Chapter 5

An Introduction to Asymptotic Theory

Our universe, though enormous, consists of fewer than 10^{82} atoms, which is a finite number. However, mathematical ideas are not bounded by secular realities. Asymptotic theory is about behaviors of statistics when the sample size is arbitrarily large up to infinity. It is a set of approximation techniques to simplify complicated finite-sample analysis. Asymptotic theory is the cornerstone of modern econometrics. It sheds lights on estimation and inference procedures under much more general conditions than what are covered by exact finite sample theory.

Nevertheless, we always have at hand a finite sample, and mostly it is difficult to increase the sample size in reality. Asymptotic theory rarely answers "how large is large", and we must be cautious about the treacher-

ous landscape of *asymptopia*. In the era of big data, albeit the sheer size of data balloons dramatically, we build more sophisticated models to better capture heterogeneity in the data. Large sample is a relative notion to the complexity of the model and underlying (in)dependence structure of the data.

Both the classical parametric approach, which is based on hard-to-verify parametric assumptions, and the asymptotic approach, which is predicated on imaginary infinite sequences, deviate from the reality. Which approach is more constructive can only be judged case by case. The prevalence of asymptotic theory is its mathematical amenability and generality. The law of evolution elevates asymptotic theory to the throne of mathematical statistics of our time.

5.1 Modes of Convergence

We first review what is *convergence* for a non-random sequence, which you learned in high school. Let $z_1, z_2, \ldots$ be an infinite sequence of non-random variables.

Definition 5.1. Convergence of this non-random sequence means that for any $\varepsilon > 0$, there exists an $N(\varepsilon)$ such that for all $n > N(\varepsilon)$, we have $|z_n - z| < \varepsilon$. We say z is the limit of z_n, and write $z_n \to z$ or $\lim_{n\to\infty} z_n = z$.

Instead of a deterministic sequence, we are interested in the convergence of a sequence of random variables. Since a random variable is "ran-

dom" thanks to the induced probability measure by the measurable function, we must be clear what *convergence* means. Several modes of convergence are widely used.

Definition 5.2 (Convergence in probability)**.** We say a sequence of random variables (z_n) converges in probability to z, where z can be either a random variable or a non-random constant, if for any $\varepsilon > 0$, the probability $P\{\omega : |z_n(\omega) - z| < \varepsilon\} \to 1$ (or equivalently $P\{\omega : |z_n(\omega) - z| \geq \varepsilon\} \to 0$) as $n \to \infty$. We can write $z_n \xrightarrow{p} z$ or $\text{plim}_{n\to\infty} z_n = z$.

Definition 5.3 (Squared-mean convergence)**.** A sequence of random variables (z_n) converges in squared-mean to z, where z can be either a random variable or a non-random constant, if $E\left[(z_n - z)^2\right] \to 0$. It is denoted as $z_n \xrightarrow{m.s.} z$.

In these definitions either $P\{\omega : |z_n(\omega) - z| > \varepsilon\}$ or $E\left[(z_n - z)^2\right]$ is a non-random quantity, and it converges to 0 as a non-random sequence.

Squared-mean convergence is stronger than convergence in probability. That is, $z_n \xrightarrow{m.s.} z$ implies $z_n \xrightarrow{p} z$ but the converse is untrue. Here is an example.

Example 5.1. (z_n) is a sequence of binary random variables: $z_n = \sqrt{n}$ with probability $1/n$, and $z_n = 0$ with probability $1 - 1/n$. Then $z_n \xrightarrow{p} 0$ but $z_n \xrightarrow{m.s.} 0$. To verify these claims, notice that for any $\varepsilon > 0$, we have $P(\omega : |z_n(\omega) - 0| < \varepsilon) = P(\omega : z_n(\omega) = 0) = 1 - 1/n \to 1$ and thereby

$z_n \xrightarrow{p} 0$. On the other hand, $E\left[(z_n - 0)^2\right] = n \cdot 1/n + 0 \cdot (1 - 1/n) = 1 \nrightarrow 0$, so $z_n \xrightarrow{m.s.} 0$.

Remark 5.1. Example 5.1 highlights the difference between the two modes of convergence. Convergence in probability does not count what happens on a subset in the sample space of small probability. Squared-mean convergence deals with the average over the entire probability space. If a random variable can take a wild value, with small probability though, it may blow away the squared-mean convergence. On the contrary, such irregularity does not undermine convergence in probability.

Both convergence in probability and squared-mean convergence are about convergence of random variables to a target random variable or constant. That is, the distribution of $z_n - z$ is concentrated around 0 as $n \to \infty$. Instead, *convergence in distribution* is about the convergence of CDF, but not the random variable. Let $F_{z_n}(\cdot)$ be the CDF of z_n and $F_z(\cdot)$ be the CDF of z.

Definition 5.4 (Convergence in distribution)**.** We say a sequence of random variables (z_n) converges in distribution to a random variable z if $F_{z_n}(a) \to F_z(a)$ as $n \to \infty$ at each point $a \in \mathbb{R}$ such that where $F_z(\cdot)$ is continuous. We write $z_n \xrightarrow{d} z$.

Convergence in distribution is the weakest mode. If $z_n \xrightarrow{p} z$, then $z_n \xrightarrow{d} z$. The converse is not true in general, unless z is a non-random constant (A constant z can be viewed as a degenerate random variables, with a

corresponding "CDF" $F_z(\cdot) = 1\{\cdot \geq z\}$.

Example 5.2. Let $x \sim N(0,1)$. If $z_n = x + 1/n$, then $z_n \xrightarrow{p} x$ and of course $z_n \xrightarrow{d} x$. However, if $z_n = -x + 1/n$, or $z_n = y + 1/n$ where $y \sim N(0,1)$ is independent of x, then $z_n \xrightarrow{d} x$ but $z_n \xrightarrow{p} x$.

Example 5.3. (z_n) is a sequence of binary random variables: $z_n = n$ with probability $1/\sqrt{n}$, and $z_n = 0$ with probability $1 - 1/\sqrt{n}$. Then $z_n \xrightarrow{d} z = 0$. Because

$$F_{z_n}(a) = \begin{cases} 0 & a < 0 \\ 1 - 1/\sqrt{n} & 0 \leq a \leq n \\ 1 & a \geq n \end{cases}.$$

$F_z(a) = \begin{cases} 0, & a < 0 \\ 1 & a \geq 0 \end{cases}$. It is easy to verify that $F_{z_n}(a)$ converges to $F_z(a)$ *pointwisely* on each point in $(-\infty, 0) \cup (0, +\infty)$, where $F_z(a)$ is continuous.

So far we have talked about convergence of scalar variables. These three modes of converges can be easily generalized to random vectors. In particular, the *Cramer-Wold device* collapses a random vector into a random vector via arbitrary linear combination. We say a sequence of K-dimensional random vectors (z_n) converge in distribution to z if $\lambda' z_n \xrightarrow{d} \lambda' z$ for any $\lambda \in \mathbb{R}^K$ and $\|\lambda\|_2 = 1$.

5.2 Law of Large Numbers

(Weak) law of large numbers (LLN) is a collection of statements about convergence in probability of the sample average to its population counterpart. The basic form of LLN is:

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - E[z_i]) \xrightarrow{p} 0$$

as $n \to \infty$. Various versions of LLN work under different assumptions about some features and/or dependence of the underlying random variables.

5.2.1 Cherbyshev LLN

We illustrate LLN by the simple example of Chebyshev LLN, which can be proved by elementary calculation. It utilizes the *Chebyshev inequality*.

- *Chebyshev inequality*: If a random variable x has a finite second moment $E\left[x^2\right] < \infty$, then we have $P\left\{|x| > \varepsilon\right\} \leq E\left[x^2\right]/\varepsilon^2$ for any constant $\varepsilon > 0$.

Exercise 5.1. Show that if $r_2 \geq r_1 \geq 1$, then $E\left[|x|^{r_2}\right] < \infty$ implies $E\left[|x|^{r_1}\right] < \infty$. (Hint: use Holder's inequality.)

The Chebyshev inequality is a special case of the *Markov inequality*.

- *Markov inequality*: If a random variable x has a finite r-th absolute

moment $E\left[|x|^r\right] < \infty$ for some $r \geq 1$, then we have $P\{|x| > \varepsilon\} \leq E\left[|x|^r\right]/\varepsilon^r$ any constant $\varepsilon > 0$.

Proof. It is easy to verify the Markov inequality.

$$\begin{aligned} E\left[|x|^r\right] &= \int_{|x|>\varepsilon} |x|^r \, dF_X + \int_{|x|\leq\varepsilon} |x|^r \, dF_X \\ &\geq \int_{|x|>\varepsilon} |x|^r \, dF_X \\ &\geq \varepsilon^r \int_{|x|>\varepsilon} dF_X = \varepsilon^r P\{|x| > \varepsilon\}. \end{aligned}$$

Rearrange the above inequality and we obtain the Markov inequality. □

Let the *partial sum* $S_n = \sum_{i=1}^n x_i$, where $\mu_i = E[x_i]$ and $\sigma_i^2 = \text{var}[x_i]$. We apply the Chebyshev inequality to the sample mean $z_n = \overline{x} - \bar{\mu} = n^{-1}(S_n - E[S_n])$.

$$\begin{aligned} P\{|z_n| \geq \varepsilon\} &= P\left\{n^{-1}|S_n - E[S_n]| \geq \varepsilon\right\} \\ &\leq E\left[\left(n^{-1}\sum_{i=1}^n (x_i - \mu_i)\right)^2\right]/\varepsilon^2 \\ &= (n\varepsilon)^{-2}\left\{E\left[\sum_{i=1}^n (x_i - \mu_i)^2\right] + \sum_{i=1}^n \sum_{j\neq i} E\left[(x_i - \mu_i)(x_j - \mu_j)\right]\right\} \\ &= (n\varepsilon)^{-2}\left\{\sum_{i=1}^n \text{var}(x_i) + \sum_{i=1}^n \sum_{j\neq i} \text{cov}(x_i, x_j)\right\}. \end{aligned} \tag{5.1}$$

Convergence in probability holds if the right-hand side shrinks to 0 as $n \to \infty$. For example, If $x_1, \ldots, x_n$ are iid with $\text{var}(x_1) = \sigma^2$, then the RHS

of (5.1) is $(n\varepsilon)^{-2}\left(n\sigma^2\right) = o\left(n^{-1}\right) \to 0$. This result gives the Chebyshev LLN:

- Chebyshev LLN: If $(z_1, \ldots, z_n)$ is a sample of iid observations, $E[z_1] = \mu$, and $\sigma^2 = \text{var}[z_1] < \infty$ exists, then $\frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{p} \mu$.

The convergence in probability can be indeed maintained under much more general conditions than under iid case. The random variables in the sample do not have to be identically distributed, and they do not have to be independent either.

Exercise 5.2. Consider an inid (independent but non-identically distributed) sample $(x_1, \ldots, x_n)$ with $E[x_i] = 0$ and $\text{var}[x_i] = \sqrt{n}\sigma_i^2$. Use the Chebyshev inequality to show that $n^{-1}\sum_{i=1}^{n} x_i \xrightarrow{p} 0$.

Exercise 5.3. Consider the time series moving average model $x_i = \varepsilon_i + \theta\varepsilon_{i-1}$ for $i = 1, \ldots, n$, where $|\theta| < 1$, $E[\varepsilon_i] = 0$, $\text{var}[\varepsilon_i] = \sigma^2$, and $(\varepsilon_i)_{i=0}^{n}$ iid. Use the Chebyshev inequality to show that $n^{-1}\sum_{i=1}^{n} x_i \xrightarrow{p} 0$.

Another useful LLN is the *Kolmogorov LLN*. Since its derivation requires more advanced knowledge of probability theory, we state the result without proof.

- Kolmogorov LLN: If $(z_1, \ldots, z_n)$ is a sample of iid observations and $E[z_1] = \mu$ exists, then $\frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{p} \mu$.

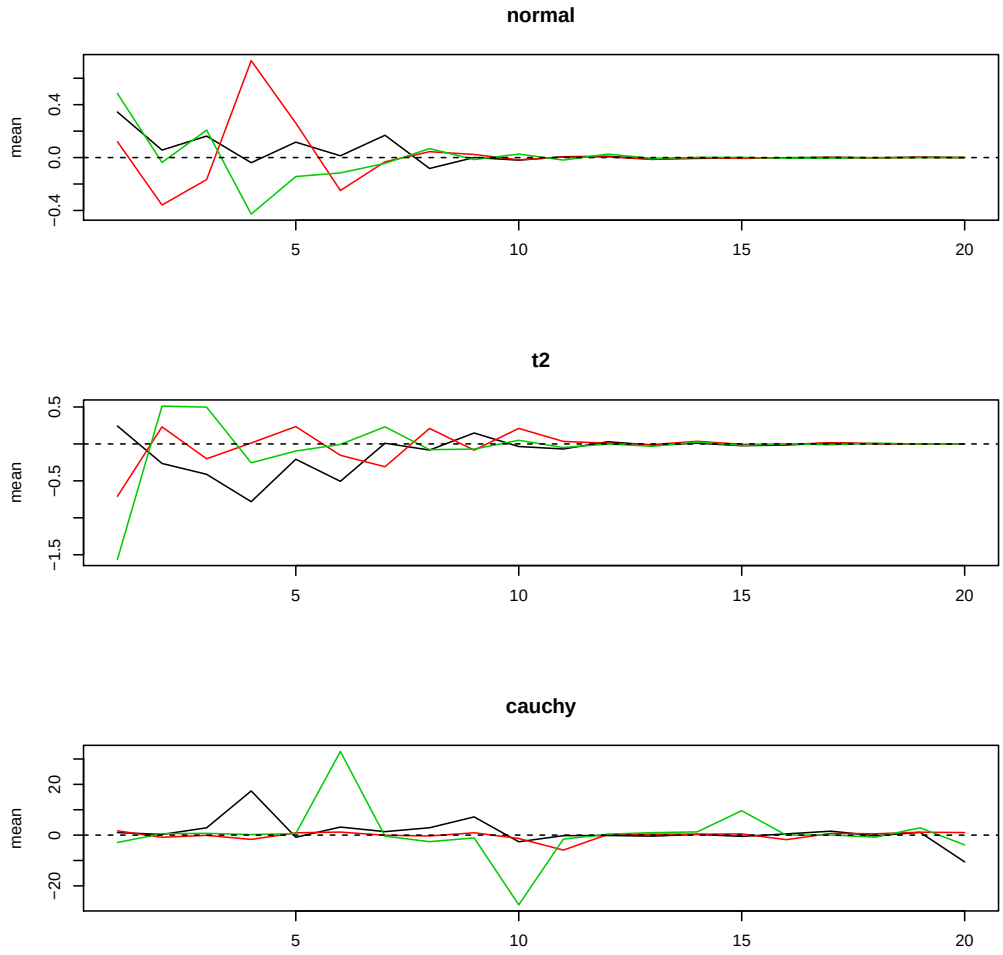
Compared with the Chebyshev LLN, the Kolmogorov LLN only requires the existence of the population mean, but not any higher moments. On the other hand, iid is essential for the Kolmogorov LLN.

Example 5.4. Consider three distributions: standard normal $N(0,1)$, $t(2)$ (zero mean, infinite variance), and the Cauchy distribution (no moments exist). We plot paths of the sample average with $n = 2^1, 2^2, \ldots, 2^{20}$. We will see that the sample averages of $N(0,1)$ and $t(2)$ converge, but that of the Cauchy distribution does not.

```
sample.mean = function( n, distribution ){
  # get sample mean for a given distribution
  if (distribution == "normal"){ y = rnorm( n ) }
  else if (distribution == "t2") {y = rt(n, 2) }
  else if (distribution == "cauchy") {y = rcauchy(n) }
  return( mean(y) )
}


LLN.plot = function(distribution){
  # draw the sample mean graph
  ybar = matrix(0, length(NN), 3 )
  for (rr in 1:3){
    for ( ii in 1:length(NN)){
      n = NN[ii]; ybar[ii, rr] = sample.mean(n, distribution)
    }
  }
  matplot(ybar, type = "l", ylab = "mean", xlab = "",
```

```
       lwd = 1, lty = 1, main = distribution)
  abline(h = 0, lty = 2)
  return(ybar)
}
# calculation
NN = 2^(1:20); set.seed(2020-10-7); par(mfrow = c(3,1))
l1 = LLN.plot("normal"); l2 = LLN.plot("t2"); l3 = LLN.plot("cauchy")
```



5.3 Central Limit Theorem

The central limit theorem (CLT) is a collection of probability results about the convergence in distribution to a stable distribution. The limiting distribution is usually the Gaussian distribution. The basic form of the CLT

is:

- *Under some conditions to be spelled out*, the sample average of *zero-mean* random variables $(z_1, \ldots, z_n)$ multiplied by $\sqrt{n}$ satisfies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \xrightarrow{d} N\left(0, \sigma^2\right)$$

  as $n \to \infty$.

Various versions of CLT work under different assumptions about the random variables. *Lindeberg-Levy CLT* is the simplest CLT.

- If the sample $(x_1, \ldots, x_n)$ is iid, $E[x_1] = 0$ and $\text{var}\left[x_1^2\right] = \sigma^2 < \infty$, then $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \xrightarrow{d} N\left(0, \sigma^2\right)$.

Lindeberg-Levy CLT can be proved by the *moment generating function*. For any random variable x, the function $M_x(t) = E[\exp(xt)]$ is called its the *moment generating function* (MGF) if it exists. MGF fully describes a distribution, just like PDF or CDF. For example, the MGF of $N\left(\mu, \sigma^2\right)$ is $\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$.

Heuristic proof of Lindeberg-Levy CLT. If $E\left[|x|^k\right] < \infty$ for a positive integer k, then

$$M_X(t) = 1 + tE[X] + \frac{t^2}{2}E\left[X^2\right] + \ldots \frac{t}{k!}E\left[X^k\right] + O\left(t^{k+1}\right).$$

Under the assumption of Lindeberg-Levy CLT,

$$M_{\frac{X_i}{\sqrt{n}}}(t) = 1 + \frac{t^2}{2n}\sigma^2 + O\left(\frac{t^3}{n^{3/2}}\right)$$

for all i, and by independence we have

$$\begin{aligned} M_{\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i}(t) = \prod_{i=1}^n M_{\frac{X_i}{\sqrt{n}}}(t) &= \left(1 + \frac{t^2}{2n}\sigma^2 + O\left(\frac{t^3}{n^{3/2}}\right)\right)^n \\ &\to \exp\left(\frac{\sigma^2}{2}t^2\right), \end{aligned}$$

where the limit is exactly the characteristic function of $N\left(0, \sigma^2\right)$. □

Remark 5.2. This proof with MGF is simple and elementary. Its drawback is that not all distributions have a well-defined MGF. A more general proof can be carried out by replacing MGF with the *characteristic function* $\varphi_x(t) = E\left[\exp\left(\mathrm{i}xt\right)\right]$, where "i" is the imaginary number. The characteristic function is the *Fourier transform* of the probability measure and it always exists. Such a proof will require background knowledge of Fourier transform and inverse transform, which we do not pursuit here.

- Lindeberg-Feller CLT: $(x_i)_{i=1}^n$ is inid. If the *Lindeberg condition* is satisfied (for any fixed $\varepsilon > 0$, $\frac{1}{s_n^2}\sum_{i=1}^n E\left[x_i^2 \cdot \mathbf{1}\left\{|x_i| \geq \varepsilon s_n\right\}\right] \to 0$ where $s_n = \sqrt{\sum_{i=1}^n \sigma_i^2}$), then we have

$$\frac{\sum_{i=1}^n x_i}{s_n} \xrightarrow{d} N(0,1).$$

- Lyapunov CLT: $(x_i)_{i=1}^n$ is inid. If $\max_{i \leq n} E\left[|x_i|^3\right] < C < \infty$, then we have

$$\frac{\sum_{i=1}^n x_i}{s_n} \xrightarrow{d} N(0,1).$$

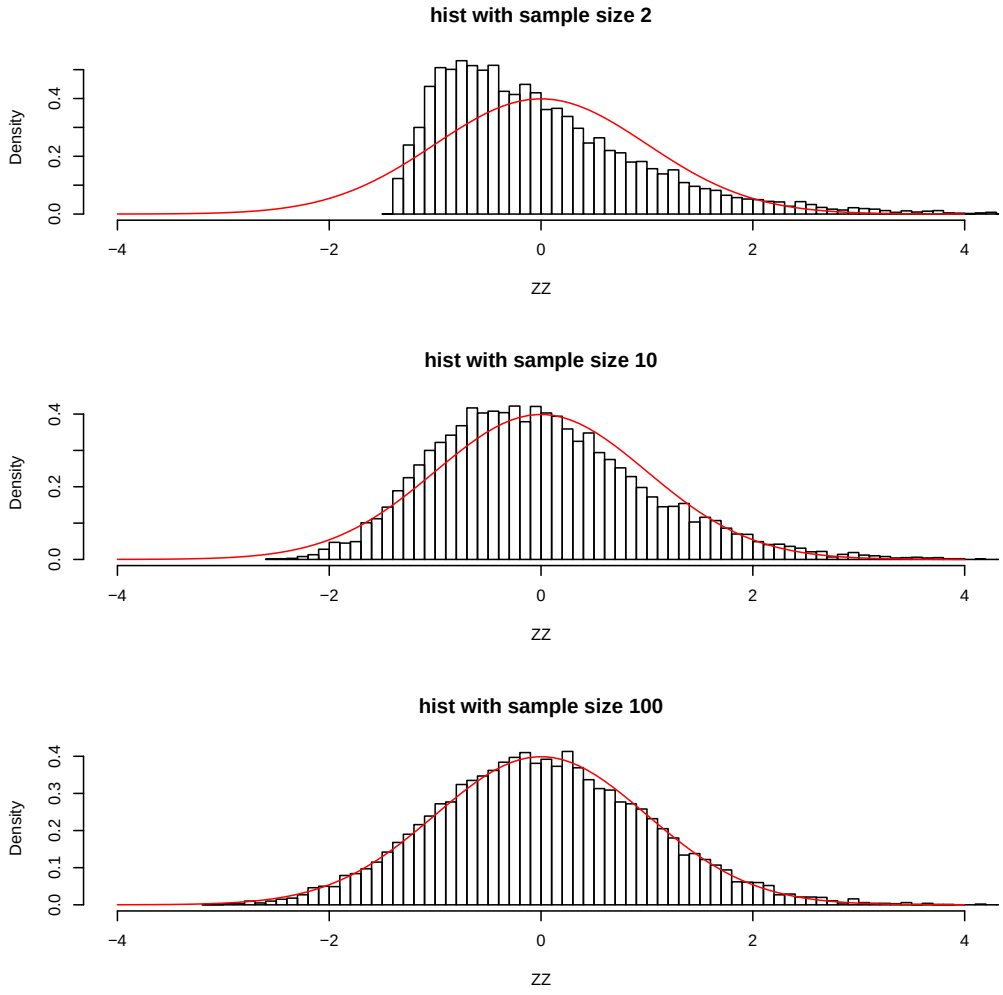
This is a simulated example.

```
Z_fun = function(n, distribution){
  if (distribution == "normal"){
      z = sqrt(n) * mean(rnorm(n))
        } else if (distribution == "chisq2") {
      df = 2;
      x = rchisq(n,2)
      z = sqrt(n) * ( mean(x) - df ) / sqrt(2*df)
      }
  return (z)
}
CLT_plot = function(n, distribution){
  Rep = 10000
  ZZ = rep(0, Rep)
  for (i in 1:Rep) {ZZ[i] = Z_fun(n, distribution)}

  xbase = seq(-4.0, 4.0, length.out = 100)
  hist( ZZ, breaks = 100, freq = FALSE,
    xlim = c( min(xbase), max(xbase) ),
```

```
    main = paste0("hist with sample size ", n) )
  lines(x = xbase, y = dnorm(xbase), col = "red")
  return (ZZ)
}


par(mfrow = c(3,1))
phist = CLT_plot(2, "chisq2")
phist = CLT_plot(10, "chisq2")
phist = CLT_plot(100, "chisq2")
```



5.4 Tools for Transformations

In their original forms, LLN deals with the sample mean, and CLT handles the scaled (by $\sqrt{n}$) and/or standardized (by standard deviation) sample mean. However, most of the econometric estimators of interest are func-

tions of sample means. For example, in the OLS estimator

$$\widehat{\beta} = \left(\frac{1}{n} \sum_i x_i x_i' \right)^{-1} \frac{1}{n} \sum_i x_i y_i$$

involves matrix inverse and the matrix-vector multiplication. We need tools to handle transformations.

- Continuous mapping theorem 1: If $x_n \xrightarrow{p} a$ and $f(\cdot)$ is continuous at a, then $f(x_n) \xrightarrow{p} f(a)$.
- Continuous mapping theorem 2: If $x_n \xrightarrow{d} x$ and $f(\cdot)$ is continuous almost surely on the support of x, then $f(x_n) \xrightarrow{d} f(x)$.
- Slutsky's theorem: If $x_n \xrightarrow{d} x$ and $y_n \xrightarrow{p} a$, then
  - $x_n + y_n \xrightarrow{d} x + a$
  - $x_n y_n \xrightarrow{d} ax$
  - $x_n / y_n \xrightarrow{d} x/a$ if $a \neq 0$.

Slutsky's theorem consists of special cases of the continuous mapping theorem 2. Only because the addition, multiplication and division are encountered so frequently in practice, we list it as a separate theorem.

- Delta method: if $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Omega)$, and $f(\cdot)$ is continuously differentiable at θ_0 (meaning $\frac{\partial}{\partial \theta} f(\cdot)$ is continuous at θ_0), then we

have

$$\sqrt{n}\left(f\left(\widehat{\theta}\right)-f\left(\theta_0\right)\right) \xrightarrow{d} N\left(0, \frac{\partial f}{\partial \theta'}\left(\theta_0\right) \Omega\left(\frac{\partial f}{\partial \theta}\left(\theta_0\right)\right)'\right).$$

Proof. Take a Taylor expansion of $f\left(\widehat{\theta}\right)$ around $f\left(\theta_0\right)$:

$$f\left(\widehat{\theta}\right)-f\left(\theta_0\right)=\frac{\partial f\left(\dot{\theta}\right)}{\partial \theta'}\left(\widehat{\theta}-\theta_0\right),$$

where $\dot{\theta}$ lies on the line segment between $\widehat{\theta}$ and θ_0. Multiply $\sqrt{n}$ on both sides,

$$\sqrt{n}\left(f\left(\widehat{\theta}\right)-f\left(\theta_0\right)\right)=\frac{\partial f\left(\dot{\theta}\right)}{\partial \theta'}\sqrt{n}\left(\widehat{\theta}-\theta_0\right).$$

Because $\widehat{\theta} \xrightarrow{p} \theta_0$ implies $\dot{\theta} \xrightarrow{p} \theta_0$ and $\frac{\partial}{\partial \theta'} f\left(\cdot\right)$ is continuous at θ_0, we have $\frac{\partial}{\partial \theta'} f\left(\dot{\theta}\right) \xrightarrow{p} \frac{\partial f\left(\theta_0\right)}{\partial \theta'}$ by the continuous mapping theorem 1. In view of $\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, \Omega\right)$, Slutsky's Theorem implies

$$\sqrt{n}\left(f\left(\widehat{\theta}\right)-f\left(\theta_0\right)\right) \xrightarrow{d} \frac{\partial f\left(\theta_0\right)}{\partial \theta'} N\left(0, \Omega\right)$$

and the conclusion follows. □

5.5 Summary

Asymptotic theory is a topic with vast breadth and depth. In this chapter we only scratch the very surface of it. We will discuss in the next chapter

how to apply the asymptotic tools we learned here to the OLS estimator.

Historical notes: Before 1980s, most econometricians did not have a good training in mathematical rigor to master asymptotic theory. A few prominent young (at that time) econometricians came to the field and changed the situation, among them were Halbert White (UCSD), Peter C.B. Phillips (Yale) and Peter Robinson (LSE), to name a few.

Further reading: Halbert White (1950-2012) wrote an accessible book (White, 2000, first edition 1984) to introduce asymptotics to econometricians. This book remains popular among researchers and graduate students in economics. Davidson (1994) is a longer and more self-contained monograph.

Bibliography

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. Oxford University Press. 5.5

White, H. (2000). *Asymptotic theory for econometricians*. Academic Press. 5.5

Zhentao Shi. October 15, 2020