Review C: Properties of Estimators

Professor Ji-Woong Chung
Korea University

This lecture note is based on Thomas Wiemann's.

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, let's return to the returns to education example.

$$E[Y_i(1) - Y_i(0) \mid D_i = 1] = E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0],$$

where $E[Y_i \mid D_i = 1]$ and $E[Y_i \mid D_i = 0]$ are features of the joint distribution of the observables (Y, D).

Note that $E[Y_i \mid D_i = 1]$ and $E[Y_i \mid D_i = 0]$ are theoretical concepts.

- Statistics forms a bridge between random variables and data.

Outline

Outline

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

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

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

Notation: Instead of the iid notation, 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 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 is often on features of F.

- Features of interest are called estimands or 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 μ is the parameter and $X_1, \ldots, X_n$ is the sample, 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.

Estimators (Contd.)

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

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

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

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

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

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

Estimators (Contd.)

Example 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

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

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

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

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2.$$

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 loss functions.

Example 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

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

Taking first-order conditions, we have

$$0 = \frac{\partial}{\partial \mu} \sum_{i=1}^{n} (X_i - \mu)^2 = -2 \sum_{i=1}^{n} (X_i - \mu).$$

Solving for μ, we get

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

which is the same as the sample mean derived earlier.

Estimators (Contd.)

For a given parameter, there are infinitely many possible estimators.

Example 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 $\lambda > 0$.

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

Outline

Bias

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

Definition (Bias)

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

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

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.)

$$\hat{\mu}_n^{(1)} = 0; \quad \hat{\mu}_n^{(2)} = X_1; \quad \hat{\mu}_n^{(3)} = \frac{1}{n}\sum_{i=1}^n X_i; \quad \hat{\mu}_n^{(4)} = \frac{1}{n+\lambda}\sum_{i=1}^n X_i$$

Example Consider the estimators $\hat{\mu}_n^{(1)}$, $\hat{\mu}_n^{(2)}$, $\hat{\mu}_n^{(3)}$, and $\hat{\mu}_n^{(4)}$ from the previous example. We have:

$$\begin{aligned}
Bias(\hat{\mu}_n^{(1)}) &= E[0] - \mu = -\mu, \\
Bias(\hat{\mu}_n^{(2)}) &= E[X_1] - \mu = 0, \\
Bias(\hat{\mu}_n^{(3)}) &= E\left[\frac{1}{n}\sum_{i=1}^n X_i\right] - \mu = \frac{1}{n}E\left[\sum_{i=1}^n X_i\right] - \mu = 0, \\
Bias(\hat{\mu}_n^{(4)}) &= E\left[\frac{1}{n+\lambda}\sum_{i=1}^n X_i\right] - \mu = -\frac{\lambda}{n+\lambda}\mu.
\end{aligned}$$

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

Bias (Contd.)

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

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 = \ldots = \left(\frac{1}{n} - \frac{1}{n^2}\right)\sum_{i=1}^{n} X_i^2 - \frac{1}{n^2}\sum_{i=1}\sum_{j\neq i} X_i X_j,$$

and

$$Bias(\hat{\sigma}_n^2) = -\frac{1}{n}Var(X).$$

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

Outline

Estimation Variance

The previous example 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:

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

- The square root of this is call the "standard error"
- 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 the earlier example.

Estimation Variance (Contd.)

$$\hat{\mu}_n^{(1)} = 0; \quad \hat{\mu}_n^{(2)} = X_1; \quad \hat{\mu}_n^{(3)} = \frac{1}{n}\sum_{i=1}^{n} X_i; \quad \hat{\mu}_n^{(4)} = \frac{1}{n+\lambda}\sum_{i=1}^{n} X_i$$

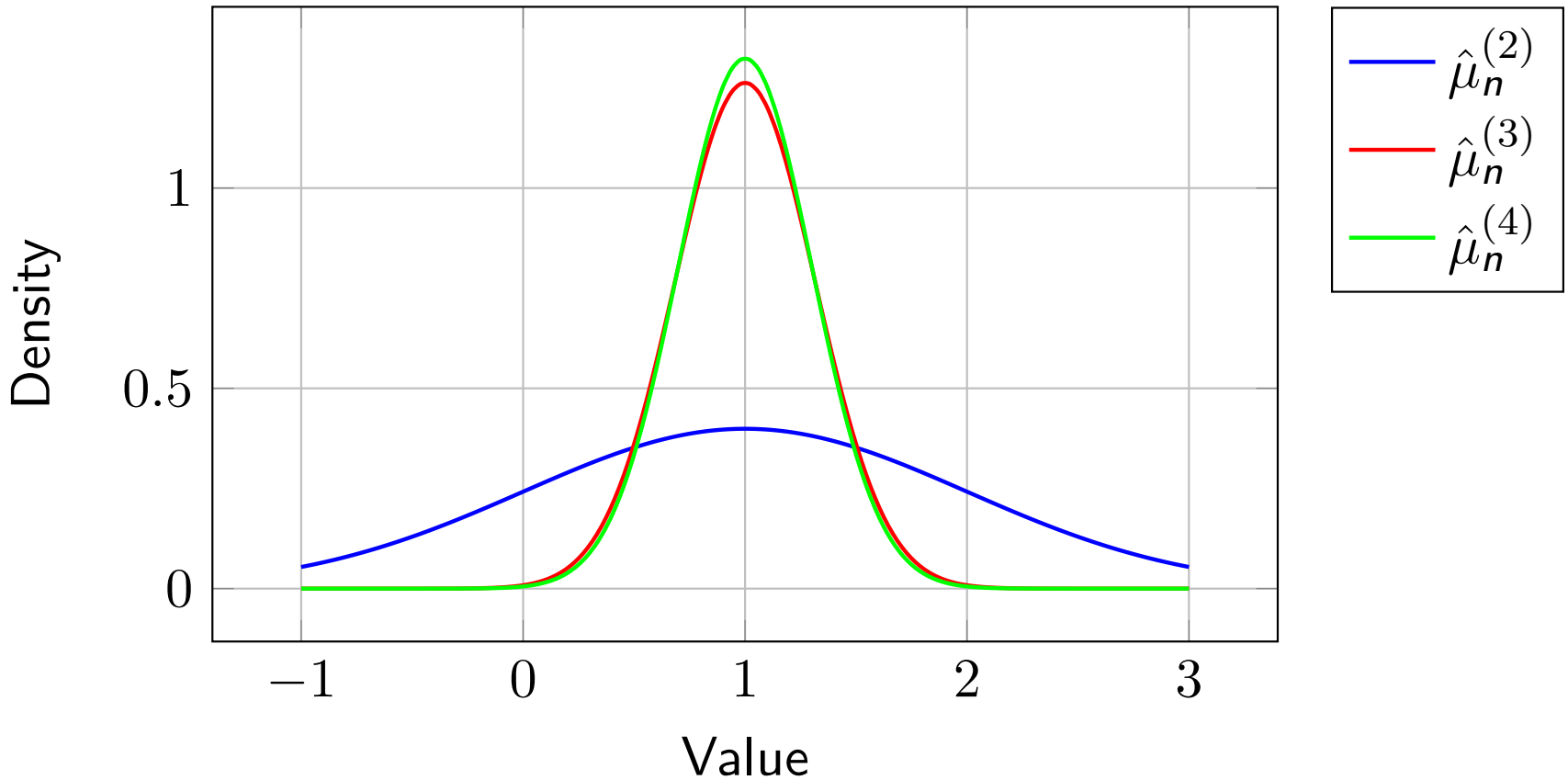

Figure 1: Draws from Sampling Distributions of Estimators
Notes: Histograms of $\hat{\mu}_n^{(2)}$, $\hat{\mu}_n^{(3)}$, and $\hat{\mu}_n^{(4)}$ where $n = 10$ and $(\mu, \sigma^2) = (1, 1)$. For $\hat{\mu}_n^{(4)}$, $\lambda = 1$.

Estimation Variance (Contd.)

$$\hat{\mu}_n^{(1)} = 0; \quad \hat{\mu}_n^{(2)} = X_1; \quad \hat{\mu}_n^{(3)} = \frac{1}{n}\sum_{i=1}^{n} X_i; \quad \hat{\mu}_n^{(4)} = \frac{1}{n+\lambda}\sum_{i=1}^{n} X_i$$

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

$$\begin{aligned}
Var(\hat{\mu}_n^{(1)}) &= 0, \\
Var(\hat{\mu}_n^{(2)}) &= \sigma^2, \\
Var(\hat{\mu}_n^{(3)}) &= \frac{\sigma^2}{n}, \\
Var(\hat{\mu}_n^{(4)}) &= \frac{\sigma^2}{(n+\lambda)^2}.
\end{aligned}$$

Note that the variances of $\hat{\mu}_n^{(2)}$, $\hat{\mu}_n^{(3)}$, and $\hat{\mu}_n^{(4)}$ depend on the unknown parameters (μ, σ^2).

Outline

Mean Squared Error

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

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

- Describes the squared deviations of $\hat{\theta}_n$ from the true parameter.
- It measure "how bad" an estimator is.

Mean Squared Error (Contd.)

Example: Consider two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of the same unknown parameter $\theta = 0$.

The distribution of $\hat{\theta}_1$: $P(\hat{\theta}_1 = -100) = P(\hat{\theta}_1 = 100) = 0.5$

The distribution of $\hat{\theta}_2$: $P(\hat{\theta}_2 = 1) = 1$

The bias:

$$E[\hat{\theta}_1] - \theta = (0.5)(-100) + (0.5)(100) - 0 = 0$$
$$E[\hat{\theta}_2] - \theta = (1)(1) - 0 = 1.$$

The MSE:

$$MSE(\hat{\theta}_1) = (0.5)(-100 - 0)^2 + (0.5)(100 - 0)^2 = 10,000$$
$$MSE(\hat{\theta}_2) = (1)(1 - 0)^2 = 1$$

Mean Squared Error (Contd.)

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

Corollary

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

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

Proof. $E[(\hat{\theta}_n - \theta)^2] = Var(\hat{\theta}_n - \theta) + [E(\hat{\theta}_n - \theta)]^2 = Var(\hat{\theta}_n) + Bias(\hat{\theta})^2$

The Bias-Variance Trade-Off

Example Consider the two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$.

$$
\begin{aligned}
Var(\hat{\theta}_1) &= E[(\hat{\theta}_1 - E(\hat{\theta}))^2] = E[(\hat{\theta}_1 - 0)^2] \\
&= (0.5)(-100)^2 + (0.5)(100)^2 = 10,000 \\
Bias(\hat{\theta}_1) &= E[\hat{\theta}_1] - \theta = 0 \\
\rightarrow MSE(\hat{\theta}_1) &= 10,000 + 0 = 10,000
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{\theta}_2) &= E[(\hat{\theta}_2 - E(\hat{\theta}))^2] = E[(\hat{\theta}_1 - 1)^2] \\
&= (1)(1-1)^2 = 0 \\
Bias(\hat{\theta}_2) &= E[\hat{\theta}_2] - \theta = 1 \\
\rightarrow MSE(\hat{\theta}_2) &= 0 + 1 = 1
\end{aligned}
$$

The Bias-Variance Trade-Off

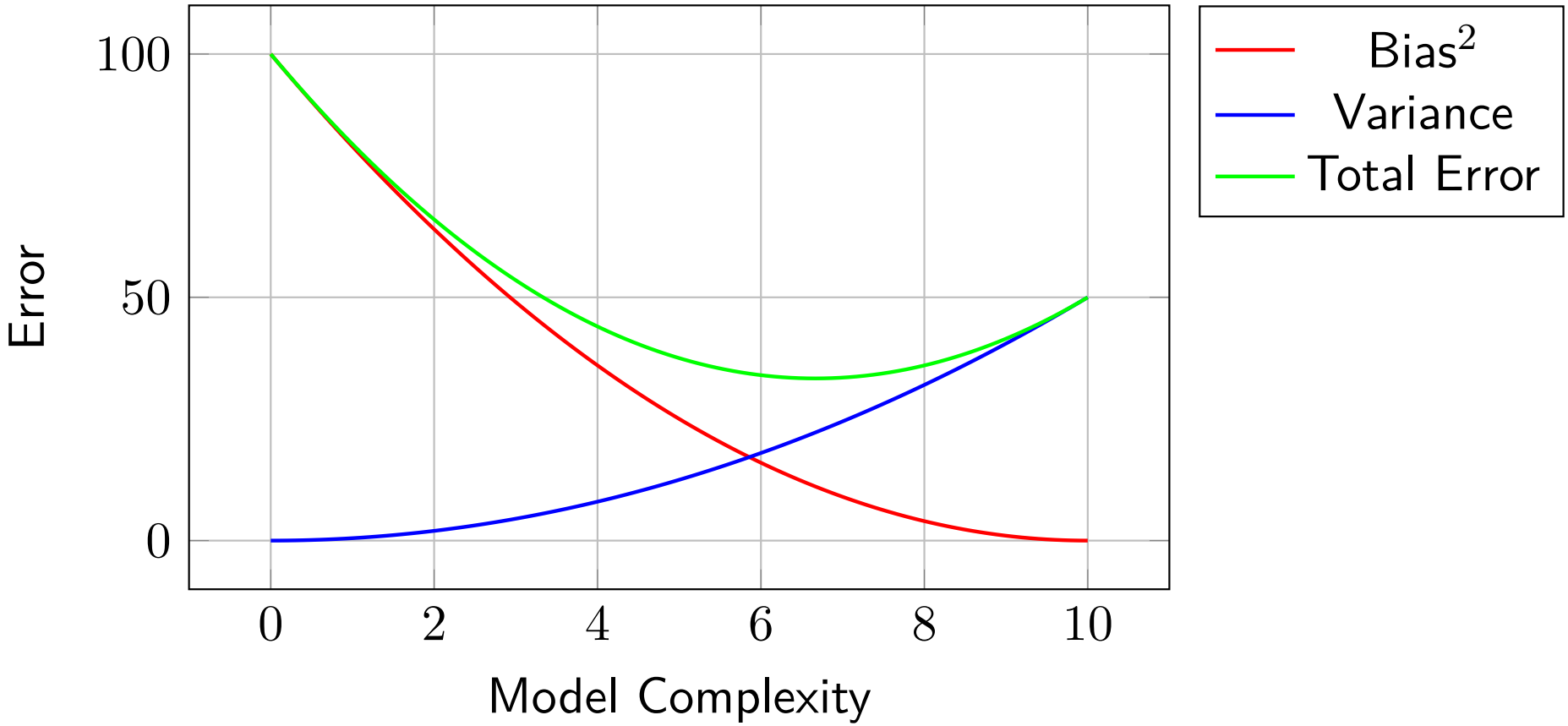

Figure: Bias-Variance Trade-off

Outline

Outline

Large Sample Properties

In the previous examples, the bias and variance 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 in 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

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

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

Definition (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

$$\forall \epsilon > 0, P(|X_n - X| > \epsilon) \to 0 \text{ as } n \to \infty.$$

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 (Consistency)

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

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

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 $\hat{\theta}_n \xrightarrow{p} \theta$ implicitly considers $n \to \infty$. Unless otherwise stated, we always consider $n \to \infty$ in this course.

Consistency (Contd.)

$$\hat{\mu}_n^{(1)} = 0; \quad \hat{\mu}_n^{(2)} = X_1; \quad \hat{\mu}_n^{(3)} = \frac{1}{n}\sum_{i=1}^n X_i; \quad \hat{\mu}_n^{(4)} = \frac{1}{n+\lambda}\sum_{i=1}^n X_i$$

Example Consider the estimators $\hat{\mu}_n^{(1)}$ and $\hat{\mu}_n^{(2)}$. We have, for all $\epsilon > 0$,

$$P(|\hat{\mu}_n^{(1)} - \mu| > \epsilon) = P(|0 - \mu| > \epsilon) \nrightarrow 0$$
$$P(|\hat{\mu}_n^{(2)} - \mu| > \epsilon) = P(|X_1 - \mu| > \epsilon) \nrightarrow 0$$

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 (WLLN):

Theorem (Weak Law of Large Numbers (WLLN))

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

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

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

Example Consider the estimator $\hat{\mu}_n^{(3)}$. 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.)

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

Note that

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

so that $\hat{\mu}_n^{(4)}$ is a function of $\frac{1}{n}\sum_{i=1}^{n} X_i$ and $\frac{n}{n+\lambda}$. What's next? Bear with me...

The WLLN provides 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 (Joint Convergence in Probability)

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

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

We won't use this directly due to the following result:

Theorem

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

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

Continuous Mapping Theorem

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

Theorem (Continuous Mapping Theorem (CMT))

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

$$g(X_n) \xrightarrow{p} g(X),$$

for any function g that is continuous at $g(x)$, $\forall x \in$ supp X.

Example 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

$$g(A_n, B_n) \xrightarrow{p} g(a, b),$$

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

Continuous Mapping Theorem (Contd.)

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

1. Define $A_n = \frac{n}{n+\lambda}$ and $B_n = \frac{1}{n}\sum X_i$
2. Define $g(a, b) = a \cdot b$. So, $g(A_n, B_n) = \hat{\mu}_n^{(4)}$
3. By the WLLN, $B_n \xrightarrow{p} E(X) = \mu$
4. $A_n \to 1$
5. By the CMT

$$g(A_n, B_n) \xrightarrow{p} \mu \cdot 1 = \mu.$$

Continuous Mapping Theorem (Contd.)

Example: Consider $\hat{\sigma}_n^2$. We show

$$\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2$$

in four steps:

Recall $\hat{\sigma}_n^2 = \frac{1}{n}\sum X_i^2 - (\frac{1}{n}\sum X_i)^2$

1. Define $A_n = \frac{1}{n}\sum X_i^2$ and $B_n = \frac{1}{n}\sum X_i$
2. Define $g(a, b) = a - b^2$. So, $g(A_n, B_n) = \hat{\sigma}_n^2$
3. By the WLLN, $A_n \xrightarrow{p} E(X^2)$ and $B_n \xrightarrow{p} E(X) = \mu$
4. By the CMT

$$g(A_n, B_n) \xrightarrow{p} E(X^2) - E(X) = \sigma$$

Outline

Convergence in Distribution

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

- But, different estimators could have different variances.

We introduce the concept of convergence in distribution:

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

Definition (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

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

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

In words: If $X_n \xrightarrow{d} X$, 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 (Central Limit Theorem (CLT))

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

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

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!

Central Limit Theorem (Contd.)

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

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

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

$$N\left(0, \sigma^2\right).$$

i.e., $\hat{\mu}_n^{(3)} \xrightarrow{d} N(\mu, \sigma^2/n)$

Note that this approximation is of little practical help unless we may substitute parameter estimates for the unknown parameters.

Delta Method

If Y_n has a limiting Normal distribution, then the delta method allows us to find the limiting distribution of $g(Y_n)$ where g is any smooth function.

Theorem (The Delta Method)

Suppose that

$$\sqrt{n}\frac{Y_n - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

and that g is a differentiable function such that $g'(\mu) \neq 0$. Then

$$\sqrt{n}\frac{g(Y_n) - g(\mu)}{\mid g'(\mu) \mid \sigma} \xrightarrow{d} N(0,1)$$

Delta Method (Contd.)

Example Given $\sqrt{n}\left(\frac{\hat{\mu}_n-\mu}{\sigma}\right) \xrightarrow{d} N(0,1)$, the asymptotic limit of $\sqrt{n}(\hat{\mu}_n^2 - \mu^2)$ is

By the Delta method, since $g(\mu) = \mu^2, g'(\mu) = 2\mu$,

$$\sqrt{n}(\hat{\mu}_n^2 - \mu^2) \xrightarrow{d} N(0, \sigma^2(2\mu)^2)$$

Slutsky's Theorem

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

Theorem (Slutsky's Theorem)

Let A_n, and B_n, 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

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

and

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

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

$$\frac{A_n}{B_n} \xrightarrow{d} \frac{A}{b}.$$

Slutsky's Theorem (Contd.)

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

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

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

By CLT, $A_n \xrightarrow{d} N(0,1)$

By WLLN & CMT, $B_n \xrightarrow{p} 1, \forall \sigma > 0$

By Slutsky's, $B_n A_n \xrightarrow{d} 1 \cdot N(0,1)$

Slutsky's Theorem (Contd.)

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

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

We have

$$\sqrt{n}\frac{\frac{n}{n+\lambda}\frac{1}{n}\sum_{i=1}^n X_i - \mu}{\hat{\sigma}_n} = \sqrt{n}\frac{\frac{n}{n+\lambda}\frac{1}{n}\sum_{i=1}^n X_i - \frac{n}{n+\lambda}\mu + \frac{n}{n+\lambda}\mu - \mu}{\hat{\sigma}_n}$$

$$= \underbrace{\frac{n}{n+\lambda}}_{\to 1}\underbrace{\sqrt{n}\frac{\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)}{\hat{\sigma}_n}}_{\xrightarrow{d} N(0,1)} + \underbrace{\sqrt{n}\frac{\mu\left(\frac{n}{n+\lambda} - 1\right)}{\hat{\sigma}_n}}_{\frac{\mu\left(\frac{-\sqrt{n}\lambda}{n+\lambda}\right)}{\hat{\sigma}_n} \to 0}$$

By Slutsky's, done.

Standard Errors

Informally, these examples show that the sampling distribution of the estimators can be approximated with $N(\mu, \sigma^2/n)$. For this purpose, practitioners often use so-called standard errors.

Definition (Standard Error)

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

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

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

$$\mathrm{se}(\hat{\theta}_n) = \frac{\hat{\sigma}_n}{\sqrt{n}}.$$

For large n, we may approximate the sampling distribution of an estimator $\hat{\theta}_n$ for θ with $\sqrt{n}$-normal asymptotic distribution by $N(\theta, \mathrm{se}(\hat{\theta}_n)^2)$.

Confidence Intervals

Researchers often construct asymptotic confidence intervals to succinctly characterize the approximate sampling distribution:

Theorem

Let $\hat{\theta}_n$ be an estimator for θ earlier. For $\alpha \in (0,1)$, consider

$$C_n = \left[\hat{\theta}_n - z_{1-\alpha/2} \cdot se(\hat{\theta}_n), \quad \hat{\theta}_n + z_{1-\alpha/2} \cdot se(\hat{\theta}_n)\right],$$

$$C_n^+ = \left[\hat{\theta}_n - z_{1-\alpha} \cdot se(\hat{\theta}_n), \infty\right),$$

$$C_n^- = \left(-\infty, \hat{\theta}_n + z_{1-\alpha} \cdot se(\hat{\theta}_n)\right],$$

where $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the $1-\alpha$ quantile of a standard normal. C_n, C_n^+, and C_n^- are asymptotically valid $1-\alpha$ confidence intervals. I.e.,

$$P(\theta \in \tilde{C}_n) \to 1-\alpha, \text{ for } \tilde{C}_n = C_n, C_n^+, C_n^-.$$

Note C_n is random (a function of sample), θ is fixed.

Confidence Intervals (Contd.)

Proof. We prove the theorem only for C_n.

$$
\begin{aligned}
P(\theta \in C_n) &= P(\hat{\theta}_n - z_{1-\alpha/2} \cdot \mathsf{se}(\hat{\theta}_n) < \theta < \hat{\theta}_n + z_{1-\alpha/2} \cdot \mathsf{se}(\hat{\theta}_n) \\
&= P(-z_{1-\alpha/2} \leq \frac{\hat{\theta}_n \theta}{se(\hat{\theta}_n)} \leq z_{1-\alpha/2}) \\
&= P\left(\frac{\hat{\theta}_n \theta}{se(\hat{\theta}_n)} \leq z_{1-\alpha/2}) - P(\frac{\hat{\theta}_n \theta}{se(\hat{\theta}_n)} \leq< -z_{1-\alpha/2}\right) \\
&\to P(Z \leq z_{1-\alpha/2}) - P(Z \leq -z_{1-\alpha/2}) \\
&= \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2}) \\
&= \Phi(z_{1-\alpha/2}) - (1 - \Phi(z_{1-\alpha/2})) \\
&= (1-\alpha/2) - (1 - 1 - \alpha/2) = 1 - \alpha \because 1 - \alpha/2 \equiv \Phi^{-1}(1-\alpha/2)
\end{aligned}
$$

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.

We need tools to understand joint convergence when both sequences converge to a random variable. Fortunately, we have:

Theorem (Bivariate Central Limit Theorem)

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

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

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

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

Bivariate Central Limit Theorem (Contd.)

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

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

By the (bivariate) CLT, we have

$$\sqrt{n}\left(\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} Y_i X_i \\ \frac{1}{n}\sum_{i=1}^{n} Y_i(1-X_i) \end{bmatrix} - \begin{bmatrix} E[YX] \\ E[Y(1-X)] \end{bmatrix}\right) \xrightarrow{d} N(0, \Sigma),$$

$$\text{where } \Sigma = \begin{bmatrix} Var(YX) & Cov(YX, Y(1-X)) \\ Cov(YX, Y(1-X)) & Var(Y(1-X)) \end{bmatrix}$$

Bivariate Slutsky's Theorem

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

Theorem (Bivariate Slutsky's Theorem)

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

$$A_n \xrightarrow{d} A \quad \text{and} \quad B_n \xrightarrow{p} b,$$

then

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

and

$$B_n^T A_n \xrightarrow{d} b^T A.$$

Bivariate Slutsky's Theorem (Contd.)

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

$$B_n^T A_n \xrightarrow{d} b^T N(0, \Sigma) \equiv N(0, b^T \Sigma b),$$

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

$$B_n^T \hat{\Sigma}_n^{-1/2} \xrightarrow{p} b^T \Sigma^{-1/2}$$

wherever Σ^{-1} exist. By Slutsky's,

$$B_n^T \hat{\Sigma}_n^{-1/2} Z_n \xrightarrow{d} b^T \Sigma^{-1/2} N(0, I_2) \stackrel{d}{=} N(0, b^T \Sigma^{-1/2} \Sigma^{-1/2} b)$$

Bivariate Slutsky's Theorem (Contd.)

Example Construct the estimator

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

Hence, by Slutsky's Theorem that

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix}^T \sqrt{n} \left(\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Y_i X_i \\ \frac{1}{n}\sum_{i=1}^n Y_i(1-X_i) \end{bmatrix} - \begin{bmatrix} E[YX] \\ E[Y(1-X)] \end{bmatrix} \right)$$
$$\xrightarrow{d} N\left(0, \begin{bmatrix} 1 \\ -1 \end{bmatrix}^T \Sigma \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right),$$

where $\Sigma = \begin{bmatrix} Var(YX) & Cov(YX, Y(1-X)) \\ Cov(YX, Y(1-X)) & Var(Y(1-X)) \end{bmatrix}$

Outline

On the Interpretation of Estimates

Thus far, we have exclusively discussed estimators $\hat{\theta}_n$ for a parameter θ.

- $\hat{\theta}_n$ is a function of the sample $X_1, \ldots, X_n \sim X$, so random.
- How does real-world data come in?

Data is a realization of our sample $X_1, \ldots, X_n$.

- The data is the collection of numbers: $x_1, \ldots, x_n$.

An *estimate* is a realization of our estimator $\hat{\theta}_n$:

- The estimator $\hat{\theta}_n(X_1, \ldots, X_n)$ is a random variable;
- The estimate $\hat{\theta}_n(x_1, \ldots, x_n)$ is a number.

This distinction between estimators and estimates can be confusing.

- We can make probabilistic statements about $\hat{\theta}_n(X_1, \ldots, X_n)$.
- We cannot make probabilistic statements about $\hat{\theta}_n(x_1, \ldots, x_n)$.

Note: To make matters worse, $\hat{\theta}_n$ often denotes both the estimator (random) and the estimate (fixed), so that you have to figure it out yourself from context!

On the Interpretation of Estimates (Contd.)

The confusion between estimators (random) and estimates (fixed) is particularly severe in the context of confidence intervals.

Recall that an asymptotic $1-\alpha$ confidence interval is such that

$$P(\theta \in C_n) \to 1-\alpha.$$

Let c_n denote a realization of C_n (i.e., what you computed using data).

- It is correct to say C_n covers θ with prob. of $1-\alpha$.
- It is incorrect to say c_n covers θ with prob. of $1-\alpha$.
- $P(\theta \in c_n) = \mathbb{1}\{\theta \in c_n\} \in \{0,1\}$. This is a comparison of numbers!

On the Interpretation of Estimates (Contd.)

Statistics courses often introduce the idea of repeated experiments to interpret confidence intervals:

- "If I were to repeat the same experiment again and again, each time computing a $1-\alpha$ confidence interval, then the confidence intervals would cover the true parameter $100(1-\alpha)\%$ of the time."

On the Interpretation of Estimates (Contd.)

Example Consider $\hat{\mu}_n^{(3)}$. Suppose that we collected data and that $\hat{\mu}_n^{(3)} = 10$, and $\text{se}(\hat{\mu}_n^{(3)}) = 3$.

Then, an asymptotic $1 - \alpha$ confidence interval is given by

$$c_n = (10 - 1.96 \times 3, 10 + 1.96 \times 3) = (4.12, 15.88)\,.$$

Here c_n denotes a realization of the confidence interval C_n:

- We've collected data (a realization of our sample);
- Computed the estimator $\hat{\mu}_n^{(3)}$ and its standard error $\text{se}(\hat{\mu}_n^{(3)})$;
- Calculated a $1 - \alpha$ confidence interval c_n.

What is $P(\theta \in c_n)$? We don't know.

Summary

This concludes the Part C 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 Part D, we discuss how estimators and their (approximate) sampling distributions can be leveraged to assess whether the true parameter θ takes a particular value, say, θ_0. This is known as hypothesis testing.