A Primer in Econometric Theory

Lecture 8: Properties of Estimators

John Stachurski
Lectures by Akshay Shanker

March 26, 2017

Evaluating Estimators

Nothing in the definition of estimators to suggest good performance

Now we start to evaluate estimators

First step – view estimators as random entities dependent on the data

- consider distributions of estimators

Estimators and Random Elements

A statistic T is a $\mathscr{B}$-measurable transformation of the sample space $Z_{\mathcal{D}}$ into some feature space S

As a function of the sample $\mathbf{z}_{\mathcal{D}} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$, T is a random element mapping outcomes in the underlying space Ω to S via

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

"Random element" could be:

- a random vector (or scalar)
- the random distribution $\hat{P}_N$
- a density estimate, which is a random function of the sample

Example. Consider a sample mean $\bar{x}_N := \frac{1}{N}\sum_{n=1}^{N} x_n$

The observations $x_1, \ldots, x_N$ are be random variables on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$

The sample mean is the random variable

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

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

Sampling Distributions

As just discussed, $T(\mathbf{z}_{\mathcal{D}})$ is a statistic (random element) taking values in S

Its distribution $\mathcal{L}(T)$ on S can be expressed as

$$\mathcal{L}(T)(B) = \mathbb{P}\left\{ T\left(\mathbf{z}_{\mathcal{D}}\right) \in B \right\}$$

$\mathcal{L}(T)$ is called the **sampling distribution** of T

Continuing to write $P_{\mathcal{D}}$ for the joint distribution of the sample $\mathbf{z}_{\mathcal{D}}$, define the sampling distribution as

$$\mathcal{L}(T)(B) = P_{\mathcal{D}}\{\mathbf{s} \in Z_{\mathcal{D}} : T(\mathbf{s}) \in B\}$$

or even $P_{\mathcal{D}} \circ T^{-1}$

Sampling distribution of T fully determined by

1. the statistic T that maps data to feature space
2. the joint distribution $P_{\mathcal{D}}$ of the data

Example. Let $x_1, \ldots, x_N$ be IID on $\mathbb{R}$ with $P = \mathcal{L}(x_n) = \mathrm{N}(\mu, \sigma^2)$ for all n

Let T be the sample mean $\bar{x}_N$

By independence, the joint distribution $P_{\mathcal{D}}$ of $\mathbf{z}_{\mathcal{D}} = \mathbf{x} := (x_1, \ldots, x_N)$ is $\mathrm{N}(\mu \mathbf{1}, \sigma^2 \mathbf{I})$

Recall fact 5.1.5:

Let $\mathbf{x}$ be a random vector in $\mathbb{R}^N$. If $\mathcal{L}(\mathbf{x}) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathcal{L}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathrm{N}(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\intercal)$$

for all constant conformable $\mathbf{A}, \mathbf{b}$

Example. (cont.) It follows that

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

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

Example. Let $x_1, \ldots, x_N$ be IID on $\mathbb{R}$ with $P = \mathcal{L}(x_n) = \mathrm{N}(\mu, \sigma^2)$ for all n

Let T be the sample mean $\bar{x}_N$

Let s_N^2 be the sample variance. If $\sigma > 0$, then

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

Proof is an exercise (or see page 249 in ET)

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

Proof is exercise 9.4.2

Unlike previous examples, mostly not possible to obtain an analytical expression for the sampling distribution of our estimator (or a transformation of the estimator) from the underlying distribution of the data

Example. A random variable x is said to be **lognormally distributed** with parameters μ, σ if $\mathcal{L}(\ln x) = \mathrm{N}(\mu, \sigma^2)$. We write $\mathcal{L}(x) = \mathrm{LN}(\mu, \sigma^2)$

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

For example, following figure shows a histogram approximation to $\mathcal{L}(\bar{x}_N)$ when $N = 20$ and $x_1, \ldots, x_N$ is IID LN$(0, 1)$

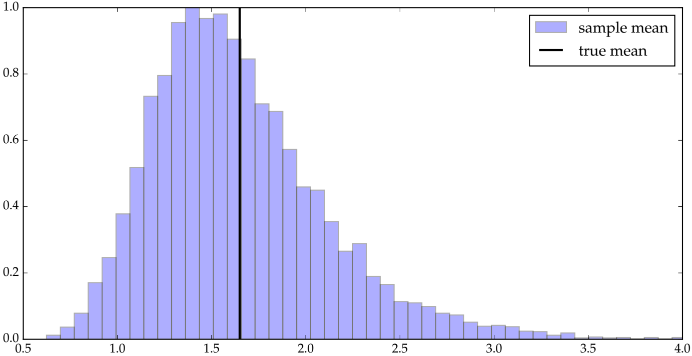

Figure: Sampling distribution of $\bar{x}_N$ when $N = 20$ and data are IID LN$(0, 1)$

```
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()
```

Desirable Properties

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

Suppose we are estimating the mean of a distribution from a random sample of size 20:

- each sample drawn from $\mathrm{LN}(0, 1)$ distribution, the mean of which is about 1.65

Consider three estimators of the mean:

- mid-range estimator
- maximum likelihood estimator
- sample mean

The last two put more probability mass in the region around 1.65 – outperform the mid-range estimator in this setting

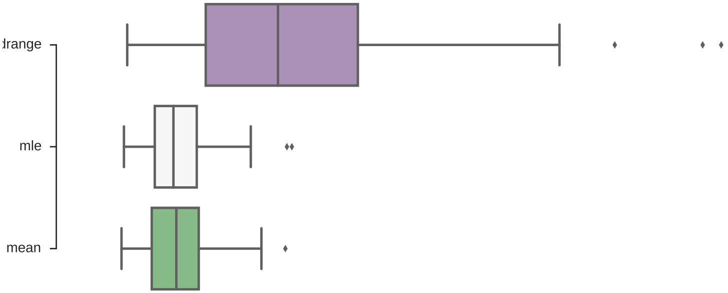

Figure: Sampling distributions of three estimators of a lognormal mean

Bootstrap

Sampling distributions in reality cannot be observed because they depend on $P_{\mathcal{D}}$

For example, suppose $x_1, \ldots, x_N$ are assumed to be $\textsc{iid}$ $\textsc{n}(\mu, \sigma^2)$ with μ and σ unknown

By (8), the distribution of $\bar{x}_N$ is $\textsc{n}(\mu, \sigma^2/N)$ — we have not pinned down the distribution because μ and σ are unknown.

In addition:

- we may not be able to easily write down the sampling distribution as a function of the parameters (IID lognormals in example above)
- we may not have assumed a parametric form for the distribution of the observations

We discuss a general technique to estimate sampling distributions given the data

Let T be any statistic and consider its sampling distribution:

$$\mathcal{L}_P(T) := \text{ the distribution of } T(x_1, \ldots, x_N)$$

when $x_1, \ldots, x_N \overset{\text{IID}}{\sim} P$

Let $x_1^o, \ldots, x_N^o$ be a sample of of IID draws from P

A good estimator of P given our sample is the empirical distribution $\hat{P}_N$ generated by $x_1^o, \ldots, x_N^o$

The **bootstrap distribution** of T:

$$\mathcal{L}_{\hat{P}_N}(T) := \text{ the distribution of } T(x_1, \ldots, x_N)$$

when $x_1, \ldots, x_N \overset{\text{IID}}{\sim} \hat{P}_N$

The bootstrap distribution is the plug-in estimator of the sampling distribution

Is $\mathcal{L}_{\hat{P}_N}(T)$ a good approximation to $\mathcal{L}_P(T)$?

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 $\mathcal{L}_P(T)$

Simulation

To generate draws from $\mathcal{L}_{\hat{P}_N}(T)$ on a computer:

```
set $\hat{P}_N =$ the empirical distribution of observed data $x_1^o, \ldots, x_N^o$
for $m$ in $1, \ldots, M$ do
    draw $x_1^b, \ldots, x_N^b$ independently from $\hat{P}_N$
    set $T_m^b = T(x_1^b, \ldots, x_N^b)$
end for
return the sample $T_1^b, \ldots, T_M^b$
```

Drawing samples from $\hat{P}_N$ easy: make repeated draws from the set $x_1^o, \ldots, x_N^o$ with equal probability on each element

Julia code to generate draws from $\mathcal{L}_{\hat{P}_N}(T)$ – implement the algorithm as a function returning M bootstrap samples

Arguments to the function:

- `xo`, the array of observed data
- `stat`, which is a function representing T

```
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
```

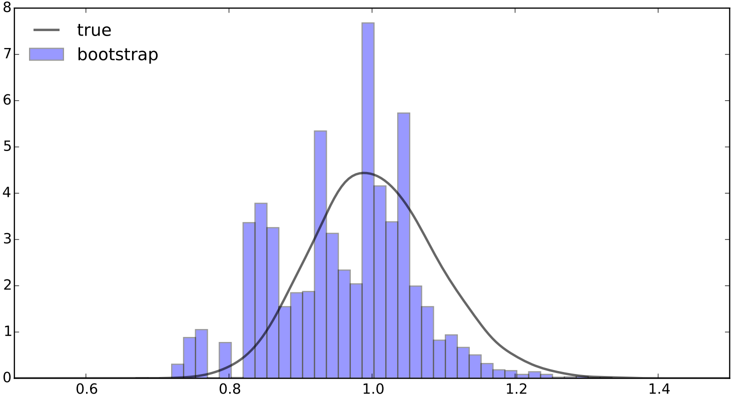

Figure: Bootstrap draws and sampling distribution for the median

Recall a good estimator is one that concentrates its probability mass close to target feature

How much probability mass concentrates usually measured by the variance or standard deviation of the sampling distribution

Replace the true standard deviation with an estimate, referred to as the **standard error**:

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

Which estimate of the standard deviation?

One way to compute a standard error of an estimator is to take the sample standard deviation of the bootstrap draws

Example. Next slide shows histogram of draws from the bootstrap distribution of the median applied to 2,118 observations of US firm sizes by sales

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

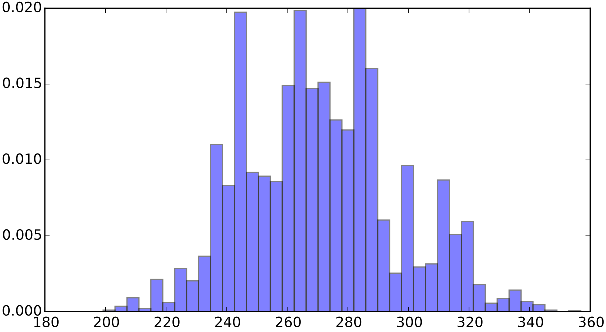

Figure: Bootstrapped draws for the median of firm sizes by sales

Evaluating Estimators

Let $\hat{\gamma}$ be an estimator of a given feature $\gamma = \gamma(P)$. The **bias** of $\hat{\gamma}$ is defined as

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

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

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

The notation $\mathbb{E}_P \hat{\gamma}$ indicates 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 is

- unbiased estimator of the mean of a uniform distribution on $\mathbb{R}$ with unknown end points
- biased for many other distributions, including the lognormal

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

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

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

Proof for fact is immediate from linearity of expectations:

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

$$= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\, h(\mathbf{x}_n) = \int h(\mathbf{s}) P(\mathrm{d}\mathbf{s}) = \gamma(P)$$

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

Example. Let $x_1, \ldots, x_N$ 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:

$$\mathbb{E}\, s_N^2 = \sigma^2 \frac{N-1}{N}$$

Variance of Estimators

Sensible estimators almost always have the property that the variance goes to zero as $N \to \infty$

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

$$\operatorname{var} \bar{x}_N = \operatorname{var}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] = \frac{\sigma^2}{N} \tag{1}$$

Example. Assume the IID Gaussian setting of the example above

We showed the sample variance satisfies
$\mathcal{L}(\sigma^{-2} N s_N^2) = \chi^2(N-1)$

The variance of a $\chi^2(k)$ distribution is $2k$

Using this fact and some algebra gives

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

Is small variance always desirable for an estimator?

If the estimator is biased, then perhaps not — probability mass might be concentrated in the wrong place

But for an unbiased estimator, low variance means probability mass is concentrated around the feature we wish to estimate

Consider any given estimator – how do we know whether its variance is too low or high?

- low variance relative

One way to approach: 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 $\mathbf{x}_1, \ldots, \mathbf{x}_N$, 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**

However

- minimizer may not exist
- may be hard to determine in practice
- may require strong assumptions on the unknown distribution of the data

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. Let $x_1, \ldots, x_N$ 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:

$$\left\{ \text{all statistics of the form } \hat{\mu} = \sum_{n=1}^{N} \alpha_n x_n, \right.$$

$$\left. \text{where } \alpha_n \in \mathbb{R} \text{ for } n = 1, \ldots, N \right\}$$

Example. (cont.) The set of linear unbiased estimators of μ:

$$U_\mu^\ell := \Bigg\{ \text{all } \hat{\mu} = \sum_{n=1}^{N} \alpha_n x_n \text{ with } \alpha_n \in \mathbb{R},$$

$$n = 1, \ldots, N \text{ and } \mathbb{E}\left[\sum_{n=1}^{N} \alpha_n x_n\right] = \mu \Bigg\}$$

Example. (cont.) 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, the variance of an element of this class is given by

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

To find the best linear unbiased estimator, solve

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

Example. (cont.) To solve this constrained optimization problem, we can use the Lagrangian, setting

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

where λ is the Lagrange multiplier

Differentiate with respect to α_n and set the result equal to zero

Example. (cont.) 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 $\alpha_n^* = 1/N$

Our estimator becomes

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

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

However, there's no convincing reason to restrict ourselves to unbiased estimators. In fact, as we'll see, there are good reasons not to

Mean Squared Error

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

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

As with bias, $\mathrm{mse}(\hat{\gamma}, \gamma)$ depends on the joint distribution of the data as well as the specification of $\hat{\gamma}$

Mean squared error takes into account both bias and variance; the scalar case:

$$\text{mse}(\hat{\gamma}, \gamma) = \operatorname{var} \hat{\gamma} + \mathsf{bias}(\hat{\gamma}, \gamma)^2 \tag{3}$$

Example. Consider an IID sample $x_1, \ldots, x_N$ with common distribution $\mathrm{N}(\mu, \sigma^2)$

Consider the sample variance s_N^2

We showed (example 9.2.4 in ET) $\operatorname{var} s_N^2 = 2\sigma^4 N^{-2}(N-1)$

Combine this result with (3) and $\mathbb{E}\, s_N^2 = \sigma^2 \frac{N-1}{N}$ to arrive at

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

Typically there is a trade-off between bias and variance to minimise MSE:

- lower variance costs more bias and vice-versa
- interior solutions: we can often reduce variance significantly by accepting a small amount of bias

Example. Let $\hat\gamma$ be any unbiased estimator of a feature $\gamma \in \mathbb{R}$

Let $v := \mathrm{var}\,\hat\gamma$

Consider the class of estimators $\{\lambda\hat\gamma : \lambda \in \mathbb{R}\}$

Value of λ that minimizes $\mathrm{mse}(\lambda\hat\gamma, \gamma)$ is

$$\lambda^* := \frac{\gamma^2}{\gamma^2 + v} \tag{4}$$

(exercise 9.4.3)

Unbiased estimator does not minimize MSE unless $v = 0$

For most sensible estimators, v converges to zero as the sample size $\rightarrow \infty$

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

Stein's Example

Consider a population of K individuals, where the kth individual has idiosyncratic ability $\mu_k \in [0, 1]$

For each agent, we observe N noisy signals of ability, each of the form $x_{kn} = \mu_k + \epsilon_{kn}$, $n = 1, \ldots, N$

The noise terms ϵ_{kn} are all independent $\mathrm{N}(0, \sigma^2)$

Objective: estimate the parameter vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ given the data from the observations x_{nk}

Assume only $\boldsymbol{\mu}$ is unknown

Vector sample mean

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

maximum likelihood estimator of ability, and best linear unbiased

However, when $K > 2$, there exists an alternative estimator $\hat{\boldsymbol{\mu}}_N$ such that

$$\mathbb{E}\left\{ \|\hat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}\|^2 \right\} < \mathbb{E}\left\{ \|\bar{\mathbf{x}}_N - \boldsymbol{\mu}\|^2 \right\} \quad \text{for all } \boldsymbol{\mu} \in \mathbb{R}^K \tag{5}$$

Estimator $\hat{\boldsymbol{\mu}}_N$ 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**:

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

Intuition from simulation:

- set $K = 200$, $N = 5$ and $\sigma = 1$
- 5 noisy observations on the ability of each of 200 agents
- for each agent, choose ability μ_k independently from a uniform distribution on $[0,1]$
- compute the estimators $\bar{\mathbf{x}}_N$ and $\hat{\boldsymbol{\mu}}_N$ and the corresponding error vectors
  See `johnstachurski.net/emet` for code

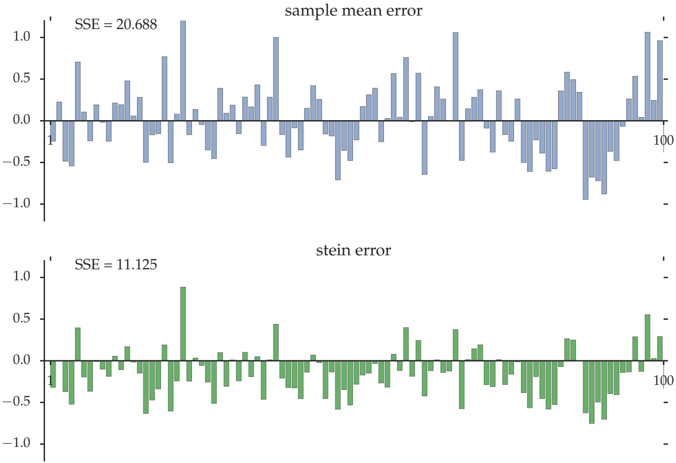

Figure: Error vectors for the two estimators

Errors for the James-Stein estimator are downward biased and also smaller than the errors of the sample mean estimator

The squared norm of the error vectors for the sample me is nearly twice as large

Asymptotic Properties

Asymptotic theory

- concerns whether or not the sampling distribution of an estimator increasingly concentrates around the feature we wish to estimate
- the rate at which an estimator “concentrates” or converges
- the shape of the sampling distribution in large samples

Let $\{\hat{\gamma}_N\}$ be a sequence of estimators of a given feature γ, based on a sample of size N

We say $\hat{\gamma}_N$ is **consistent** for γ if:

$$\hat{\gamma}_N \xrightarrow{p} \gamma \quad \text{as} \quad N \to \infty \tag{6}$$

The definition is the same whether the estimators are vector or scalar-valued

The statement that $\hat{\gamma}_N$ is consistent means 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 $\mathscr{P}$, we mean:

$$\hat{\gamma}_N \xrightarrow{p} \gamma \quad \text{as} \quad N \to \infty$$

Holds true for any distribution in $\mathscr{P}$

Example. 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 $x_1, \ldots, x_N$ are IID draws from P, then the law of large numbers gives

$$\frac{1}{N} \sum_{n=1}^{N} x_n \xrightarrow{p} \int s P(\mathrm{d}s) \quad \text{as} \quad N \to \infty$$

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

1. $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is an IID sample with $\mathcal{L}(\mathbf{x}_n) = P$ for all n and
2. $\int h(\mathbf{s}) P(\mathrm{d}\mathbf{s})$ exists.

For example, the kth sample moment is consistent for the kth moment over the class of distributions with finite kth moment

Example. For any IID sample $x_1, \ldots, x_N$ with finite variance, the sample variance s_N^2 is consistent for the variance

Can be established from:

$$s_N^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2 - (\bar{x}_N - \mu)^2 \tag{7}$$

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 in ET

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 $g(\gamma)$

Example. 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

Asymptotic Distributions

Central limit theorem tells us (remarkably!) the sample mean is always asymptotically Gaussian, provided that the underlying observations have finite second moment

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

$$\sqrt{N}(\hat{\gamma}_N - \gamma) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \Sigma) \quad \text{as} \quad N \to \infty \tag{8}$$

When (8) holds, Σ is called the **asymptotic variance–covariance matrix** of $\hat{\gamma}_N$

Example. Let $x_1, \ldots, x_N$ be IID with common mean μ and variance σ^2

If $\mathbb{E}\left[x_n^4\right] < \infty$, then the sample variance s_N^2 is asymptotically normal with

$$\sqrt{N}(s_N^2 - \sigma^2) \xrightarrow{d} \mathrm{N}(0, m_4 - \sigma^4) \tag{9}$$

where $\quad m_4 := \mathbb{E}\left[(x_n - \mu)^4\right]$

To show this, recall

$$s_N^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2 - (\bar{x}_N - \mu)^2 \tag{10}$$

Example. (cont.) We can modify this expression to get

$$
\begin{aligned}
&\sqrt{N}(s_N^2 - \sigma^2) \\
&\qquad = \sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2 - \sigma^2\right] - \sqrt{N}(\bar{x}_N - \mu)^2 \qquad (11)
\end{aligned}
$$

The last term on the right-hand side of (11) converges to zero in probability:

$$\sqrt{N}(\bar{x}_N - \mu)^2 = a_N b_N$$

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

Example. (cont.) By the CLT, the LLN and fact 6.1.8 on page 168 of ET, we have $a_N b_N \xrightarrow{p} 0$

To complete the proof, let $Y_n := (x_n - \mu)^2$

The first term in (11) can then be written as $\sqrt{N}(\bar{Y}_N - \mathbb{E}[Y_n])$

Apply the CLT, then the expression converges to a zero-mean normal with variance

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

The claim follows

Example. Under the same assumptions as the example above, the sample standard deviation is also asymptotically normal, with

$$\sqrt{N}(s_N - \sigma) \xrightarrow{d} \mathrm{N}\left(0, \frac{m_4 - \sigma^4}{4\sigma^2}\right) \tag{12}$$

See exercise 9.4.10 for a proof

Example. Asymptotic normality implies consistency. In particular, if $\{\hat{\gamma}_N\}$ satisfies

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

then $\hat{\gamma}_N \xrightarrow{p} \gamma$ as $N \to \infty$ (see ex. 9.4.12)

Asymptotic normality implies a rate at which $\hat{\gamma}_N$ converges to γ

If $\hat{\gamma}_N$ is asymptotically normal, then $\sqrt{N}(\hat{\gamma}_N - \gamma)$ does not diverge

- $\hat{\gamma}_N - \gamma$ goes to zero fast enough to offset the diverging term $\sqrt{N}$

We say that an asymptotically normal estimator $\hat{\gamma}_N$ is **$\sqrt{N}$-consistent** for γ

Important! Asymptotic normality gives us an approximation to the entire sampling distribution — means of performing inference

To be discussed in Ch. 10.

Decision Theory

We still lack convincing general theory of optimality of estimators that rests on objective criteria

For example, consider 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* – we discuss below

Without a fully objective theory of “good” estimators, we are forced to take a stand on our criteria for assessing estimators in different situations

The estimation problem is not fully specified until we make our preferences explicit

- specify preference in terms of loss
- 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

1. a sample space $\mathcal{X}$,
2. an action space $\mathcal{A}$,
3. a set $\mathscr{D}$ of decision rules, which are maps $d \colon \mathcal{X} \to \mathcal{A}$
4. a universe Θ of possible states of the world, with typical element θ,
5. a set of probability distributions $\{P_\theta\}$ over the sample space $\mathcal{X}$, and
6. a loss function $L \colon \mathcal{A} \times \Theta \to \mathbb{R}$

The loss function has the following interpretation:

$$L(a,\theta) = \text{loss from choosing action } a \text{ when state of world is } \theta$$

The state of the world treated as unknown

The problem is to choose a decision rule $d \in \mathscr{D}$ generating "low loss"

Define **risk**:

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

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 we defined in Equation (8.16) in ET

Example. (Knightian uncertainty) 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

$\Pi(p, \pi)$ represents profits from choosing price p when the inflation rate is π – let loss be negative profits

Risk:

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

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**

Problem of choosing estimators a special case of the decision theory described above

Suppose we observe data $\mathbf{z}_{\mathcal{D}}$ taking values in $Z_{\mathcal{D}}$ with joint distribution P_θ indexed by $\theta \in \Theta$ (The set Θ can be infinite dimensional.)

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 $Z_{\mathcal{D}}$ 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

$$\mathcal{R}(\hat{\gamma}, \theta) = \int_{Z_{\mathcal{D}}} L(\hat{\gamma}(\mathbf{z}), \gamma_\theta) P_\theta(\mathrm{d}\mathbf{z})$$

Example. Suppose we want to estimate expected returns $\mu_\theta := \int s P_\theta(\mathrm{d}s)$ to holding an asset based on IID observations $x_1, \ldots, x_N$ from P_θ

With quadratic loss for prediction error and the sample mean as our estimator, the risk is

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

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

In the preceding example, the risk of the estimator is its MSE

But no reason to confine ourselves to quadratic loss:

Example. Common to use the sample median rather the sample mean as an estimate of central tendency of housing prices at a location

Large upper tails distort the price faced by a "typical buyer"

Choice of median over mean can be understood as minimization of a different loss function when estimating price given pricing data

Choosing Decision Rules

Natural idea for making optimal decisions and for choosing optimal estimators: choose decision rules with low risk

However, risk depends on the unknown state of the world θ

- risk-minimizing decision rule also depends on the unknown value θ

Knightian uncertainty – we can't evaluate risk because we don't know the probabilities

Example. Consider again expected returns $\mu_\theta := \int s P_\theta(\mathrm{d}s)$ to holding an asset based on IID observations $x_1, \ldots, x_N$ from P_θ

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: predicting 100% return on our investment

Great estimator when true mean returns are 1:

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

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

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

Inadmissibility

We cannot, in general, choose estimators based on lowest risk in a classical setting

We can exclude estimators always dominated in terms of risk — inadmissible estimators

A decision rule d is called **inadmissible** if there exists a second rule e such that

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

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

Example. Consider again estimating the mean when L is quadratic and the universe of distributions is the multivariate Gaussians

Vector sample mean

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

Recall that when $K > 2$, there exists estimator $\hat{\boldsymbol{\mu}}_N$ such that

$$\mathbb{E}\left\{ \|\hat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}\|^2 \right\} < \mathbb{E}\left\{ \|\bar{\mathbf{x}}_N - \boldsymbol{\mu}\|^2 \right\} \quad \text{for all } \boldsymbol{\mu} \in \mathbb{R}^K \qquad (13)$$

Thus, in this case, the vector sample mean is inadmissible as an estimator of the mean

Minimax Rule

A decision rule d_m is called a **minimax** rule if

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

Minimax rule performs well in the worst possible state of the world

Bayes Rules and Bayes Risk

Consider now a prior distribution $\pi = \pi(\theta)$ over states of the world

In addition each P_θ is now restricted to be a density $p(x \mid \theta)$

As in §8.3.4, we obtain the posterior for θ from the prior given data x via Bayes' law:

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

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

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

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**

Attractive features of Bayes rules, and Bayes estimators in particular:

- Bayes rules are always admissible under mild regularity conditions
- 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, write the joint density of (x, θ) as either $p(x \mid \theta)\pi(\theta)$ or $\pi(\theta \mid x)p(x)$

Use a change in order of integration followed by this identity, the Bayes risk becomes:

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

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

$$d(x) \in \operatorname*{argmin}_{a \in \mathcal{A}} \int_{\Theta} L(a, \theta) \pi(\theta \mid x) \, \mathrm{d}\theta \quad \text{for all } x \in \mathcal{X} \tag{14}$$

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. Let θ be scalar, let $L(a,\theta) = (a-\theta)^2$ and let $\pi(\theta \mid x)$ be the posterior distribution of θ given data x. The Bayes estimator of θ is, by (14) above, the solution to

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

The minimizer is the mean of $\pi(\theta \mid x)$ (exercise 8.5.1 on page 244)

With quadratic loss, the Bayes estimator is the mean of the posterior distribution

Example. If we repeat the setting of above but with the absolute loss function $L(a, \theta) = |a - \theta|$ instead of quadratic loss, the Bayes estimator of θ is

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

$\pi(\theta \mid x)$ (exercise 9.4.14)