Chapter 1

Asymptotics

In practice, what we are interested is the finite sample behavior of estimators or test statistics. However, the finite sample exact distributions count on underlying data generating processes and are often too mathematically difficult to deduce and hard to generalize. Asymptotic theory, on the other hand, uses a thought experiment to imagine what happens when the sample size is arbitrarily large. It is a surprising fact that despite the multitude of finite sample settings, under some reasonable conditions, the different starting points often end up with the same result. For example, the limit distribution of the central limit theorems is usual a normal distribution. The convenience of the large sample theory makes it the prevalent apparatus of modern statistical theory and econometric theory.

Though the large sample theory is popular, we must keep in mind that it is a mathematical tool for approximation, and the quality of approximation depend on unknown implicit quantities. Monte Carlo simulations and numerical experiments are useful to deepen our understanding of asymptotic results in finite sample.

1.1 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 1.1. 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 "random", we must be clear what *convergence* means. Several modes of convergence are widely used.

Definition 1.2 (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 1.3 (r-th moment 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 $E\left[|z_n - z|^r\right] \to$

0 for some $r \geq 1$. It is denoted as $z_n \xrightarrow{rth.m.} z$. In particular, when $r = 2$ it is called square-mean convergence, written as $z_n \xrightarrow{m.s.} z$.

In these definitions either $P\{\omega : |z_n(\omega) - z| > \varepsilon\}$ or $E\left[|z_n - z|^r\right]$ 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, $z_n \xrightarrow{rth.m.} z$ implies $z_n \xrightarrow{p} z$ but the converse is untrue. Here is an example.

Example 1.1. (z_n) is a sequence of binary random variables: $z_n = n$ with probability $1/n$, and $z_n = 0$ with probability $1 - 1/n$. Then $z_n \xrightarrow{p} 0$ but $z_n \overset{1st.m.}{\nrightarrow} 0$. To verify these claims, notice that for any $\varepsilon > 0$, we have $P(\omega : |z_n(\omega) - 0| > \varepsilon) = P(\omega : z_n(\omega) = n) = 1/n \to 0$ and thereby $z_n \xrightarrow{p} 0$. On the other hand, $E[|z_n - 0|] = n \cdot 1/n + 0 \cdot (1 - 1/n) = 1 \nrightarrow 0$, so $z_n \overset{m.s.}{\nrightarrow} 0$.

Remark 1.1. Example 1.1 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 1.4 (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 1.2. 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 $z_n \overset{p}{\nrightarrow} x$.

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 $\|\lambda\|_2 = 1$.

1.2 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.

1.2.1 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 $P\{|x| > \varepsilon\} \leq E\left[x^2\right]/\varepsilon^2$ for any constant $\varepsilon > 0$.

Exercise 1.1. 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.

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

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 $z_n = \bar{x} - \bar{\mu} = n^{-1}(S_n - E[S_n])$.

$$\begin{aligned}
P\{|z_n| \geq \varepsilon\} &= P\left\{n^{-1}|S_n - E[S_n]| \geq \varepsilon\right\} \\
&\leq E\left[\left(n^{-1}\sum_{i=1}^n (x_i - \mu_i)\right)^2\right]/\varepsilon^2 \\
&= (n\varepsilon)^{-2}\left\{E\left[\sum_{i=1}^n (x_i - \mu_i)^2\right] + \sum_{i=1}^n \sum_{j\neq i} E\left[(x_i - \mu_i)(x_j - \mu_j)\right]\right\} \\
&= (n\varepsilon)^{-2}\left\{\sum_{i=1}^n \text{var}(x_i) + \sum_{i=1}^n \sum_{j\neq i} \text{cov}(x_i, x_j)\right\}. \qquad (1.1)
\end{aligned}$$

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 (1.1) is $(n\varepsilon)^{-2}(n\sigma^2) = o(n^{-1}) \to 0$. 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 1.2. Consider an inid (independent but non-identically distributed) sample $(x_1, \ldots, x_n)$ with $E[x_i] = 0$ and $\text{var}[x_i] = \sqrt{n}c$ for some constant $c > 0$. 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 $\frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{a.s.} \mu$.

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.

1.3 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 $\operatorname{var}[x_1] = \sigma^2 < \infty$, 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 $\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$.

1.4 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 functions 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

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

To see why the Delta method works, take a Taylor expansion of $f\left(\widehat{\theta}\right)$ around $f(\theta_0)$:

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

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

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

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

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

and the conclusion follows.

1.5 OLS

We apply these results to study asymptotic properties of the OLS estimator. We will show (i) $\widehat{\beta} = (X'X)^{-1} X'Y$ is a consistent estimator of the linear projection coefficient β; (ii) $\widehat{\beta}$ is asymptotically normal; (iii) the asymptotic normality allows asymptotic inference of β.

1.5.1 Consistency

Consistency is a basic requirement for estimators in large sample. Intuitively, it says that when the sample size is arbitrarily large, a desirable estimator should be arbitrarily close (in the sense of convergence in probability) to the population quantity of interest. Otherwise, if an estimator still deviates from the object of interest under infinite sample size, it is hard to persuade other researchers to use such an estimator unless compelling justification is provided.

Definition 1.5 (Consistency)**.** For a generic estimator $\widehat{\theta}$, we say $\widehat{\theta}$ is *consistent* for θ if $\widehat{\theta} \xrightarrow{p} \theta$, where θ is some non-random object.

In OLS, we say $\widehat{\beta}$ is *consistent* if $\widehat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$, where β is the linear projection coefficient of the population model $y_i = x_i'\beta + e_i$ with $E[x_i e_i] = 0$. To verify consistency, we write

$$\widehat{\beta} - \beta = \left(X'X\right)^{-1} X'e = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i e_i. \tag{1.2}$$

For simplicity, in this chapter we discuss the iid setting only. The first term, by LLN,

$$\widehat{Q} := \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} Q := E\left[x_i x_i'\right].$$

Here $\widehat{Q}$ is the sample mean of $x_i x_i'$ and Q is the population mean of $x_i x_i'$. The second term, again by LLN,

$$\frac{1}{n}\sum_{i=1}^{n} x_i e_i \xrightarrow{p} 0.$$

The continuous mapping theorem immediately implies

$$\widehat{\beta} - \beta \xrightarrow{p} Q^{-1} \times 0 = 0.$$

The OLS estimator $\widehat{\beta}$ is a consistent estimator of β.

Remark 1.2. No matter whether $(y_i, x_i)_{i=1}^n$ is an iid, or inid, or dependent sample, consistency holds as long as the convergence in probability holds for the above two expressions and Q is an invertible matrix.

1.5.2 Asymptotic Distribution

In finite sample, $\widehat{\beta}$ is a random variable. We have shown the distribution of $\widehat{\beta}$ under normality before. Without restrictive distribtional assumptions, how can we characterize the randomness of the OLS estimator? We know from the previous section that $\hat{\beta} - \beta \xrightarrow{p} 0$ degenerates to a constant. To study its distribution, we must scale it up by a proper multiplier so that in the limit it neither degenerates nor explodes. The suitable scaling factor is $\sqrt{n}$, as in a CLT.

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i.$$

Since $E[x_i e_i] = 0$, we apply a CLT to obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i \xrightarrow{d} N(0, \Sigma)$$

where $\Sigma = E\left[x_i x_i' e_i^2\right]$. By the continuous mapping theorem,

$$\sqrt{n}\left(\widehat{\beta}-\beta\right) \xrightarrow{d} Q^{-1} \times N(0,\Sigma) \sim N(0,\Omega) \tag{1.3}$$

where $\Omega = Q^{-1}\Sigma Q^{-1}$ is called the *asymptotic variance*. This result is the *asymptotic normality* of the OLS estimator.

The asymptotic variance $\Omega = Q^{-1}\Sigma Q^{-1}$ is called of the *sandwich form*. It can be simplified under conditional homoskedasticity $E\left[e_i^2|x_i\right] = \sigma^2$ for all i, which gives

$$\Sigma = E\left[x_i x_i' e_i^2\right] = E\left[x_i x_i' E\left[e_i^2|X\right]\right] = \sigma^2 E\left[x_i x_i'\right] = \sigma^2 Q.$$

In this case, $\Omega = Q^{-1}\Sigma Q^{-1} = \sigma^2 Q^{-1}$, and thus

$$\sqrt{n}\left(\widehat{\beta}-\beta\right) \xrightarrow{d} N\left(0,\sigma^2 Q^{-1}\right). \tag{1.4}$$

Remark 1.3. If we are interested in the k-th parameter β_k, then the joint distribution in (1.3) implies

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}_k-\beta_k\right) &= \sqrt{n}\eta_k'\left(\widehat{\beta}-\beta\right) \\ &\xrightarrow{d} N\left(0,\sigma^2\eta_k' Q^{-1}\eta_k\right) \sim N\left(0,\sigma^2[Q^{-1}]_{kk}\right),\end{aligned} \tag{1.5}$$

where $\eta_k = (0,\ldots,0,1,0\ldots,0)'$ is the selector of the k-th element.

Remark 1.4. If $\Omega^{-1/2}$ is multiplied on both sides of (1.3), we have

$$\Omega^{-1/2}\sqrt{n}\left(\widehat{\beta}-\beta\right) \xrightarrow{d} N(0,I_K). \tag{1.6}$$

We say the asymptotic distribution in (1.6), $N(0,I_K)$, is *pivotal* because it does not involve any unknown parameter. In contrast, the asymptotic distribution in (1.3) is not pivotal because Ω is unknown in $N(0,\Omega)$. If we are interested in the k-th parameter β_k, we can write (1.6) into the pivotal form as

$$\frac{\sqrt{n}\left(\widehat{\beta}_k-\beta_k\right)}{\sqrt{\sigma^2[Q^{-1}]_{kk}}} \xrightarrow{d} N(0,1). \tag{1.7}$$

In reality the variance is unknown. We can use a plug-in estimator to approximate it. Under regularization conditions, the plug-in estimator will deliver feasible and asymptotically valid inference.