Advanced Econometric Methods
EMET3011/8014

Lecture 6

John Stachurski

Semester 1, 2011

Announcements/Reminders

- Please get yourself fresh copy of the course notes PDF
- Assignment 1
  - Posted on course homepage later today
  - Read instructions carefully
  - Due 14th April

Today's Lecture

- Estimators
- Maximum Likelihood
- Parametric vs Nonparametric
- Empirical Distributions

Estimators

Estimators are statistics—observable functions of the data—used to estimate specific quantity of interest

Example: Observations $x_1, \ldots, x_N \overset{\text{IID}}{\sim} F$ where F unknown

We wish to estimate the mean

$$\mu := \int s \, F(ds)$$

LLN suggests using sample mean $\bar{x}_N$

We are using $\bar{x}_N$ as an **estimator** of μ

Evaluating Estimators

How do we choose between estimators?

Wish to estimate mean $\mu := \int s\, F(ds)$ from $x_1, \ldots, x_N \overset{\text{IID}}{\sim} F$

Option 1: Sample mean

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

Option 2: Mid-range estimator

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

Which is a better estimator of μ?

Depends on how you define "better"

We now outline some standard notions of "goodness"

Two main types:

- Finite sample criteria
- Asymptotic criteria

Finite Sample Criteria

Let $\hat{\theta}$ be an estimator of some quantity $\theta \in \mathbb{R}$

The **bias** of $\hat{\theta}$ is $\mathbb{E}\left[\hat{\theta}\right] - \theta$

Estimator $\hat{\theta}$ called **unbiased** for θ if bias is zero

- Quantity we want to estimate is "most likely" value
- Unbiased not necessarity better than biased—see below

Example: Sample mean of $x_1, \ldots, x_N \sim F$ unbiased for mean

Example: Mid-range estimator may be biased for mean

Consider m_N when $x_1, \ldots, x_N \overset{\text{IID}}{\sim}$ lognormal and $N = 20$

Then $\mathbb{E}\,[m_N] \neq \mu := \mathbb{E}\,[x_n] = e^{1/2}$

```
# Function to compute mid-range estimator
> mr <- function(x) return((min(x) + max(x)) / 2)

# Generate 5,000 observations of m_N
> observations <- replicate(5000, mr(rlnorm(20)))

# Compare (estimated) mean of m_N with mean of x_n
> mean(observations)
[1] 3.800108
> exp(1/2)
[1] 1.648721
```

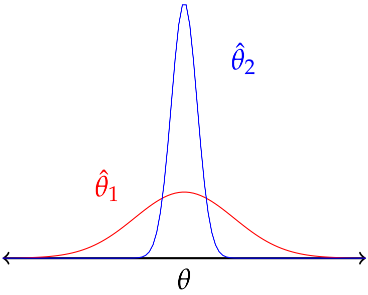
Exercise: Show that if $x_1, \ldots, x_N \stackrel{\text{IID}}{\sim} F$, then sample variance unbiased for variance

That is,

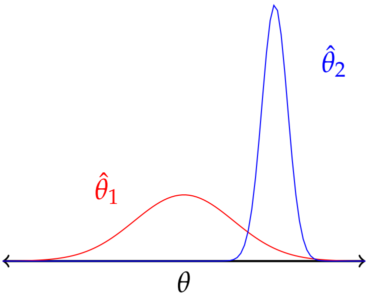
$$\mathbb{E}\left[s_N^2\right] := \mathbb{E}\left[\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2\right] = \sigma^2 := \mathrm{var}[x_n]$$

- Actually, $\mathrm{cov}[x_i, x_j] = 0$ whenever $i \neq j$ is enough

For unbiased estimator of θ, low variance is desirable



For biased estimator of θ, low variance more ambiguous...



Exercise: When unbiased, Chebychev's inequality yields

$$\mathbb{P}\{|\hat{\theta} - \theta| > \delta\} \leq \frac{\mathrm{var}[\hat{\theta}]}{\delta^2} \quad \text{for any } \delta > 0$$

Check: The argument depends on $\hat{\theta}$ being unbiased for θ.

But when is variance "low"?

- Is $\mathrm{var}[\hat{\theta}] = 23$ low or high?

One approach: Find the unbiased estimator with lowest variance

Define

$$U_\theta := \{\text{all statistics } \hat{\theta} \text{ with } \mathbb{E}[\hat{\theta}] = \theta\}$$

The **minimum variance unbiased estimator** of θ is

$$\text{MVUE} := \operatorname*{argmin}_{\hat{\theta} \in U_\theta} \operatorname{var}[\hat{\theta}]$$

Potential problems

- Might not exist
- Might be hard to find in practice
- Look up "Cramér-Rao lower bound" if you're interested

Hence often focus on smaller classes than U_θ

For example, set of linear unbiased estimators

Given sample $x_1, \ldots, x_N$, set of linear statistics given by

$$\left\{ \text{all } \hat{\theta} = \sum_{n=1}^{N} c_n x_n, \text{ where } c_n \in \mathbb{R} \text{ for } n = 1, \ldots, N \right\}$$

Let

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

Element with lowest variance called **best linear unbiased estimator** of θ:

$$\text{BLUE} := \operatorname*{argmin}_{\hat{\theta} \in U_\theta^\ell} \operatorname{var}[\hat{\theta}]$$

Example:

- Data $x_1, \ldots, x_N \overset{\text{IID}}{\sim} F$
- Wish to estimate mean $\mu \neq 0$

Set of linear unbiased estimators of μ given by

$$U_\mu^\ell := \left\{ \text{all } \hat{\mu} = \sum_{n=1}^N c_n x_n \text{ with } \mathbb{E}\left[\sum_{n=1}^N c_n x_n\right] = \mu \right\}$$

Using linearity of expectations, U_μ^ℓ can be re-written as

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

Variance of element of U^ℓ_μ given by

$$\operatorname{var}\left[\sum_{n=1}^N c_n x_n\right] = \sum_{n=1}^N c_n^2 \operatorname{var}[x_n] + 2\sum_{n<m} c_n c_m \operatorname{cov}[x_n, x_m]$$

$$= \sigma^2 \sum_{n=1}^N c_n^2 \qquad \text{where } \sigma^2 := \operatorname{var}[x_n]$$

To find the BLUE, need to solve

$$\text{minimize } \sigma^2 \sum_{n=1}^N c_n^2 \text{ over all } c_1, \ldots, c_N \text{ with } \sum_{n=1}^N c_n = 1$$

To solve use Lagrangian, setting

$$L(c_1, \ldots, c_N; \lambda) := \sigma^2 \sum_{n=1}^N c_n^2 - \lambda \left[\sum_{n=1}^N c_n - 1\right]$$

Exercise: Differentiating w.r.t. c_n and setting result to zero,

$$c_n^* = \lambda/(2\sigma^2) \qquad n = 1, \ldots, N$$

Therefore, each c_n^* takes the same value

From the constraint $\sum_n c_n^* = 1$, we have $c_n^* = 1/N$

BLUE is therefore

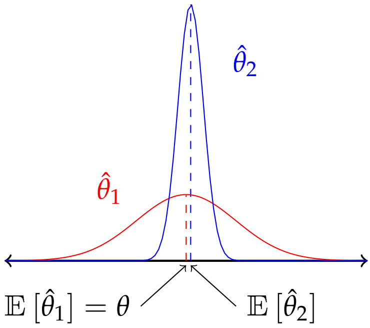
$$\sum_{n=1}^{N} c_n^* x_n = \sum_{n=1}^{N} (1/N) x_n = (1/N) \sum_{n=1}^{N} x_n =: \bar{x}$$

Conclusion: For uncorrelated data, sample mean is BLUE of μ

Note: May or may not be MVUE, depending on distribution F

Mean squared error

Unbiased estimator not necessarily better than biased one



May want to admit some bias to obtain lower variance

We need a criterion that accounts for bias and variance

Mean squared error of estimator $\hat{\theta}$ of θ is

$$\mathrm{mse}(\hat{\theta}) := \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \tag{1}$$

Low MSE implies probability mass concentrated around θ

Exercise: Decompose MSE into variance plus squared bias:

$$\mathrm{mse}(\hat{\theta}) = \mathrm{var}[\hat{\theta}] + (\mathbb{E}\left[\hat{\theta}\right] - \theta)^2 \tag{2}$$

Example: Estimating variance σ^2 from IID sample $x_1, \ldots, x_N$

$$\mathrm{mse}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2\right] \leq \mathrm{mse}\left[\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2\right]$$

Asymptotic Criteria

Consider estimator $\hat{\theta}_N$ of θ as $N \to \infty$

Key definitions:

- **asymptotically unbiased** if $\mathbb{E}\left[\hat{\theta}_N\right] \to \theta$
- **consistent** if $\hat{\theta}_N \xrightarrow{p} \theta$
- **asymptotically normal** if $\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d}$ centered normal

Example: What properties does $\bar{x}_N$ have as estimator of mean?

Example: Assume $x_1, \ldots, x_N \overset{\text{IID}}{\sim} F$ with mean μ, standard deviation σ

Sample standard deviation

$$s_N := \sqrt{s_N^2} = \left[\frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2 \right]^{1/2} \tag{3}$$

is consistent for standard deviation σ

If g is continuous and $s_N^2 \xrightarrow{p} \sigma^2$, then $g(s_N^2) \xrightarrow{p} g(\sigma^2)$

Taking $g(x) = \sqrt{x}$, we see that

$$s_N^2 \xrightarrow{p} \sigma^2 \text{ implies } s_N \xrightarrow{p} \sigma$$

Hence suffices to show that sample variance consistent for σ^2

Note that

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

But

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

Some rearranging (an exercise) gives

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

Applying various facts about convergence in probability, we get

$$
\begin{aligned}
s_N^2 &= \frac{N}{N-1} \left[\frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2 - (\mu - \bar{x}_N)^2 \right] \\
&\xrightarrow{p} 1 \times [\sigma^2 - 0] \\
&= \sigma^2
\end{aligned}
$$

Exercise: Precisely which facts are we using here?

Conclusion: For IID data,

- sample variance consistent for variance
- sample standard deviation consistent for standard deviation

Inductive Principles

How to find good estimators?

Need systematic approach that leads to good estimators

Let's start with the traditional paradigm—parametric estimation

Parametric Classes

A **parametric class of densities** is set of densities

$$\mathcal{D} = \{p_\theta : \theta \in \Theta\}$$

indexed by a vector of parameters $\theta \in \Theta \subset \mathbb{R}^K$ with $K < \infty$

Example: Let

$$p(s;\mu,\sigma) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(s-\mu)^2}{2\sigma^2}\right\} \qquad (s \in \mathbb{R})$$

and consider the set $\mathcal{D}_n$ of all normal densities:

$$\mathcal{D}_n := \{\text{all densities } p(\cdot\,;\mu,\sigma) \ \text{ s.t. } \ \mu \in \mathbb{R},\ \sigma > 0\}$$

Not all classes of densities are parametric

Example: Set $\mathcal{D}_2$ of densities with finite second moment:

$$\mathcal{D}_2 := \left\{ \text{all } p\colon \mathbb{R} \to \mathbb{R} \text{ s.t. } p \geq 0, \int p(s)ds = 1, \int s^2 p(s)ds < \infty \right\}$$

Classical estimation methods are **parametric** in nature:

- Data generated by unknown density
- Density belongs to parametric class $\mathcal{D} = \{p_\theta\}_{\theta \in \Theta}$
- We know the class, but $\theta \in \Theta$ is unknown
- Task is to estimate θ

Maximum Likelihood

One parametric approach: **Principle of maximum likelihood**

Motivation: Suppose have one draw x_1 from a normal distribution

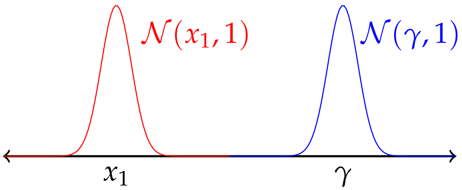
- variance known and equal to one for simplicity
- mean μ is unknown, to be estimated

How to choose $\hat{\mu}$?

A reasonable approach:

- A choice of $\hat{\mu}$ determines a density $\mathcal{N}(\hat{\mu}, 1)$
- So choose $\hat{\mu}$ such that x_1 is a likely realization from $\mathcal{N}(\hat{\mu}, 1)$

Setting $\hat{\mu} = \gamma$ makes x_1 an "unlikely" realization:



Max likelihood: Set $\hat{\mu} = x_1$

More formally:

Unknown density of x_1 is

$$p(s;\mu) := (2\pi)^{-1/2} \exp\left\{-\frac{(s-\mu)^2}{2}\right\} \qquad (s \in \mathbb{R})$$

Think of $p(x_1;u)$ as "probability" of realizing x_1 when mean $= u$

Principle of maximum likelihood: Maximize this probability

$$\hat{\mu} = \operatorname*{argmax}_{-\infty<u<\infty} p(x_1;u)$$

Exercise: Solution is $\hat{\mu} = x_1$

Same principle when $x_1, \ldots, x_N \overset{\text{IID}}{\sim} \mathcal{N}(\mu, 1)$, where μ unknown

By independence, joint density = product of marginals

Applying the principle of maximum likelihood,

1. Plug sample values into joint density
2. Maximize with respect to mean:

$$\hat{\mu} := \operatorname*{argmax}_{-\infty < u < \infty} (2\pi)^{-N/2} \prod_{n=1}^{N} \exp\left\{ -\frac{(x_n - u)^2}{2} \right\}$$

Exercise: Show that maximizer $\hat{\mu}$ is sample mean of $x_1, \ldots, x_N$

General Case

Assume $x_1, \ldots, x_N$ has joint density in class $\{p_\theta\}_{\theta \in \Theta}$

Likelihood function is p_θ evaluated at $(x_1, \ldots, x_N)$, regarded as function of θ:

$$L(\theta) := p_\theta(x_1, \ldots, x_N) \qquad (\theta \in \Theta)$$

Maximum likelihood estimate of θ is the maximizer of L:

$$\hat{\theta} := \operatorname*{argmax}_{\theta \in \Theta} L(\theta)$$

Sometimes convenient to use **log likelihood function**

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

Increasing transforms don't change maximizers:

$$\hat{\theta} = \operatorname*{argmax}_{\theta \in \Theta} \ell(\theta)$$

Example: If $x_1, \ldots, x_N \overset{\text{IID}}{\sim}$ density p_θ (marginal, not joint), then

$$L(\theta) = \prod_{n=1}^{N} p_\theta(x_n) \quad \text{and} \quad \ell(\theta) = \sum_{n=1}^{N} \ln p_\theta(x_n)$$

General comments on maximum likelihood

In theory, provides good estimators in many settings

In particular, under certain regularity conditions,

- asymptotically unbiased
- consistent
- asymptotically normal
- asymptotically "efficient" (small asymptotic variance)

However:

All these nice properties depend on the assumption

True model generating data is p_θ for some $\theta \in \Theta$

This is a huge assumption, no?

Parametric vs Nonparametric

Let's look at example to illustrate difference between parametric and nonparametric estimation

In the example, we let

$$f(s) := \frac{1}{2}(2\pi)^{-1/2} \exp\left\{-\frac{(s+1)^2}{2}\right\} + \frac{1}{2}(2\pi)^{-1/2} \exp\left\{-\frac{(s-1)^2}{2}\right\}$$

Interpretation of $x \sim f$: Flip a fair coin,

- if heads then draw x from $\mathcal{N}(-1, 1)$
- if tails then draw x from $\mathcal{N}(1, 1)$

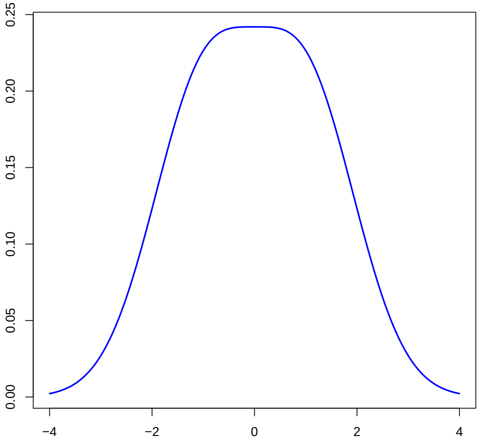


Figure: The true density f

Suppose don't know f, observe only $x_1, \ldots, x_N \overset{\text{IID}}{\sim} f$

Objective: Estimate f based on the sample

Classical method:

1. Choose parametric class for f
2. Estimate unknown parameters

But which parametric class?

If no pointers from theory, might assume $f \in \mathcal{D}_n =$ normal densities

To estimate $\hat{f} \in \mathcal{D}_n$, must determine two parameters μ and σ

Obvious way:

- estimate μ via $\hat{\mu} := \bar{x}$,
- estimate σ via $\hat{\sigma} := s$

These determine estimate $\hat{f} := \mathcal{N}(\hat{\mu}, \hat{\sigma}^2) \in \mathcal{D}_n$

Illustration: We generate $x_1, \ldots, x_{200} \overset{\text{IID}}{\sim} f$ and compute $\hat{f}$:

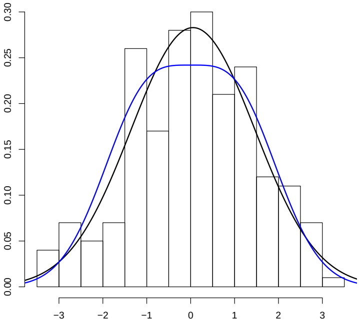


Figure: Estimate $\hat{f}$ (black line) and true f (blue line)

Is $\hat{f}$ a good estimator of f?

Not really: $\hat{f} \to f$ as $N \to \infty$ fails

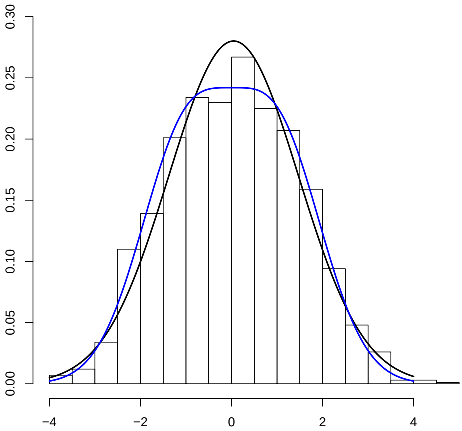


Figure: Estimating $\hat{f}$ again, sample size = 2,000

The problem: Mistaken assumption that $f \in \mathcal{D}_n$

There is no element of $\mathcal{D}_n$ that can approximate f well

A Nonparametric Estimate

Same estimation problem

Now let's not presume to know parametric class of f

How can we proceed?

One approach: a **kernel density estimator**

Letting K be a density and $\delta > 0$, define

$$\hat{f}(s) := \frac{1}{N\delta} \sum_{n=1}^{N} K\left(\frac{s - x_n}{\delta}\right) \tag{4}$$

Exercise: Show that $\hat{f}$ is a density for any realization of sample

Here K called **kernel function**, δ called **bandwidth**

Example:

- Two sample points x_1 and x_2
- $\delta = 1$, K attains max at zero

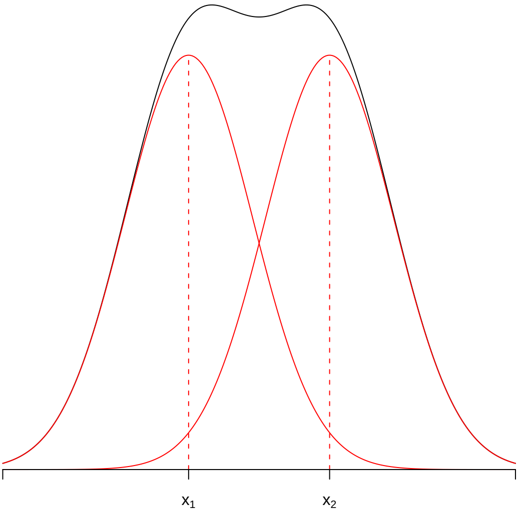
$$\hat{f}(s) := \frac{1}{N\delta} \sum_{n=1}^{N} K\left(\frac{s - x_n}{\delta}\right) = \frac{1}{2} K(s - x_1) + \frac{1}{2} K(s - x_2)$$

On each x_n we place a smooth "bump", which is

$$g_n(s) := \frac{1}{2} K(s - x_n)$$

Summing these two bumps gives $\hat{f} = g_1 + g_2$

On next slide, $\hat{f}$ in black, g_n in red


x_1
x_2

Let's compare this technique with the parametric one

Next two figures show estimates of f using default settings in R

Sample sizes are 200 and 2,000 respectively

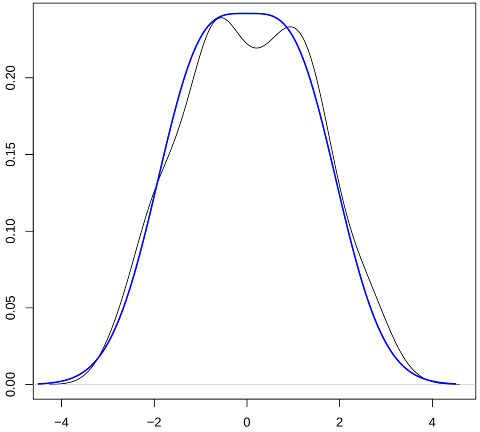


Figure: Nonparametric, sample size = 200

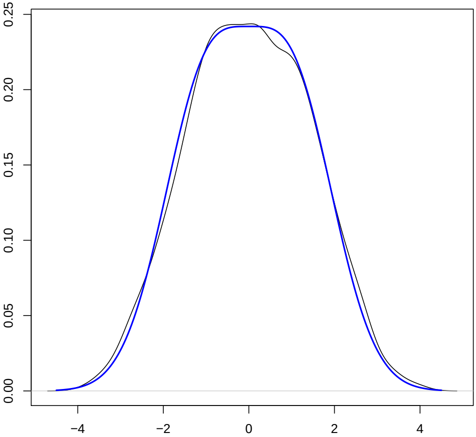


Figure: Nonparametric, sample size = 2,000

```r
set.seed(1234)
fden <- function(x) {  # The density function of f
   return(0.5 * dnorm(x, mean=-1)
               + 0.5 * dnorm(x, mean=1))
}
fsamp <- function(N) { # Generates N draws from f
   observations <- numeric(N)
   u <- runif(N)
   for (i in 1:N) {
       if (u[i] < 0.5) {
          observations[i] <- rnorm(1, mean=-1)
       }
       else observations[i] <- rnorm(1, mean=1)
   }
   return(observations)
}
observations <- fsamp(2000)
xgrid <- seq(-4.5, 4.5, length=200)
plot(density(observations), main="", xlab="", ylab="")
lines(xgrid, fden(xgrid), col="blue", lwd=2)
```

Comments:

With $N = 2000$, nonparametric fit better than parametric fit

Further increases in N continue to improve fit:

Theorem. Let f and K be densities on $\mathbb{R}$, and let $\{x_n\}_{n=1}^{\infty} \overset{\text{IID}}{\sim} f$. If $\delta_N \to 0$ and $N\delta_N \to \infty$ as $N \to \infty$, then

$$\mathbb{E} \left[\int |\hat{f}_N(s) - f(s)| \, ds \right] \to 0 \qquad \text{as } N \to \infty$$

In parametric example above, this was not true

On other hand, for small N, nonparametric estimate can be worse

Nonparametric methods generally need more data

Reason: Nonparametric methods have little structure imposed

Parametric methods: more structure, but structure might be wrong

General comments on parametric vs nonparametric estimation:

In some fields of science, underlying theory yields parametric class

(Example: Brownian motion)

In this case, parametric paradigm is excellent

Economics messier, rarely have such strong theory

Empirical Distributions

Let $x_1, \ldots, x_N \sim F$ where F unknown

Empirical distribution of the sample:

> Discrete distribution putting equal probability $1/N$ on each x_n

Below, x^e denotes RV with empirical distribution of $x_1, \ldots, x_N$

The cdf F_N of the empirical distribution called the **empirical cumulative distribution function**, or **ecdf**

Put differently

- $F_N :=$ cdf of x^e

To compute ecdf, recall that if x is a discrete RV taking values $s_1, \ldots, s_J$ with probabilities $p_1, \ldots, p_J$, then

$$F(s) = \mathbb{P}\{x \leq s\} = \sum_{j=1}^{J} \mathbb{1}\{s_j \leq s\} p_j$$

Thus, ecdf is

$$F_N(s) = \mathbb{P}\{x^e \leq s\} = \sum_{n=1}^{N} \mathbb{1}\{x_n \leq s\} \frac{1}{N} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{x_n \leq s\}$$

Alternatively

$$F_N(s) := \text{fraction of sample less than or equal to } s$$

Graphically, F_N is a step function

Upward jump of $1/N$ at each data point x_n.

Next slide: Example with $N = 10$ and $x_1, \ldots, x_N \overset{\text{IID}}{\sim}$ Beta(5, 5)

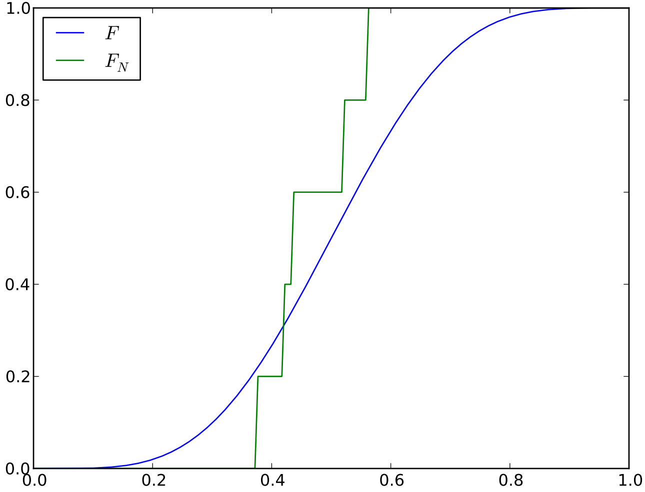


Figure: F_N and F with $N = 10$

Plug in Estimators

If $h\colon \mathbb{R} \to \mathbb{R}$, then expectation w.r.t. the ecdf is

$$\int h(s) F_N(ds) = \mathbb{E}\left[h(x^e)\right] = \sum_{n=1}^{N} h(x_n) \frac{1}{N} = \frac{1}{N} \sum_{n=1}^{N} h(x_n)$$

Example 1: Mean of ecdf is sample mean

$$\int s F_N(ds) = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Example 2: Second moment of ecdf is sample second moment

$$\int s^2 F_N(ds) = \frac{1}{N} \sum_{n=1}^{N} x_n^2$$

If sample IID, then for N large

$$\frac{1}{N}\sum_{n=1}^{N} h(x_n) \approx \mathbb{E}\left[h(x_n)\right]$$

Put differently,

$$\int h(s) F_N(ds) \approx \int h(s) F(ds)$$

Suggests approach for producing estimators:

To estimate $\theta = \int h(s)F(ds)$, replace cdf F with ecdf F_N:

$$\hat{\theta}_N := \int h(s) F_N(ds) = \frac{1}{N}\sum_{n=1}^{N} h(x_n)$$

Estimator $\hat{\theta}_N$ called the **plug in** estimator

Example: Plug in estimator of mean $\int sF(ds)$ is sample mean:

$$\int sF_N(ds) = \frac{1}{N}\sum_{n=1}^{N} x_n =: \bar{x}_N$$

Example: Plug in estimator of the variance

$$\mathbb{E}\left[(x - \mathbb{E}[x])^2\right] = \int \left[t - \int sF(ds)\right]^2 F(dt)$$

is

$$\int \left[t - \int sF_N(ds)\right]^2 F_N(dt) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x}_N)^2$$

(Approximately equal to sample variance s_N^2 when N large)

Properties of the ecdf

Let $x_1, \ldots, x_N \stackrel{\text{IID}}{\sim} F$, where F is any cdf

Let F_N be the corresponding ecdf

Fact: For all $s \in \mathbb{R}$, we have $F_N(s) \stackrel{p}{\to} F(s)$ as $N \to \infty$

Proof: By LLN,

$$F_N(s) := \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{x_n \leq s\} \stackrel{p}{\to} \mathbb{P}\{x_n \leq s\} =: F(s)$$

Fundamental Theorem of Statistics

In fact, stronger statement is true:

Theorem If $x_1, \ldots, x_N \overset{\text{IID}}{\sim} F$ and F_N is the corresponding ecdf, then

$$\sup_{s \in \mathbb{R}} |F_N(s) - F(s)| \xrightarrow{p} 0$$

(In fact, stronger result true—details omitted)

Next few figures illustrate

Each figure shows 10 observations of F_N

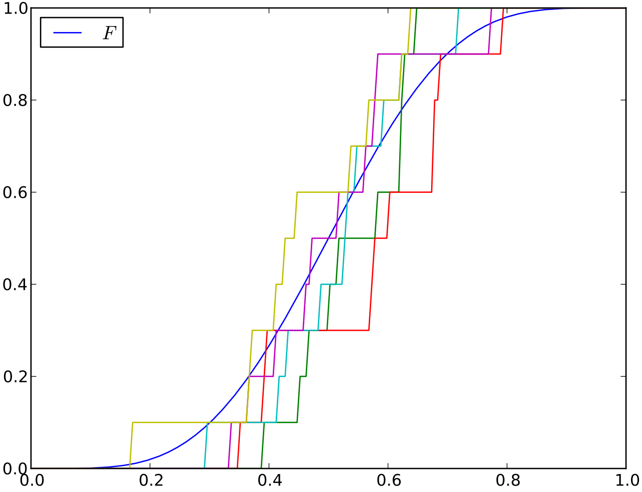


Figure: Realizations of F_N with $N = 10$

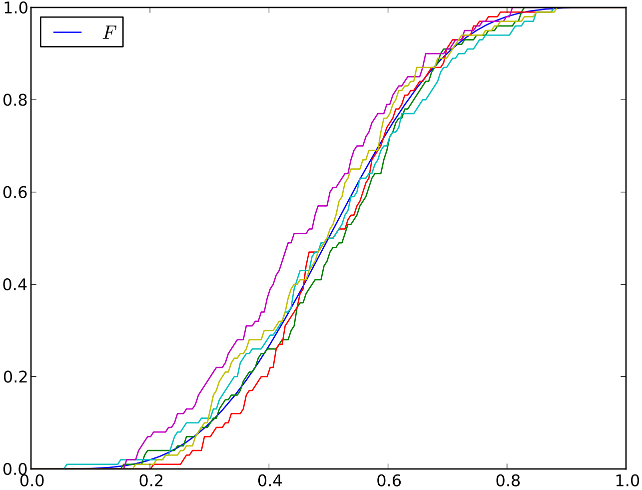


Figure: Realizations of F_N with $N = 100$

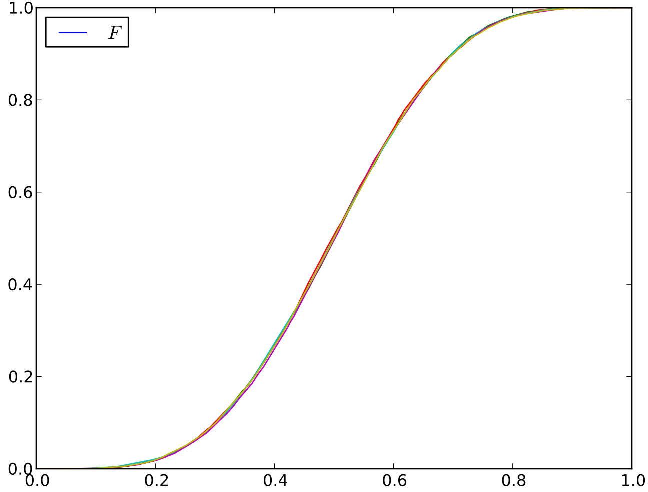


Figure: Realizations of F_N with $N = 1000$

The FTS tells us that in IID setting:

"If we have an infinite amount of data, then we can learn the underlying distribution without having to impose any assumptions"

However, bear in mind that we only ever have finite amount of data

Thus, assumptions still required to make inferences from data