Review of Statistics
Part A: Properties of Estimators

Thomas Wiemann
University of Chicago

Econometrics
Econ 21020

Updated: April 12, 2022

Recap

The review of probability theory introduced a formal language for characterizing uncertainty.

- ▷ Introduced random variables and their probability distributions;
- ▷ Developed concepts to describe features of random variables;
- ▷ Discussed restrictions on the joint distribution of random variables.

With our toolbox, we can return to the returns to education example.

- ▷ Under the random assignment assumption, we can show that

$$E_U[g(1, U) - g(0, U)|W = 1] = E_Y[Y|W = 1] - E_Y[Y|W = 0],$$

  where $E_Y[Y|W = 1]$ and $E_Y[Y|W = 0]$ are features of the joint distribution of the observables (Y, W).

Note that $E_Y[Y|W = 1]$ and $E_Y[Y|W = 0]$ are *theoretical* concepts.

- ▷ Statistics forms a bridge between random variables and data.

Outline

1. Estimators

2. Finite Sample Properties
   - ▷ Bias
   - ▷ Variance
   - ▷ The Bias-Variance Trade-off

3. Large Sample Properties
   - ▷ Consistency
   - ▷ Asymptotic Distribution

These notes benefit greatly from the exposition in Wasserman (2003) and the lecture notes of Prof. Max Tabord-Meehan.

Outline

1. **Estimators**

2. Finite Sample Properties
   - ▷ Bias
   - ▷ Variance
   - ▷ The Bias-Variance Trade-off

3. Large Sample Properties
   - ▷ Consistency
   - ▷ Asymptotic Distribution

These notes benefit greatly from the exposition in Wasserman (2003) and the lecture notes of Prof. Max Tabord-Meehan.

Random Sampling

Consider independent random variable $X_1, \ldots, X_n$ with $X_i \sim F_i, \forall i$.

- When $F_i = F, \forall i = 1, \ldots, n$, we say that $X_1, \ldots, X_n$ are *independent and identically distributed* (iid).
- To denote an iid sample of size n from F, we write

$$X_1, \ldots, X_n \overset{iid}{\sim} F. \tag{1}$$

Example 1

Consider $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$.

- If $X_1 \perp\!\!\!\perp X_2$, then X_1 and X_2 are independent.
- If $(\mu_1, \sigma_1^2) = (\mu_2, \sigma_2^2)$, then X_1 and X_2 are identically distributed.
- If $X_1 \perp\!\!\!\perp X_2$ *and* $(\mu_1, \sigma_1^2) = (\mu_2, \sigma_2^2)$, then X_1 and X_2 are iid.

Notation*: Instead of* (1)*, we also sometimes write* $X_1, \ldots, X_n \overset{iid}{\sim} X$*. So* X *may denote a random variable or its distribution.*

Estimators

Statistics is concerned with learning about the distribution from F using a sample $X_1, \ldots, X_n \sim F$.

- ▷ We will (for the most part), consider iid-samples.

Instead of fully characterizing F, the focus often lies on features of F.

- ▷ Features of interest are called *parameters*.
- ▷ For example, we may be interested in $\mu \equiv E[X]$ where $X \sim F$. Here, μ is the parameter of interest.

An *estimate* is a "guess" for the value of the parameter of interest.

- ▷ An *estimator* is a function of the sample whose value serves as a "guess" for a parameter of interest.
- ▷ For example, if $\mu \in \mathbb{R}$ and $\operatorname{supp} X_i = \mathbb{R}, \forall i$, then an estimator for μ is a function $\hat{\mu}_n(X_1, \ldots, X_n)$.
- ▷ Importantly: μ is a number but $\hat{\mu}_n$ is a random variable.

Notation*: Subscripts on expectation operators or distribution functions are omitted from now on whenever the context is clear.*

Example 2

Consider a sample $X_1, \ldots, X_n \overset{iid}{\sim} F$. An estimator for $F(x) = P(X \leq x)$ is given by

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}, \tag{2}$$

that is, the share of the sample below x is a “guess” for $P(X \leq x)$.

The estimator $\widehat{F}_n$ is called the *empirical CDF*.

The empirical CDF leads to a class of estimators that are know under the *sample analogue principle*.

- Suppose we are interested in a feature of F. The sample analogue principle suggests using the analogous feature of $\widehat{F}_n$ as an estimate.

Estimators (Contd.)

Example 3

Consider a sample $X_1, \ldots, X_n \overset{iid}{\sim} F$. Let $\mu = E[X]$ denote the parameter of interest. The sample analogue principle suggests the estimator

$$\hat{\mu}_n \equiv E_n[X] = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{3}$$

where E_n denotes the expectation with respect to the empirical CDF $\widehat{F}_n$.

Similarly, if the parameter of interest is $\sigma^2 = Var(X)$, the sample analogue principle suggests the estimator

$$\hat{\sigma}_n^2 \equiv \tag{4}$$

Estimators (Contd.)

The sample analogue principle is not the only approach to constructing estimators. Another frequently encountered class of estimators are extremum estimators, defined as the minimizers of a loss-functions.

Example 4

Consider a sample $X_1, \ldots, X_n \overset{iid}{\sim} F$ and let $\mu = E[X]$ denote the parameter of interest. Define an estimator

$$\hat{\mu}_n = \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (X_i - \mu)^2. \tag{5}$$

Taking first order conditions, we have

$$0 =$$

Estimators (Contd.)

For a given parameter, there infinitely many possible estimators.

Example 5

Consider a sample $X_1, \ldots, X_n \overset{iid}{\sim} F$ and let $\mu = E[X]$ denote the parameter of interest. Each of the following are estimators for μ :

- $\hat{\mu}_n^{(1)} = 0$;
- $\hat{\mu}_n^{(2)} = X_1$;
- $\hat{\mu}_n^{(3)} = \frac{1}{n} \sum_{i=1}^{n} X_i$.
- $\hat{\mu}_n^{(4)} = \frac{1}{n+\lambda} \sum_{i=1}^{n} X_i$ for some fixed $\lambda \in \mathbb{R}_+$.

Which one do you like best?

Statistics provides tools that allow for comparisons of estimators.

- Allows for selecting the "best" (or – at least – a "good enough") estimator.

Sampling Distribution

Recall that an estimator is a function of random variables and hence itself a random variable.

- The *sampling distribution* of an estimator is a name for its distribution.

Comparisons of estimators are analogous to comparisons of (features of) their sampling distribution.

- The sampling distribution often depends on the sample size n.

Consider an estimator $\hat{\theta}_n$ for some parameter θ of a distribution F.

- *Finite sample properties* describe features of the distribution of $\hat{\theta}_n$. These properties hold for any sample size $n \in \mathbb{N}$.
- *Large sample properties* describe features of the *asymptotic* distribution of $\hat{\theta}_n$. These properties hold approximately for large enough sample sizes n.

Outline

1. Estimators
2. **Finite Sample Properties**
   - **Bias**
   - Variance
   - The Bias-Variance Trade-off
3. Large Sample Properties
   - Consistency
   - Asymptotic Distribution

Bias

We begin with describing the expected deviations of the estimator from the true parameter.

Definition 1

The *bias* of an estimator $\hat{\theta}_n$ for θ is defined as

$$Bias(\hat{\theta}_n) = E\left[\hat{\theta}_n\right] - \theta. \tag{6}$$

The estimator is said to be

- *unbiased* if $Bias(\hat{\theta}_n) = 0$;
- *downwards biased* if $Bias(\hat{\theta}_n) < 0$;
- *upwards biased* if $Bias(\hat{\theta}_n) > 0$.

Bias (Contd.)

Example 6

Consider the estimators $\hat{\mu}_n^{(1)}, \hat{\mu}_n^{(2)}, \hat{\mu}_n^{(3)}$ and $\hat{\mu}_n^{(4)}$ of Example 5. We have

$$Bias(\hat{\mu}_n^{(1)}) =$$

$$Bias(\hat{\mu}_n^{(2)}) =$$

$$Bias(\hat{\mu}_n^{(3)}) =$$

$$Bias(\hat{\mu}_n^{(4)}) =$$

Note that the Bias of $\hat{\mu}_n^{(4)}$ depends on the unknown parameter μ.

Bias (Contd.)

Example 7

Consider the estimator $\hat{\sigma}_n^2$ defined in Example 3. We have

$$\hat{\sigma}_n^2 =$$

and

$$Bias(\hat{\sigma}_n^2) =$$

Can you construct an unbiased estimate for $Var(X)$?

Outline

1. Estimators
2. **Finite Sample Properties**
   - ▷ Bias
   - ▷ **Variance**
   - ▷ The Bias-Variance Trade-off
3. Large Sample Properties
   - ▷ Consistency
   - ▷ Asymptotic Distribution

Estimation Variance

Example 6 showed that very different estimators can have the same bias.

- ▷ Require other features of the sampling distribution to make comparison useful.

Another key property of an estimator is its variance:

$$Var\left(\hat{\theta}_n\right) = E\left[\left(\hat{\theta}_n - E[\hat{\theta}_n]\right)^2\right] \tag{7}$$

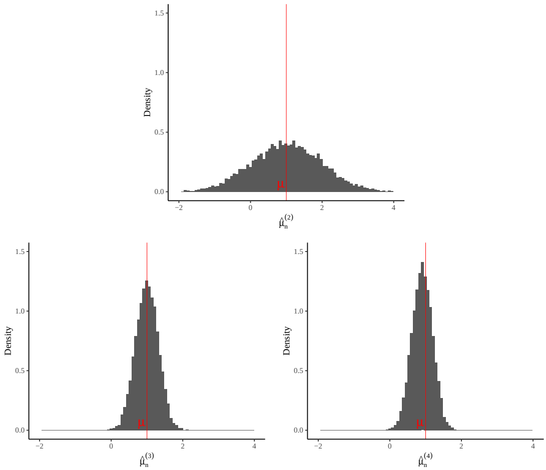
- ▷ Describes deviations from the expected value of the estimator.
- ▷ The expected value of a biased estimator is *not* the true parameter.

Figure 1 illustrates why considering both bias and variance is useful for distinguishing estimators.

- ▷ Draws from the sampling distribution of the estimators of Example 5.

Estimation Variance (Contd.)

Figure 1: Draws from Sampling Distributions of Estimators



Notes. Histograms of $\hat{\mu}_n^{(2)}$, $\hat{\mu}_n^{(3)}$ and $\hat{\mu}_n^{(4)}$ of Example 5 where $n = 10$ and $(\mu, \sigma^2) = (1, 1)$. For $\hat{\mu}_n^{(4)}$, I set $\lambda = 1$. You can find the corresponding code on GitHub: `lecture_plots.R`.

Estimation Variance (Contd.)

Example 8

Consider the estimators $\hat{\mu}_n^{(1)}, \hat{\mu}_n^{(2)}, \hat{\mu}_n^{(3)}$ and $\hat{\mu}_n^{(4)}$ of Example 5. We have

$$Var(\hat{\mu}_n^{(1)}) =$$

$$Var(\hat{\mu}_n^{(2)}) =$$

$$Var(\hat{\mu}_n^{(3)}) =$$

$$Var(\hat{\mu}_n^{(4)}) =$$

Note that the variances of $\hat{\mu}_n^{(2)}, \hat{\mu}_n^{(2)}$, and $\hat{\mu}_n^{(4)}$ depend on the unknown parameters (μ, σ^2).

Outline

1. Estimators

2. **Finite Sample Properties**
    - Bias
    - Variance
    - **The Bias-Variance Trade-off**

3. Large Sample Properties
    - Consistency
    - Asymptotic Distribution

The Bias-Variance Trade-Off

A popular criterion for evaluating estimators is the mean-squared error:

$$MSE\left(\hat{\theta}_n\right) = E\left[\left(\hat{\theta}_n - \theta\right)^2\right]. \tag{8}$$

- ▷ Describes the squared deviations of $\hat{\theta}_n$ from the true parameter.

The next result shows that the MSE is a one-number summary of the bias and variance of an estimator.

Corollary 1

Let $\hat{\theta}_n$ be an estimator for θ. We have

$$MSE\left(\hat{\theta}_n\right) = Bias\left(\hat{\theta}_n\right)^2 + Var\left(\hat{\theta}_n\right). \tag{9}$$

Proof. □

Example 9

Our analysis suggests that we may prefer $\hat{\mu}_n^{(3)}$ to $\hat{\mu}_n^{(2)}$.

- ▷ Both are unbiased but $Var(\hat{\mu}_n^{(2)}) > Var(\hat{\mu}_n^{(3)})$.

But Figure 1 also suggests that we may prefer $\hat{\mu}_n^{(4)}$ to $\hat{\mu}_n^{(2)}$ for small λ.

- ▷ Even though $Bias(\hat{\mu}_n^{(2)}) < Bias(\hat{\mu}_n^{(4)})$, we may find the difference in $Var(\hat{\mu}_n^{(2)})$ and $Var(\hat{\mu}_n^{(4)})$ sufficiently large to prefer the latter.

Calculations in R show that for the setting of Figure 1, we have:

- ▷ $MSE\left(\hat{\mu}_n^{(1)}\right) = 1.00$; $MSE\left(\hat{\mu}_n^{(2)}\right) \approx 0.97$;
- ▷ $MSE\left(\hat{\mu}_n^{(3)}\right) \approx 0.10$; $MSE\left(\hat{\mu}_n^{(4)}\right) \approx 0.09$.

Note: These are results for a *specific parameter values* (μ, σ^2).
Simulation are not mathematical proofs!

Outline

1. Estimators
2. Finite Sample Properties
   - Bias
   - Variance
   - The Bias-Variance Trade-off
3. **Large Sample Properties**
   - **Consistency**
   - Asymptotic Distribution

Large Sample Properties

Note that in Examples 5 and 8 depended on unknown parameters (μ, σ^2).

- ▷ $Bias(\hat{\mu}_n^{(4)})$ depends on μ;
- ▷ $Var(\hat{\mu}_n^{(2)})$ and $Var(\hat{\mu}_n^{(3)})$ depend on σ^2;
- ▷ $Var(\hat{\mu}_n^{(4)})$ depends on (μ, σ^2).

Without knowledge of the parameters that we want to estimate, we can't rank our estimators is terms of the MSE!

Instead of the (often) impossible question

- ▷ "Which estimator *is* best (or: 'good enough')?"

we instead attempt to answer the question

- ▷ "Which estimator *will eventually be* best? (or: 'good enough')"

Here, "eventually" considers gathering more and more observations.

Large Sample Properties (Contd.)

It turns out that we can make statements about the *eventual* characteristics of estimators in many settings *without* knowledge of the parameters of interest.

We rely heavily on two notions of convergence of random variables:

- Convergence in Probability;
- Convergence in Distribution.

Using these concepts, we study

- the consistency of an estimator, which checks whether it will eventually be arbitrarily "close" to the true parameter value;
- the asymptotic distribution of an estimator, which approximates its sampling distribution when n is large.

Convergence in Probability

Recall convergence in the context of sequences of real numbers:

- ▷ Consider $x, x_1, \ldots, x_n \in \mathbb{R}$. We write $x_n \to x$ if

$$\forall \varepsilon > 0, \, \exists N_\varepsilon \in \mathbb{N} : \quad |x_n - x| < \varepsilon, \quad \forall n \geq N_\varepsilon .$$

Convergence in probability generalizes this notion of convergence to sequences of random variables.

Definition 2 (Convergence in Probability)

Let $X_1, \ldots, X_n$ be a sequence of random variables, and let X be another random variable. We say X_n *converges in probability to* X if

$$\forall \varepsilon > 0, \quad P\left(|X_n - X| > \varepsilon\right) \to 0, \quad \text{as } n \to \infty. \tag{10}$$

We write $X_n \xrightarrow{p} X$.

In words: If $X_n \xrightarrow{p} X$, then X_n deviates from X by no more than ε with large probability as $n \to \infty$.

Consistency

We consider convergence in probability to analyze whether an estimator $\hat{\theta}_n$ for θ will eventually be arbitrarily close to the true parameter value.

Definition 3

We say an estimator $\hat{\theta}_n$ for a parameter θ is *consistent* if

$$\hat{\theta}_n \xrightarrow{p} \theta. \tag{11}$$

Consistency is often considered a minimum requirement for an estimator.

- ▷ If the estimator is not arbitrarily close to the true parameter even with infinitely many observations, then there is little hope that it will be reasonably close when the sample size n is finite.
- ▷ *No* inconsistent estimator is considered to be "good enough."

Note: *Equation* (11) *implicitly considered $n \to \infty$. Unless otherwise stated, we always consider $n \to \infty$ in this course.*

Consistency (Contd.)

Example 10

Consider the estimators $\hat{\mu}_n^{(1)}$ and $\hat{\mu}_n^{(2)}$ of Example 5. We have, $\forall \varepsilon > 0$,

$$P\left(|\hat{\mu}_n^{(1)} - \mu| > \varepsilon\right) =$$

$$P\left(|\hat{\mu}_n^{(2)} - \mu| > \varepsilon\right) =$$

Hence, neither $\hat{\mu}_n^{(1)}$ nor $\hat{\mu}_n^{(2)}$ are consistent estimators of μ.

- ▷ Since neither estimator meets the minimum requirement, we won't consider them any further.

Weak Law of Large Numbers

To show consistency of less trivial estimators, we need new technical tools. The most important is the Weak Law of Large Numbers:

Theorem 1 (Weak Law of Large Numbers; WLLN)

Let $X_1, \ldots, X_n \overset{iid}{\sim} X$ be a random sample. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E[X]. \tag{12}$$

In words: As $n \to \infty$, the sample average concentrates around its mean.

Example 11

Consider the estimator $\hat{\mu}_n^{(3)}$ of Example 5. By the WLLN,

$$\hat{\mu}_n^{(3)} \xrightarrow{p} \mu,$$

so that $\hat{\mu}_n^{(3)}$ is a consistent estimator of μ.

Weak Law of Large Numbers (Contd.)

To proof the WLLN, we make use of the following intermediate result:

Lemma 1 (Chebyshev's Inequality)

Let X be a random variable. Then,

$$\forall \varepsilon > 0, \quad P\left(|X| > \varepsilon\right) \leq \frac{E\left[X^2\right]}{\varepsilon^2}. \tag{13}$$

Proof. $\square$

Weak Law of Large Numbers (Contd.)

We now return to the proof of the WLLN.

Proof.

□

Weak Law of Large Numbers (Contd.)

Examples 10 and 11 discussed consistency of the estimators $\hat{\mu}_n^{(1)}$, $\hat{\mu}_n^{(2)}$, and $\hat{\mu}_n^{(3)}$ of Example 5. What about $\hat{\mu}_n^{(4)}$?

Note that

$$\hat{\mu}_n^{(4)} = \frac{1}{n+\lambda}\sum_{i=1}^{n} X_i = \frac{n}{n+\lambda}\frac{1}{n}\sum_{i=1}^{n} X_i, \tag{14}$$

so that $\hat{\mu}_n^{(4)}$ is a function of $\frac{1}{n}\sum_{i=1}^{n} X_i$ and $\frac{n}{n+\lambda}$.

The WLLN provides considers convergence in probability of the sample average. Now, we need tools to:

- derive convergence in probability of *random vectors*;
- derive convergence in probability of *functions* of random vectors.

Joint Convergence in Probability

Definition 4

Take $k \in \mathbb{N}$ and let $\tilde{X}_n = (X_{1,n}, \ldots, X_{k,n}), n \geq 1$, be a sequence of random vectors, and let $\tilde{X} = (X_1, \ldots, X_k)$ be another random vector. We say $\tilde{X}_n$ converges in probability to $\tilde{X}$ if

$$\forall \varepsilon > 0, \quad P\left(\sqrt{\sum_{j=1}^{k} (X_{j,n} - X_j)^2} > \varepsilon\right) \to 0, \quad \text{as } n \to \infty. \tag{15}$$

We won't require using Equation (15) directly due to the following result:

Theorem 2

Take $k \in \mathbb{N}$ and let $\tilde{X}_n = (X_{1,n}, \ldots, X_{k,n}), n \geq 1$, be a sequence of random vectors, and let $\tilde{X} = (X_1, \ldots, X_k)$ be another random vector. Then

$$X_{j,n} \xrightarrow{p} X_j, \forall j = 1, \ldots, k \quad \Rightarrow \quad \tilde{X}_n \xrightarrow{p} \tilde{X}. \tag{16}$$

Continuous Mapping Theorem

The following theorem delivers a powerful tool for proving convergence of any continuous functions of sample averages.

Theorem 3 (Continuous Mapping Theorem; CMT)

Let $X_n, n \geq 1$, be a sequence of random vectors, and let and X be another random vector. If $X_n \xrightarrow{p} X$, then

$$g\left(X_n\right) \xrightarrow{p} g\left(X\right), \tag{17}$$

for any function g that is continuous at $g\left(x\right), \forall x \in \operatorname{supp} X$.

Example 12

Let $A_n \xrightarrow{p} a \in \mathbb{R}$ and $B_n \xrightarrow{p} b \in \mathbb{R}$. Consider $g(a, b) = a/b$. Then

$$g\left(A_n, B_n\right) \xrightarrow{p} g(a, b), \tag{18}$$

by the CMT as long as $b \neq 0$.

Example 13

Consider $\hat{\mu}_n^{(4)}$ from Example 5. We show $\hat{\mu}_n^{(4)} \xrightarrow{p} \mu$ in four steps:

Example 14

Consider $\hat{\sigma}_n^2$ defined in Example 3. We show $\sqrt{\hat{\sigma}_n^2} \xrightarrow{p} \sigma$ in four steps:

Outline

1. Estimators
2. Finite Sample Properties
   - ▷ Bias
   - ▷ Variance
   - ▷ The Bias-Variance Trade-off
3. **Large Sample Properties**
   - ▷ Consistency
   - ▷ **Asymptotic Distribution**

Convergence in Distribution

Examples 11 and 13 showed that both $\hat{\mu}_n^{(3)}$ and $\hat{\mu}_n^{(4)}$ are consistent for θ.

- ▷ But: Consistency does not imply that the choice of estimator is irrelevant even for large n: Could have different variances.

We introduce the concept of convergence in distribution:

- ▷ Allows to assess dispersion of estimators as n grows large.
- ▷ Allows to make approximate probability statements about estimators.

Definition 5 (Convergence in Distribution)

Let $X_n, n \geq 1$, be a sequence of random variables, and let X be another random variable. We say X_n *converges in distribution to* X if

$$P\left(X_n \leq t\right) \to P\left(X \leq t\right), \quad \forall t \in \mathbb{R}. \tag{19}$$

We write $X_n \stackrel{d}{\to} X$.

In words: If $X_n \stackrel{d}{\to} X$, then the distribution of X_n is approximately equal to the distribution of X for large n.

Central Limit Theorem

The next result is a powerful tool for deriving the asymptotic distribution of sample averages.

Theorem 4 (Central Limit Theorem; CLT)

Let $X_1, \ldots, X_n \stackrel{iid}{\sim} X$ *be a random sample. Then*

$$\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)}{\sigma} \stackrel{d}{\to} N(0,1), \tag{20}$$

where $\mu \equiv E[X]$ *and* $\sigma \equiv sd(X) > 0$.

In words: As n grows large, the distribution of the sample average is approximately normal.

- ▷ Remarkable because we have *not* assumed that X is normal!

Notation*: We could have stated Equation* (20) *instead as* $\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)}{\sigma} \stackrel{d}{\to} Z$, *where* $Z \sim N(0,1)$. *As before, we may occasionally use random variables and their distributions interchangeably.*

Example 15

Consider $\hat{\mu}_n^{(3)}$ from Example 5. By the CLT, we have

$$\frac{\sqrt{n}\left(\hat{\mu}_n^{(3)} - \mu\right)}{\sigma} \xrightarrow{d} N(0,1). \tag{21}$$

Hence, for large n, we may approximate the distribution of $\hat{\mu}_n^{(3)}$ with

$$N\left(\mu,\ \sigma^2/n\right). \tag{22}$$

Note that (22) is of little practical help unless we may substitute parameter estimates for the unknown parameters.

Slutsky's Theorem

Good news: The result of the CLT continues to hold when parameter estimates are substituted for unknown parameter values.

Theorem 5 (Slutsky's Theorem)

Let $A_n, n \geq 1$, and $B_n, n \geq 1$, be sequences of random variables. Let A be another random variable and $b \in \mathbb{R}$. If $A_n \xrightarrow{d} A$ and $B_n \xrightarrow{p} b$, then

$$B_n + A_n \xrightarrow{d} b + A, \tag{23}$$

and

$$B_n A_n \xrightarrow{d} bA. \tag{24}$$

If in addition $b \neq 0$, then also

$$A_n / B_n \xrightarrow{d} A/b. \tag{25}$$

Slutsky's Theorem (Contd.)

Example 16

Consider $\hat{\sigma}_n^2$ and $\hat{\mu}_n^{(3)}$ from Example 3 and 5. Consider

$$Z_n \equiv \frac{\sqrt{n}\left(\hat{\mu}_n^{(3)} - \mu\right)}{\hat{\sigma}_n} = \frac{\sigma}{\hat{\sigma}_n}\frac{\sqrt{n}\left(\hat{\mu}_n^{(3)} - \mu\right)}{\sigma},$$

so that Slutsky's suggests taking $A_n \equiv \frac{\sqrt{n}\left(\hat{\mu}_n^{(3)} - \mu\right)}{\sigma}$ and $B_n \equiv \frac{\sigma}{\hat{\sigma}_n}$. Then,

Example 17

Consider $\hat{\sigma}_n^2$ and $\hat{\mu}_n^{(4)}$ from Example 3 and 5. We want to show that

$$\frac{\sqrt{n}\left(\hat{\mu}_n^{(4)} - \mu\right)}{\hat{\sigma}_n} \xrightarrow{d} N(0,1).$$

We have

Standard Errors

Examples 16 and 17 show that approximate probabilistic statements about estimators can be made using their asymptotic distribution. For this purpose, practitioners often use so-called *standard errors*.

Definition 6 (Standard Error)

Let $\hat{\theta}_n$ and $\hat{\sigma}_n$ be estimators such that

$$\frac{\sqrt{n}\left(\hat{\theta}_n - \theta\right)}{\hat{\sigma}_n} \xrightarrow{d} N(0,1). \tag{26}$$

The *standard error* of $\hat{\theta}_n$ is defined as

$$se\left(\hat{\theta}_n\right) = \frac{\hat{\sigma}_n}{\sqrt{n}}. \tag{27}$$

Standard errors are an approximation to the standard deviation of an estimator based on its asymptotic distribution.

Bivariate Central Limit Theorem

Slutsky's Theorem considered the joint convergence of sequences of random variables when one of the sequences converges to a constant.

- ▷ Need tools to understand joint convergence when *both* sequences converge to a random variable. Fortunately, we have the next result:

Theorem 6 (Bivariate Central Limit Theorem)

Let $\tilde{X}_1, \ldots, \tilde{X}_n \overset{iid}{\sim} Y$ *be a sample or bivariate random vectors where* $\tilde{X}_i = (X_{1,i}, X_{2,i})$ *and* $\tilde{X} = (X_1, X_2)$. *Then*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i - \mu\right) \xrightarrow{d} N(0, \Sigma), \tag{28}$$

where $\mu \equiv E[\tilde{X}]$ *and*

$$\Sigma \equiv Var(\tilde{X}) = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{bmatrix}. \tag{29}$$

Bivariate Central Limit Theorem (Contd.)

Example 18

Consider a sample $(Y_1, X_1), \ldots, (Y_1, X_1) \overset{iid}{\sim} (Y, X)$ where $X \sim \mathrm{Bernoulli}(p)$ with unknown $p \in (0, 1)$. Suppose we are interested in the joint distribution of the estimators

$$E_n[YX] = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i, \quad \text{and} \quad E_n[Y(1-X)] = \frac{1}{n}\sum_{i=1}^{n} Y_i(1-X_i). \tag{30}$$

By the (bivariate) CLT, we have

Bivariate Slutsky's Theorem

As was the case with the univariate CLT, it's bivariate analogue is particularly useful when combined with a Slutsky-type result:

Theorem 7 (Bivariate Slutsky's Theorem)

Let $A_n, n \geq 1$, and $B_n, n \geq 1$, be sequences of bivariate random vectors variables. Let A be another bivariate random vector and $b \in \mathbb{R}^2$. If $A_n \xrightarrow{d} A$ and $B_n \xrightarrow{p} b$, then

$$A_n + B_n \xrightarrow{d} A + b, \tag{31}$$

and

$$B_n^\top A_n \xrightarrow{d} b^\top A. \tag{32}$$

Bivariate Slutsky's Theorem (Contd.)

Example 19

Let $A_n, n \geq 1$ and $B_n, n \geq 1$ be sequences of bivariate random vectors such that $A_n \stackrel{d}{\to} N(0, \Sigma)$ and $B_n \stackrel{p}{\to} b \in \mathbb{R}^2$. By Slutsky's Theorem,

$$B_n^\top A_n \stackrel{d}{\to} b^\top N(0, \Sigma) \stackrel{d}{=} N(0, b^\top \Sigma b),$$

where the last equation follows from Lemma 4c of Lecture 2A.

Suppose now that $Z_n, n \geq 1$, such that $Z_n \stackrel{d}{\to} N(0, I_2)$, and $\hat{\Sigma}_n, n \geq 1$ is a sequence of estimators such that $\hat{\Sigma}_n^{-1}$ exists and $\hat{\Sigma}_n \stackrel{p}{\to} \Sigma$. By the CMT,

whenever Σ^{-1} exists. Hence, by Slutsky's Theorem,

Bivariate Slutsky's Theorem (Contd.)

Example 20

Consider the setting of Example 18 and construct the estimator

$$E_n[YX] - E_n[Y(1-X)] = \begin{bmatrix} 1 \\ -1 \end{bmatrix}^\top \begin{bmatrix} E_n[YX] \\ E_n[Y(1-X)] \end{bmatrix}. \tag{33}$$

Hence, it follows from Example 18 and Slutsky's Theorem that

Summary

This concludes the first part of our statistics review.

- ▷ Introduced the sample analogue principle to develop estimators;
- ▷ Discussed finite sample properties of estimators, in particular, their bias, variance, and MSE;
- ▷ Generalized the concept of convergence to random variables via convergence in probability and convergence in distribution;
- ▷ Studied large sample properties of estimators, in particular, their consistency and asymptotic distribution.

A key insight was that under fairly general conditions, approximate probabilistic statements about estimators can be made using their asymptotic distribution.

- ▷ In the second part of our review, we focus on statements of the form: "If the true parameter were to be $\theta \in \Theta_0$, what is the (approximate) probability our estimator would take its realized value?"
- ▷ This is known as *statistical hypothesis testing*.

References

Wasserman, L. (2003). *All of statistics*. Springer.