = \int h(\mathbf{s})P(d\mathbf{s})$ whenever

- (i) $\mathbf{x}_1, \dots, \mathbf{x}_N$ is an IID sample with $\mathcal{L}(\mathbf{x}_n) = P$ for all n and
- (ii) $\int h(\mathbf{s})P(d\mathbf{s})$ exists.

(See (6.7) on page 164.) For example, the k th sample moment is consistent for the k th moment over the class of distributions with finite k th moment.

Here's an example of consistency that requires a bit more work.

Example 9.2.9 For any IID sample $x_1, \dots, x_N$ with finite variance, the sample variance s_N^2 is consistent for the variance. This can be established from the expression

$$s_N^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 - (\bar{x}_N - \mu)^2 \quad (9.14)$$

where μ is the common mean (see ex. 9.4.1 and its solution). Applying the law of large numbers to (9.14), the first term converges to σ^2 , while $(\bar{x}_N - \mu) \xrightarrow{P} 0$, and hence $(\bar{x}_N - \mu)^2 \xrightarrow{P} 0$ by fact 6.1.1 on page 161. Consistency of s_N^2 follows.

Fact 9.2.2 If $\hat{\gamma}_N$ is consistent for γ and g is any continuous function, then $g(\hat{\gamma}_N)$ is consistent for $g(\gamma)$.

Fact 9.2.2 is immediate from fact 6.1.1 on page 161.

Example 9.2.10 The sample standard deviation $s_N = \sqrt{s_N^2}$ is consistent for the standard deviation whenever s_N^2 is consistent for the variance.

Reasonable estimators should be consistent, at least for the IID case. In fact, as we saw in theorem 8.1.1, the entire unknown distribution can be recovered asymptotically with repeated independent samples.

9.2.4.1 Asymptotic Distributions

In the IID Gaussian setting of example 9.1.3, the distribution of the sample mean is $N(\mu, \sigma^2/N)$. Beyond the information contained in the mean and variance, the fact that the sampling distribution is normal gives us important additional knowledge. We know that the distribution is symmetric, that it has thin tails, and that it puts

around 95% of its mass within two standard deviations of the mean. We can determine probabilities of intervals or the expectation of any function of $\bar{x}_N$.

The only concern is our assumption that the observations are Gaussian. Economic theory rarely provides any convincing basis for this assumption. Fortunately, we have the central limit theorem to fall back on, which tells us (remarkably!) that the sample mean is always asymptotically Gaussian, provided that the underlying observations have finite second moment. Even better, the same approximation is valid for many estimators beyond the sample mean.

To clarify, let $\{\hat{\gamma}_N\}$ be a sequence of estimators for some feature γ . We say that $\hat{\gamma}_N$ is **asymptotically normal** if there exists a positive definite matrix Σ such that

$$\sqrt{N}(\hat{\gamma}_N - \gamma) \xrightarrow{d} N(\mathbf{0}, \Sigma) \quad \text{as } N \rightarrow \infty \quad (9.15)$$

When (9.15) holds, Σ is called the **asymptotic variance–covariance matrix** of $\hat{\gamma}_N$.

Example 9.2.11 Let $x_1, \dots, x_N$ be IID with common mean μ and variance σ^2 . If $\mathbb{E}[x_n^4] < \infty$, then the sample variance s_N^2 is asymptotically normal with

$$\sqrt{N}(s_N^2 - \sigma^2) \xrightarrow{d} N(0, m_4 - \sigma^4) \quad \text{where } m_4 := \mathbb{E}[(x_n - \mu)^4] \quad (9.16)$$

To show this, we can modify (9.14) to get

$$\sqrt{N}(s_N^2 - \sigma^2) = \sqrt{N} \left[\frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 - \sigma^2 \right] - \sqrt{N}(\bar{x}_N - \mu)^2 \quad (9.17)$$

The last term on the right-hand side of (9.17) converges to zero in probability, since

$$\sqrt{N}(\bar{x}_N - \mu)^2 = a_N b_N \quad \text{where } a_N := \sqrt{N}(\bar{x}_N - \mu), \quad b_N := \bar{x}_N - \mu$$

By the CLT, the LLN and fact 6.1.8 on page 168, we have $a_N b_N \xrightarrow{p} 0$.

To complete the proof, let $Y_n := (x_n - \mu)^2$. The first term in (9.17) can then be written as $\sqrt{N}(\bar{Y}_N - \mathbb{E}[Y_n])$. Applying the CLT, this expression converges to a zero-mean normal with variance

$$\text{var } Y_n = \mathbb{E}[(Y_n - \sigma^2)^2] = \mathbb{E}[(x_n - \mu)^4 - 2(x_n - \mu)^2 \sigma^2 + \sigma^4]$$

The claim follows.

Example 9.2.12 Under the same assumptions as example 9.2.11, the sample standard

deviation is also asymptotically normal, with

$$\sqrt{N}(s_N - \sigma) \xrightarrow{d} N\left(0, \frac{m_4 - \sigma^4}{4\sigma^2}\right) \quad (9.18)$$

Exercise 9.4.10 asks you to confirm this.

Fact 9.2.3 Asymptotic normality implies consistency. In particular, if $\{\hat{\gamma}_N\}$ satisfies (9.15), then $\hat{\gamma}_N \xrightarrow{p} \gamma$ as $N \rightarrow \infty$ (see ex. 9.4.12).

Asymptotic normality also implies a rate at which $\hat{\gamma}_N$ converges to γ . Indeed, if $\hat{\gamma}_N$ is asymptotically normal, then $\sqrt{N}(\hat{\gamma}_N - \gamma)$ does not diverge. This means that $\hat{\gamma}_N - \gamma$ goes to zero fast enough to offset the diverging term $\sqrt{N}$. To emphasize this point, we say that an asymptotically normal estimator $\hat{\gamma}_N$ is **$\sqrt{N}$ -consistent** for γ .

Still more significant is the fact that asymptotic normality gives us an approximation to the entire sampling distribution. This will provide us with a means of performing inference. Chapter 10 gives details.

9.2.5 Decision Theory

We've now reviewed the most common methods of assessing estimators. Next we'll look more deeply at the theory of estimation and decisions under uncertainty. Our focus will be on finite sample properties. This is the most interesting case, since it captures our need to generalize on the basis of limited information.

Despite the efforts of many brilliant statisticians and econometricians, we still lack a convincing general theory of optimality of estimators that rests on purely objective criteria. For example, while the theory of minimum variance unbiased estimators initially seems attractive because it can potentially yield estimators without any subjective judgment on the part of the statistician, there are simple settings where minimum variance unbiased estimators are *inadmissible*, which means that they fail a minimal requirement for what constitutes a sensible estimator (see below).

Without a fully objective theory of "good" estimators, there are competing notions of what good means, and we are forced to take a stand on what our criteria for assessing estimators are in different situations. This means that the estimation problem is not fully specified until we make our preferences explicit.

For economists it's natural to think of specifying preference in terms of utility or loss. To fit with modern statistical terminology, we will think in terms of loss. Thus a full description of an estimation problem includes specification of the loss incurred when an estimator performs poorly.

Let's frame these ideas in line with the seminal work of Wald (1939) and subsequent research on decision theory. A general decision problem consists of

- (i) a sample space $\mathcal{X}$,
- (ii) an action space $\mathcal{A}$,
- (iii) a set $\mathcal{D}$ of decision rules, which are maps $d: \mathcal{X} \rightarrow \mathcal{A}$
- (iv) a universe Θ of possible states of the world, with typical element θ ,
- (v) a set of probability distributions $\{P_\theta\}$ over the sample space $\mathcal{X}$, and
- (vi) a loss function $L: \mathcal{A} \times \Theta \rightarrow \mathbb{R}$, with interpretation

$L(a, \theta)$ = loss incurred by choosing action a when the state of the world is θ .

The state of the world is treated as unknown. The problem is to choose a decision rule $d \in \mathcal{D}$ that generates low loss in some sense to be made precise. Central to this discussion is the notion of **risk**, which is defined as

$$\mathcal{R}(d, \theta) := \int L[d(x), \theta] P_\theta(dx) \quad (9.19)$$

The integral is over all $x \in \mathcal{X}$. The notation $\mathcal{R}$ is used to distinguish this decision-theoretic concept of risk from $R(f)$, the prediction risk of policy f as defined in (8.16).

Example 9.2.13 Let π be an inflation rate and let x be a noisy signal of inflation received by firms. Each firm responds by choosing a price $p = p(x)$ for their product. Let $\Pi(p, \pi)$ represent profits from choosing price p when the inflation rate is π , and let loss be negative profits (see Drenik and Perez (2015) for details and motivation). The risk is then

$$\mathcal{R}(p, \theta) = - \int \Pi(p(x), \pi) P_\theta(dx, d\pi)$$

If the distribution P_θ is known to the firm (think of $\{P_\theta\}$ as a singleton), then we have a standard problem of choosing price to maximize expected profits under known probabilities for outcomes. If, however, P_θ is not known, we have a problem of **Knightian uncertainty**. The risk is unknown because we have uncertainty over probabilities.

The problem of choosing estimators is a special case of the decision theory described above. Suppose that we observe data $\mathbf{z}_D$ taking values in Z_D with joint distribution P_θ indexed by $\theta \in \Theta$. (The set Θ can be infinite dimensional and θ is just an index.) We wish to estimate some S -valued feature $\gamma_\theta = \gamma(P_\theta)$ using this data. An estimator in this context is a decision rule $\hat{\gamma}$ mapping Z_D to the action space $\mathcal{A} = S$. If we specify the loss $L(\hat{\gamma}, \theta)$ of choosing $\hat{\gamma}$ when the state of the world is θ , our risk becomes

$$\mathcal{R}(\hat{\gamma}, \theta) = \int_{Z_D} L(\hat{\gamma}(\mathbf{z}), \gamma_\theta) P_\theta(d\mathbf{z}) \quad (9.20)$$

Example 9.2.14 Suppose we want to estimate expected returns $\mu_\theta := \int sP_\theta(ds)$ to holding an asset based on IID observations $x_1, \dots, x_N$ from P_θ . With quadratic loss for prediction error and the sample mean as our estimator, the risk is

$$\mathcal{R}(\bar{x}_N, \theta) = \mathbb{E}_\theta \left[(\bar{x}_N - \mu_\theta)^2 \right] \quad (9.21)$$

Since $\bar{x}_N$ is unbiased for μ_θ , this evaluates to the variance of $\bar{x}_N$ under P_θ . By (9.8), this is equal to $\sigma_\theta^2/N = \int (s - \mu_\theta)^2 P_\theta(ds)/N$.

In the preceding example, the risk of the estimator is its MSE. More generally, if $\hat{\gamma}$ and γ in (9.20) are vectors and L measures squared vector deviation, then $\mathcal{R}(\hat{\gamma}, \theta) = \text{mse}(\hat{\gamma}, \gamma)$. Thus, risk gives us a generalization of the idea of evaluating estimators based on MSE, as described in §9.2.3. Since loss is subjective, there is no reason we should confine ourselves to quadratic loss.

Example 9.2.15 As an estimate of central tendency of housing prices in a given location, it's common to use the sample median rather the sample mean. This is because large upper tails distort the price faced by a "typical buyer." The choice of median over mean can be understood as minimization of a different loss function when estimating price given pricing data. See example 9.2.18.

9.2.5.1 Choosing Decision Rules

In the current setup, a natural idea for making optimal decisions—and for choosing optimal estimators, in particular—is to choose decision rules with low risk. The problem with this approach is that the risk depends, in general, on the unknown state of the world θ , and hence the risk-minimizing decision rule also depends on the unknown value θ . In other words, we face Knightian uncertainty. We can't evaluate risk because we don't know the probabilities.

Example 9.2.16 Let's return to the setting of example 9.2.14. The risk of the sample mean is equal to σ_θ^2/N for every value of θ . Consider now a new estimator $\hat{\gamma}$ that is identically equal to 1. In other words, we are predicting 100% return on our investment. This is a great estimator when true mean returns are 1, since

$$\mu_\theta = 1 \quad \implies \quad \mathcal{R}(\hat{\gamma}, \theta) = \mathbb{E} \left[(1 - \mu_\theta)^2 \right] = (1 - 1)^2 = 0$$

Compared with (9.21), we have $\mathcal{R}(\hat{\gamma}, \theta) < \mathcal{R}(\bar{x}_N, \theta)$ whenever $\sigma_\theta > 0$. But if μ_θ diverges sufficiently from 1, then

$$\mathcal{R}(\hat{\gamma}, \theta) = (1 - \mu_\theta)^2 > \frac{\sigma_\theta^2}{N} = \mathcal{R}(\bar{x}_N, \theta)$$

This example illustrates the fact that we cannot, in general, choose estimators based on lowest risk in a classical setting. One useful thing we can do, however, is exclude estimators that are always dominated in terms of risk. These are the so-called inadmissible estimators. In particular, a decision rule d is called **inadmissible** for the general decision problem if there exists a second rule e such that

- (i) $\mathcal{R}(e, \theta) \leq \mathcal{R}(d, \theta)$ for all $\theta \in \Theta$ and
- (ii) $\mathcal{R}(e, \theta) < \mathcal{R}(d, \theta)$ for at least one $\theta \in \Theta$.

A decision rule that is not inadmissible is called **admissible**

Example 9.2.17 Equation (9.12) tells us that the vector sample mean is inadmissible as an estimator of the mean when L is quadratic and the universe of distributions is the multivariate Gaussians.

Another idea that comes out of the desire to minimize risk is minimax. A decision rule d_m is called a **minimax** rule if

$$r(d_m) \leq r(d) \text{ for all } d \in \mathcal{D} \quad \text{where} \quad r(d) := \sup_{\theta \in \Theta} \mathcal{R}(d, \theta)$$

A minimax rule is one that performs well in the worst possible state of the world. Minimax decision rules are an important component of robust control and the study of policy under uncertainty. See, for example, Hansen and Sargent (2008). On the other hand, minimax decision rules can be inadmissible (see, e.g., Young and Smith 2005, p. 9).

9.2.5.2 Bayes Rules and Bayes Risk

The Bayesian approach to statistics leads naturally to a coherent and attractive decision theory. The horse race is not entirely fair, however, since we are granted additional information expressed as a prior. Let's review the main ideas.

Consider again the generic decision problem discussed above, but now with a prior distribution $\pi = \pi(\theta)$ over states of the world. In addition each P_θ is now restricted to be a density $p(x | \theta)$. As in §8.3.4, we obtain the posterior for θ from the prior given data x via Bayes' law:

$$\pi(\theta | x) = \frac{p(x | \theta)\pi(\theta)}{p(x)} \quad (x \in \mathcal{X}, \theta \in \Theta)$$

Given decision rule d with risk function $\mathcal{R}(d, \theta)$, the **posterior risk** or **Bayes risk** of d is

$$r_\pi(d) := \int \mathcal{R}(d, \theta)\pi(\theta) d\theta$$

A decision rule d that minimizes the Bayes risk is called a **Bayes rule**. If d is an estimator, then a Bayes rule is called a **Bayes estimator**. Existence of such rules is not guaranteed, but in many practical instances they are well defined and unique. We give some simple examples below.

There are two particularly attractive features of Bayes rules, and Bayes estimators in particular. One is that Bayes rules are always admissible under mild regularity conditions. See, for example, theorem 10.4.2 of Casella and Berger (2002). A second is that Bayes rules can be computed as a function of observed data, meaning that we don't have to concern ourselves with how to act in situations that potentially never occur.

To understand the second point, note that we can write the joint density of (x, θ) as either $p(x | \theta)\pi(\theta)$ or $\pi(\theta | x)p(x)$. Using a change in order of integration followed by this identity, the Bayes risk can be expressed as

$$\begin{aligned} r_\pi(d) &= \int_{\Theta} \left\{ \int_{\mathcal{X}} L(d(x), \theta) p(x | \theta) dx \right\} \pi(\theta) d\theta \\ &= \int_{\mathcal{X}} \left\{ \int_{\Theta} L(d(x), \theta) p(x | \theta) \pi(\theta) d\theta \right\} dx \\ &= \int_{\mathcal{X}} \left\{ \int_{\Theta} L(d(x), \theta) \pi(\theta | x) d\theta \right\} p(x) dx \end{aligned}$$

It follows from this last expression that d will be a Bayes rule whenever

$$d(x) \in \operatorname{argmin}_{a \in \mathcal{A}} \int_{\Theta} L(a, \theta) \pi(\theta | x) d\theta \quad \text{for all } x \in \mathcal{X} \quad (9.22)$$

This is akin to the scenario encountered in dynamic programming after applying Bellman's principle of optimality: once we know the value function, we can choose the optimal action just by responding to any observed state.

Example 9.2.18 Let θ be scalar, let $L(a, \theta) = (a - \theta)^2$ and let $\pi(\theta | x)$ be the posterior distribution of θ given data x . The Bayes estimator of θ is, by (9.22), the solution to

$$\min_{a \in \mathbb{R}} \int (a - \theta)^2 \pi(\theta | x) d\theta$$

As shown in a different context in exercise 8.5.1 on page 244, the minimizer is the mean of $\pi(\theta | x)$. In other words, with quadratic loss, the Bayes estimator is the mean of the posterior distribution.

Example 9.2.19 If we repeat the setting of example 9.2.18 but with the absolute loss

function $L(a, \theta) = |a - \theta|$ instead of quadratic loss, the Bayes estimator of θ is

$$\operatorname{argmin}_{a \in \mathbb{R}} \int |a - \theta| \pi(\theta | x) d\theta$$

As exercise 9.4.14 asks you to show, the solution is the median of $\pi(\theta | x)$.

9.3 Further Reading

There are many excellent general texts on statistical inference, including Casella and Berger (2002), Wasserman (2013), and Young and Smith (2005). For a discussion of estimation theory from the perspective of machine learning, see Friedman et al. (2009) or Abu-Mostafa et al. (2012).

9.4 Exercises

Ex. 9.4.1 Prove (9.14) in example 9.2.9.²

Ex. 9.4.2 Confirm fact 9.1.1 on page 249.³

Ex. 9.4.3 Verify the claim in (9.11) of example 9.2.7.

Ex. 9.4.4 Let $x_1, \dots, x_N$ be IID random variables uniformly distributed on $[0, \theta]$. Consider the estimator of θ defined by $\hat{\theta}_N := \frac{2}{N} \sum_{n=1}^N x_n$.

- (i) Calculate the bias of $\hat{\theta}_N$.
- (ii) Calculate the mean squared error of $\hat{\theta}_N$.
- (iii) Is $\hat{\theta}_N$ consistent? Why or why not?

Ex. 9.4.5 As in ex. 9.4.4, let $x_1, \dots, x_N$ be IID draws from $U[0, \theta]$, where θ is an unknown positive parameter. Let $\mu := \mathbb{E}[x_n]$. Show that the estimator

$$\hat{\mu} := \frac{N+1}{2N} \cdot \max\{x_1, \dots, x_N\}$$

is an unbiased estimator of μ .⁴

Ex. 9.4.6 Let $x_1, \dots, x_N$ be IID and standard normal. Let $\bar{x}$ be the sample mean. Show that $\bar{x}$ and $x_i - \bar{x}$ are independent for $i = 1, \dots, N$.

2. Hint: Review example 2.2.3 on page 33.

3. Hints: Look at fact 5.1.13 on page 137 and the proofs following example 9.1.4 on page 249.

4. Hint: If you're stuck, review ex. 5.4.6 and its solution.

Ex. 9.4.7 Following on from ex. 9.4.6, let $x_1, \dots, x_N$ be IID and standard normal, let $\bar{x}$ be the sample mean, and s^2 be the sample variance. Show that $\bar{x}$ and s^2 are independent.

Ex. 9.4.8 Confirm (9.10): Show that, for any estimator $\hat{\gamma}$ of γ , we have $\text{mse}(\hat{\gamma}, \gamma) = \text{var } \hat{\gamma} + \text{bias}(\hat{\gamma}, \gamma)^2$.

Ex. 9.4.9 Let $\hat{\gamma}_N$ be an estimator of $\gamma \in \mathbb{R}$, based on a sample of size N .

- (i) Is it true that if $\text{mse}(\hat{\gamma}_N, \gamma) \rightarrow 0$ as $N \rightarrow \infty$, then $\hat{\gamma}_N$ is consistent for γ ? Why or why not?
- (ii) Is it true that if $\hat{\gamma}_N$ is asymptotically unbiased and $\text{var } \hat{\gamma}_N \rightarrow 0$, then $\hat{\gamma}_N$ is consistent for γ ? Why or why not?

Ex. 9.4.10 Assuming the conditions of examples 9.2.11–9.2.12, show that (9.18) on page 265 is valid.

Ex. 9.4.11 Let $x_1, \dots, x_N$ be IID with mean μ and variance σ^2 . Let $\bar{x}_N$ be the sample mean, and let σ_N be a consistent estimator of σ . What is the limiting distribution of

$$y_N := N \left(\frac{\bar{x}_N - \mu}{\sigma_N} \right)^2$$

Ex. 9.4.12 Let $\hat{\theta}_N$ be an estimator of θ . Show that if $\hat{\theta}_N$ is asymptotically normal, then $\hat{\theta}_N$ is consistent for θ . (You can restrict attention to the scalar case.)

Ex. 9.4.13 Consider the following estimation problem. Let $\{p_\theta\}_{\theta \in \Theta}$ be a parametric class of densities. Now let a particular $\theta_0 \in \Theta$ be fixed, and suppose that an asymptotically normal sequence of estimators $\{\hat{\theta}_N\}$ exists for it, in the sense that

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

for some positive definite Σ . Fix a point y and consider $p(y, \hat{\theta}_N)$ as an estimator of $p(y, \theta_0)$. State conditions under which this estimator is consistent and asymptotically normal. Derive the asymptotic distribution under these conditions.

Ex. 9.4.14 Confirm the claim in example 9.2.19 that the minimizer is the median of $\pi(\theta | x)$. In other words, it is the value $a \in \mathbb{R}$ that satisfies

$$\int_{-\infty}^a \pi(\theta | x) = \int_a^{\infty} \pi(\theta | x) = \frac{1}{2} \quad (9.23)$$

9.4.1 Solutions to Selected Exercises

Solution to Ex. 9.4.1. Let s_N^2 , $\mathbf{x}$ and μ be as stated in the exercise. Here's one way to prove (9.14). Let $S = \text{span}\{\mathbf{1}\}$, let $\mathbf{P} = \text{proj } S$, and let $\mathbf{M}$ be the residual projection. Recalling example 2.2.3 on page 33, when $\mathbf{y} := \mathbf{x} - \mu\mathbf{1}$, we have $\mathbf{P}\mathbf{y} = \bar{y}\mathbf{1}$ and $\mathbf{M}\mathbf{y} = \mathbf{y} - \bar{y}\mathbf{1}$. Applying $\|\mathbf{y}\|^2 = \|\mathbf{P}\mathbf{y}\|^2 + \|\mathbf{M}\mathbf{y}\|^2$ now gives

$$\sum_{n=1}^N (x_n - \mu)^2 = N(\bar{x}_N - \mu)^2 + \sum_{n=1}^N (x_n - \bar{x}_N)^2$$

Some minor rearrangements give (9.14). $\square$

Solution to Ex. 9.4.2. Let $\mathcal{L}(\mathbf{x}) = N(\mu\mathbf{1}, \sigma^2\mathbf{I})$ and let $\boldsymbol{\xi} := \sigma^{-1}(\mathbf{x} - \mu\mathbf{1})$. Since the images of a standard normal random vector under a projection and its residual are independent (fact 5.1.13) and since arbitrary deterministic functions preserve independence (fact 5.1.10), it is enough to find functions f_1 and f_2 and a subspace S and its corresponding projection $\mathbf{P}$ and residual $\mathbf{M}$ such that

$$\bar{x}_N = f_1(\mathbf{P}\boldsymbol{\xi}) \quad \text{and} \quad s_N^2 = f_2(\mathbf{M}\boldsymbol{\xi}) \quad (9.24)$$

To this end, recall from example 2.2.3 (page 33) that if $S := \text{span}\{\mathbf{1}\}$ and $\mathbf{P} = \text{proj } S$, then $\mathbf{P}\mathbf{x} = \bar{x}_N\mathbf{1}$. In addition, $\mathbf{P}\mathbf{1} = \mathbf{1}$. From this you will be able to show that $\bar{x}_N = \sigma\mathbf{e}_1^T\mathbf{P}\boldsymbol{\xi} + \mu$, where $\mathbf{e}_1$ is the first canonical basis vector. This confirms the first part of (9.24). For the second part, apply (9.5) on page 249 to get

$$s_N^2 = \frac{\sigma^2\boldsymbol{\xi}^T\mathbf{M}\boldsymbol{\xi}}{N} \quad \square$$

Solution to Ex. 9.4.5. Let $z := \max\{x_1, \dots, x_N\}$. By the definition of the maximum, we have

$$z \leq s \quad \text{if and only if} \quad x_1 \leq s \text{ and } x_2 \leq s \text{ and } \dots \text{ and } x_N \leq s$$

$$\therefore \quad \{z \leq s\} = \cap_{n=1}^N \{x_n \leq s\}$$

$$\therefore \quad \mathbb{P}\{z \leq s\} = \mathbb{P} \cap_{n=1}^N \{x_n \leq s\} = \prod_{n=1}^N \mathbb{P}\{x_n \leq s\}$$

For $s \in [0, \theta]$, the CDF of x_n is $F(s) = s/\theta$, and we obtain

$$\mathbb{P}\{z \leq s\} = \left(\frac{s}{\theta}\right)^N$$

Differentiating shows that density p of z on the interval $[0, \theta]$ is $p(s) = N \frac{s^{N-1}}{\theta^N}$. The expectation of z is therefore $\int_0^\theta s p(s) ds = \frac{N}{N+1} \theta$. We conclude that

$$\mathbb{E}[\hat{\mu}] = \frac{N+1}{2N} \mathbb{E}[z] = \frac{N+1}{2N} \frac{N}{N+1} \theta = \frac{\theta}{2} = \mu$$

In other words, the estimator is unbiased. $\square$

Solution to Ex. 9.4.8. Adding and subtracting $\mathbb{E}[\hat{\gamma}]$, we get

$$\text{mse}(\hat{\gamma}, \gamma) := \mathbb{E}[(\hat{\gamma} - \gamma)^2] = \mathbb{E}[(\hat{\gamma} - \mathbb{E}[\hat{\gamma}] + \mathbb{E}[\hat{\gamma}] - \gamma)^2]$$

Expanding the square and minor manipulations yield the desired result. $\square$

Solution to Ex. 9.4.10. The aim is to verify the claim

$$\sqrt{N}(s_N - \sigma) \xrightarrow{d} N\left(0, \frac{m_4 - \sigma^4}{4\sigma^2}\right)$$

under the conditions of example 9.2.12 on page 264. Letting $g(x) = \sqrt{x}$ and applying theorem 6.2.2 on page 173 along with (9.16), we have

$$\sqrt{N}(s_N - \sigma) = \sqrt{N}(g(s_N^2) - g(\sigma^2)) \xrightarrow{d} N\left(0, g'(\sigma^2)^2(m_4 - \sigma^4)\right)$$

The result follows. $\square$

Solution to Ex. 9.4.11. Let

$$w_N := \sqrt{N} \frac{\bar{x}_N - \mu}{\sigma_N} = \frac{\sigma}{\sigma_N} \sqrt{N} \frac{\bar{x}_N - \mu}{\sigma}$$

Since $\sigma_N \xrightarrow{p} \sigma$ by assumption, fact 6.1.1 on page 161 yields $\sigma/\sigma_N \xrightarrow{p} \sigma/\sigma = 1$. Applying the central limit theorem and Slutsky's theorem (fact 6.1.7 on page 168) together, we then have $w_N \xrightarrow{d} z \sim N(0, 1)$. By the continuous mapping theorem (fact 6.1.6 on page 167), $y_N = w_N^2$ converges in distribution to z^2 . As discussed in §5.1.6, the distribution of z^2 is $\chi^2(1)$. $\square$

Solution to Ex. 9.4.12. Fix $\delta > 0$. It suffices to show that for any positive number ϵ we have

$$\lim_{N \rightarrow \infty} \mathbb{P}\{|\hat{\theta}_N - \theta| > \delta\} \leq \epsilon \quad (9.25)$$

(Note: If a is any nonnegative real number and $a \leq \epsilon$ for any $\epsilon > 0$, then $a = 0$.) To establish (9.25), fix $\epsilon > 0$. Let z be zero-mean Gaussian with variance equal to the

asymptotic variance of $\hat{\theta}_N$. Choose M such that $\mathbb{P}\{|z| \geq M\} \leq \epsilon$. For N such that $\sqrt{N}\delta \geq M$ we have

$$\mathbb{P}\{|\hat{\theta}_N - \theta| > \delta\} = \mathbb{P}\{\sqrt{N}|\hat{\theta}_N - \theta| > \sqrt{N}\delta\} \leq \mathbb{P}\{\sqrt{N}|\hat{\theta}_N - \theta| > M\}$$

Taking $N \rightarrow \infty$, applying asymptotic normality, the continuous mapping theorem (fact 6.1.6 on page 167) and the definition of M gives (9.25). $\square$

Solution to Ex. 9.4.13. First observe that, under the conditions of the exercise, $p(y, \hat{\theta}_N)$ is consistent for $p(y, \theta_0)$ whenever $\theta \mapsto p(y, \theta)$ is continuous at θ_0 . Indeed, since $\hat{\theta}_N$ is asymptotically normal, it is also consistent. Under this continuity assumption, in probability convergence of $p(y, \hat{\theta}_N)$ to $p(y, \theta_0)$ now follows from fact 6.1.1 on page 161.

Regarding asymptotic normality, suppose that $\theta \mapsto p(y, \theta)$ is differentiable at θ_0 , and that the vector of partial derivatives $\nabla_{\theta} p(y, \theta_0)$ is not the zero vector. Applying (6.14) from page 173, we then have

$$\sqrt{n}\{p(y, \hat{\theta}_N) - p(y, \theta_0)\} \xrightarrow{d} N(0, \nabla_{\theta} p(y, \theta_0)^{\top} \Sigma \nabla_{\theta} p(y, \theta_0)) \quad \square$$

Solution to Ex. 9.4.14. The minimization problem in example 9.2.19 can alternatively be written as

$$\operatorname{argmin}_{a \in \mathbb{R}} \left\{ \int_{-\infty}^a (a - \theta) \pi(\theta | x) d\theta + \int_a^{\infty} (\theta - a) \pi(\theta | x) d\theta \right\}$$

Differentiating with respect to a , setting the result to zero, and using the fact that the posterior is a density gives (9.23). $\square$