November 15, 2021

# Large Sample Theory

## Limits and convergence concepts: in probability and in mean

Let $\{a_n : n = 1, 2, \ldots\}$ be a sequence of non-random real numbers. We say that $a$ is the limit of $\{a_n\}$ if for *all* real $\delta > 0$ we can find an integer $N_\delta$ such that for *all* $n \geq N_\delta$ we have that $|a_n - a| < \delta$. When the limit exists, we say that $\{a_n\}$ converges to $a$, and write $a_n \to a$ or $\lim_{n\to\infty} a_n = a$. In this case, we can make the elements of $\{a_n : n \geq N\}$ arbitrary close to $a$ by choosing $N$ sufficiently large. Naturally, if $a_n \to a$, we have that $a_n - a \to 0$.

The concept can be extended to vectors or matrices as well. Let $\{\boldsymbol{A}_n : n = 1, 2, \ldots\}$ be a $m \times k$ matrix. Then $\boldsymbol{A}_n \to \boldsymbol{A}$ if for all $i = 1, \ldots, m$ and $j = 1, \ldots, k$ we have that the $(i, j)$-th element of $\boldsymbol{A}_n$ converges to the $(i, j)$-th element of $\boldsymbol{A}$.

The concept of convergence cannot be applied in a straightforward way to sequences of random variables. This is so because a random variable is a *function* from the sample space $\Omega$ to the real line. The solution is to consider convergence of a *non-random* sequence derived from the random one. Since there are many ways to derive non-random sequences, there exist several stochastic convergence concepts. Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of *random* variables. Let $X$ be random or non-random (i.e. it is possible that $X(\omega)$ is the same for all $\omega \in \Omega$). We will consider *non-random* sequences with the following typical elements: (i) $\mathbb{E}\left|X_n - X\right|^r$, and (ii) $\Pr\left(|X_n - X| > \varepsilon\right)$ for some $\varepsilon > 0$. These are sequences of non-random real numbers, and, consequently, the usual definition of convergence applies to each of them leading to a corresponding definition of stochastic convergence:

**(i) Convergence in $r$-th mean.** $X_n$ converges to $X$ in $r$-th mean if $\mathbb{E}\left|X_n - X\right|^r \to 0$ as $n \to \infty$.

**(ii) Convergence in probability.** $X_n$ converges in probability to $X$ if for all $\varepsilon > 0$, $\Pr\left(|X_n - X| \geq \varepsilon\right) \to 0$ as $n \to \infty$. It is denoted as $X_n \to_p X$ or $p\lim X_n = X$. Alternatively, convergence in probability can be defined as $\Pr\left(|X_n - X| < \varepsilon\right) \to 1$ for all $\varepsilon > 0$. The two definitions are equivalent.

Next, we show that convergence in $r$-th mean implies convergence in probability. The proof requires the following Lemma.

**Lemma 1.** ***(Markov's Inequality)*** *Let $X$ be a random variable. For $\varepsilon > 0$ and $r > 0$,*

$$\Pr\left(|X| \geq \varepsilon\right) \leq \mathbb{E}\left|X\right|^r / \varepsilon^r.$$

**Proof.** Let $f_X$ be the PDF of $X$ (the proof is similar for the discrete case). Let $1(\cdot)$ be an indicator function, i.e. it is equal one if the condition inside the parenthesis is satisfied, and zero otherwise. For example,

$$1\left(|x| \geq \varepsilon\right) = \begin{cases} 1, & |x| \geq \varepsilon, \\ 0, & |x| < \varepsilon. \end{cases}$$

Note that $1(|x| \geq \varepsilon) + 1(|x| < \varepsilon) = 1$. Next,

$$\begin{aligned}
\mathbb{E}|X|^r &= \mathbb{E}\left(|X|^r 1(|X| \geq \varepsilon)\right) + \mathbb{E}\left(|X|^r 1(|X| < \varepsilon)\right) \\
&\geq \mathbb{E}\left(|X|^r 1(|X| \geq \varepsilon)\right) \\
&\geq \varepsilon^r \mathbb{E}\left(1(|X| \geq \varepsilon)\right) \\
&= \varepsilon^r \int_{-\infty}^{\infty} 2(|x| \geq \varepsilon) f_X(x) dx \\
&= \varepsilon^r \left( \int_{-\infty}^{-\varepsilon} 1 \cdot f_X(x) dx + \int_{-\varepsilon}^{\varepsilon} 0 \cdot f_X(x) dx + \int_{\varepsilon}^{\infty} 1 \cdot f_X(x) dx \right) \\
&= \varepsilon^r \left( \int_{-\infty}^{-\varepsilon} f_X(x) dx + \int_{\varepsilon}^{\infty} f_X(x) dx \right) \\
&= \varepsilon^r \Pr(|X| \geq \varepsilon).
\end{aligned}$$

Now, suppose that $X_n$ converges to $X$ in $r$-th mean, $\mathbb{E}|X_n - X|^r \to 0$. Then,

$$\begin{aligned}
\Pr(|X_n - X| \geq \varepsilon) &\leq \mathbb{E}|X_n - X|^r / \varepsilon^r \\
&\to 0.
\end{aligned}$$

The following are some rules for manipulation of probability limits. Suppose that $X_n \to_p a$ and $Y_n \to_p b$, where $a$ and $b$ are some finite constants. Let $c$ be another constant. Then,

**(i)** $cX_n \to_p ca$.

**(ii)** $X_n + Y_n \to_p a + b$.

**(iii)** $X_n Y_n \to_p ab$.

**(iv)** $X_n / Y_n \to_p a/b$, provided that $b \neq 0$.

**Proof of (ii):**

$$\begin{aligned}
\Pr(|(X_n + Y_n) - (a + b)| \geq \varepsilon) &= \Pr(|(X_n - a) + (Y_n - b)| \geq \varepsilon) \\
&\leq \Pr(|X_n - a| + |Y_n - b| \geq \varepsilon) \\
&\leq \Pr(|X_n - a| \geq \varepsilon/2 \text{ or } |Y_n - b| \geq \varepsilon/2) \\
&\leq \Pr(|X_n - a| \geq \varepsilon/2) + \Pr(|Y_n - b| \geq \varepsilon/2) \\
&\to 0.
\end{aligned}$$

It is easy to show the following “Squeeze Rule”: If $0 \leq X_n \leq Y_n$ and $Y_n \to_p 0$, then $X_n \to_p 0$. It is also clear that $X_n \to_p 0$ if and only if $|X_n| \to_p 0$.

The following result shows that if a sequence of random variables converges in probability to a constant, then their continuous functions converge in probability as well.

**Theorem 2.** ***(Continuous Mapping Theorem (CMT) or Slutsky’s Lemma)*** *Suppose that $X_n \to_p c$, a constant, and let $h(\cdot)$ be a continuous function at $c$. Then, $h(X_n) \to_p h(c)$.*

**Proof:** By continuity of $h(\cdot)$, given $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that $|u - c| < \delta_\varepsilon$ implies that $|h(u) - h(c)| < \varepsilon$. Consequently, we have the following relation between the two events:

$$\{\omega : |h(X_n(\omega)) - h(c)| < \varepsilon\} \supset \{\omega : |X_n(\omega) - c| < \delta_\varepsilon\},$$

and, therefore,

$$\begin{aligned} \Pr(|h(X_n) - h(c)| < \varepsilon) &\geq \Pr(|X_n - c| < \delta_\varepsilon) \\ &\to 1. \end{aligned}$$

For example, suppose that $\widehat{\beta}_n \to_p \beta$. Then $\widehat{\beta}_n^2 \to_p \beta^2$, and $1/\widehat{\beta}_n \to_p 1/\beta$, provided $\beta \neq 0$.

The random vectors/matrices converge in probability if their elements converge in probability. Alternatively, one may consider convergence in probability of norms. Consider the vector case. Let $\{\boldsymbol{X}_n : n = 1, 2, \ldots\}$ be a sequence of random $k$-vectors. We will show that $\boldsymbol{X}_n - \boldsymbol{X} \to_p 0$ element-by-element, where $\boldsymbol{X}$ is a possibly random $k$-vector, if and only if $\|\boldsymbol{X}_n - \boldsymbol{X}\| \to_p 0$, where $\|\cdot\|$ denotes the Euclidean norm. First, suppose that for all $i = 1, \ldots k$ we have that $X_{n,i} - X_i \to_p 0$. Then,

$$\begin{aligned} \|\boldsymbol{X}_n - \boldsymbol{X}\| &= \sqrt{\sum_{j=1}^{k} (X_{n,j} - X_j)^2} \\ &\to_p 0, \end{aligned}$$

due to the CMT and property (ii) above. Next, suppose that $\|\boldsymbol{X}_n - \boldsymbol{X}\| \to_p 0$. By CMT, $\|\boldsymbol{X}_n - \boldsymbol{X}\|^2 \to_p 0$. Since $\|\boldsymbol{X}_n - \boldsymbol{X}\|^2 = \sum_{j=1}^{k} (X_{n,j} - X_j)^2$, and $(X_{n,j} - X_j)^2 \geq 0$ for all $n$ and $j = 1, \ldots, k$, by Squeeze Rule, $(X_{n,j} - X_j)^2 \to_p 0$. By CMT, $|X_{n,j} - X_j| \to_p 0$.

The rules for manipulation of probability limits in the vector/matrix case are similar to those in the scalar case, (i) - (iv) above with corresponding definitions of multiplication and division. The CMT is valid in vector/matrix case as well.

## Weak Law of Large Numbers (WLLN)

The WLLN is one of the most important examples of convergence in probability.

**Theorem 3.** ***(WLLN)*** *Let $X_1, \ldots X_n$ be a sample of iid random variables such that $\mathbb{E}|X_1| < \infty$. Then, $n^{-1}\sum_{i=1}^{n} X_i \to_p \mathbb{E}X_1$ as $n \to \infty$.*

Note that due to iid assumption, we have that $\mathbb{E}X_i = \mathbb{E}X_1$ for all $i = 1, \ldots, n$. We will prove the result assuming instead that $\mathbb{E}X_1^2 < \infty$, which implies that $\mathbb{E}|X_1| < \infty$, and $\mathrm{Var}(X_1) < \infty$.

**Theorem 4.** *Let $X_1, \ldots X_n$ be a sample of iid random variables such that $\mathrm{Var}(X_1) < \infty$. Then, $n^{-1}\sum_{i=1}^{n} X_i \to_p \mathbb{E}X_1$ as $n \to \infty$.*

**Proof:**

$$\Pr\left(\left|n^{-1}\sum_{i=1}^{n} X_i - \mathbb{E}X_1\right| \geq \varepsilon\right) = \Pr\left(\left|n^{-1}\sum_{i=1}^{n} (X_i - \mathbb{E}X_1)\right| \geq \varepsilon\right)$$

$$
\begin{aligned}
&\leq \frac{\mathbb{E}\left|\sum_{i=1}^{n}\left(X_i-\mathbb{E}X_1\right)\right|^2}{n^2\varepsilon^2} \\
&= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left(X_i-\mathbb{E}X_1\right)\left(X_j-\mathbb{E}X_1\right)}{n^2\varepsilon^2} \\
&= \frac{\sum_{i=1}^{n}\mathbb{E}\left(X_i-\mathbb{E}X_1\right)^2}{n^2\varepsilon^2} \\
&= \frac{n\mathrm{Var}\left(X_1\right)}{n^2\varepsilon^2} \\
&\to 0 \text{ as } n\to\infty.
\end{aligned}
$$

## Convergence in distribution

Convergence in distribution is another stochastic convergence concept used to approximate the distribution of a random variable $X_n$ in large samples. Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of random variables. Let $F_n(x)$ denote the marginal CDF of $X_n$, i.e. $F_n(x) = \Pr\left(X_n \leq x\right)$. Let $F(x)$ be another CDF. We say that $X_n$ *converges in distribution* if $F_n(x) \to F(x)$ for all $x$ where $F(x)$ is continuous. In this case, we write $X_n \to_d X$, where $X$ is *any* random variable with the distribution function $F(x)$. Note that while we say that $X_n$ converges to $X$, the convergence in distribution is not convergence of random variables, but of the distribution functions.

The extension to the vector case is straightforward. Let $\boldsymbol{X}_n$ and $\boldsymbol{X}$ be two random $k$-vectors. We say that $\boldsymbol{X}_n \to_d \boldsymbol{X}$ if the joint CDF of $\boldsymbol{X}_n$ converges to that of $\boldsymbol{X}$ at all continuity points, i.e.

$$
\begin{aligned}
F_n\left(x_1,\ldots,x_k\right) &= \Pr\left(X_{n,1}\leq x_1,\ldots,X_{n,k}\leq x_k\right) \\
&\to \Pr\left(X_1\leq x_1,\ldots,X_k\leq x_k\right) \\
&= F\left(x_1,\ldots,x_k\right),
\end{aligned}
$$

for all points $(x_1,\ldots,x_k)$ where $F$ is continuous. In this case, we say that the elements of $\boldsymbol{X}_n$, $X_{n,1},\ldots X_{n,k}$, *jointly* converge in distribution to $X_1,\ldots X_k$, the elements of $\boldsymbol{X}$.

The rules for manipulation of convergence in distribution results are as follows.

**(i) Cramer Convergence Theorem (Slutsky's Theorem):** Suppose that $X_n \to_d X$, and $Y_n \to_p c$. Then,

- **(a)** $X_n + Y_n \to_d X + c$.
- **(b)** $Y_n X_n \to_d cX$,
- **(c)** $X_n/Y_n \to_d X/c$, provided that $c \neq 0$.

Similar results hold in the vector/matrix case with proper definitions of multiplication and division.

**(ii)** If $X_n \to_p X$, then $X_n \to_d X$. Converse is not true with one exception:

**(iii)** If $X_n \to_d c$, a constant, then $X_n \to_p c$.

**(iv)** If $X_n - Y_n \to_p 0$, and $Y_n \to_d Y$, then $X_n \to_d Y$.

The following theorem extends convergence in distribution of random variables/vectors to convergence of their continuous functions.

**Theorem 5.** ***(Continuous Mapping Theorem (CMT))*** *Suppose that* $X_n \to_d X$*, and let* $h(\cdot)$ *be a function continuous on a set* $\mathcal{X}$ *such that* $\Pr(X \in \mathcal{X}) = 1$*. Then,* $h(X_n) \to_d h(X)$*.*

Examples:

- Suppose that $X_n \to_d X$. Then $X_n^2 \to_d X^2$. For example, if $X_n \to_d N(0,1)$, then $X_n^2 \to_d \chi_1^2$.
- Suppose that $(X_n, Y_n) \to_d (X, Y)$ (joint convergence in distribution), and set $h(x,y) = x$. Then $X_n \to_d X$. Set $h(x,y) = x^2+y^2$. Then $X_n^2+Y_n^2 \to_d X^2+Y^2$. For example, if $(X_n, Y_n) \to_d N(0, I_2)$ (bivariate standard normal distribution), then $X_n^2 + Y_n^2 \to_d \chi_2^2$.

Note that contrary to convergence in probability, $X_n \to_d X$ and $Y_n \to_d Y$ does not imply that, for example, $X_n + Y_n \to_d X + Y$, unless a joint convergence result holds. This is due to the fact that the individual convergence in distribution is convergence of the *marginal* CDFs. In order to characterize the limiting distribution of $X_n + Y_n$ one has to consider the limiting behavior of the *joint* CDF of $X_n$ and $Y_n$.

## The Central Limit Theorem (CLT)

Various versions of the CLT are used to establish convergence in distribution of re-scaled sums of random variables.

**Theorem 6.** ***(CLT)*** *Let* $X_1, \ldots, X_n$ *be a sample of iid random variables such that* $\mathbb{E}X_1 = 0$ *and* $0 < \mathbb{E}X_1^2 < \infty$*. Then, as* $n \to \infty$*,* $n^{-1/2}\sum_{i=1}^n X_i \to_d N(0, \mathbb{E}X_1^2)$*.*

For example, the CLT can be used to approximate the distribution of the average in large samples as follows. Let $X_1, \ldots X_n$ be a sample of iid random variables with $\mathbb{E}X_1 = \mu$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Define

$$\overline{X}_n = n^{-1}\sum_{i=1}^n X_i.$$

Consider $n^{-1/2}\sum_{i=1}^n (X_i - \mu)$. We have that $(X_1 - \mu), \ldots, (X_n - \mu)$ are iid with the mean $\mathbb{E}(X_1 - \mu) = 0$, and the variance $\mathbb{E}(X_1 - \mu)^2 = \sigma^2 < \infty$. Therefore, by the CLT,

$$\begin{aligned} n^{1/2}\left(\overline{X}_n - \mu\right) &= n^{-1/2}\sum_{i=1}^n (X_i - \mu) \\ &\to_d N\left(0, \sigma^2\right). \end{aligned}$$

In practice, we use convergence in distribution as an *approximation.* Let $\overset{a}{\sim}$ denote "approximately in large samples". Informally, one can say that $n^{1/2}\left(\overline{X}_n - \mu\right) \overset{a}{\sim} N\left(0, \sigma^2\right)$ or

$$\overline{X}_n \overset{a}{\sim} N\left(\mu, \sigma^2/n\right),$$

Note that under the normality assumption for $X_i$'s, the above result is obtained *exactly* for any sample size $n$.

The CLT can be extended to the vector case by the means of the following result.

**Lemma 7.** ***(Cramer-Wold device)*** *Let* $\boldsymbol{X}_n$ *be a random k-vector. Then,* $\boldsymbol{X}_n \to_d \boldsymbol{X}$ *if and only if* $\boldsymbol{\lambda}'\boldsymbol{X}_n \to_d \boldsymbol{\lambda}'\boldsymbol{X}$ *for all non-zero* $\boldsymbol{\lambda} \in \mathbb{R}^k$*.*

**Corollary 8.** *(**Multivariate CLT**) Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a sample of iid random $k$-vectors such that $\mathbb{E}\boldsymbol{X}_1 = \boldsymbol{0}$ (denote $\boldsymbol{X}_i = (X_{i,1}, ..., X_{i,k})'$) and $\mathbb{E}X_{1,j}^2 < \infty$ for all $j = 1, \ldots, k$, and $\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)$ is positive definite. Then, $n^{-1/2}\sum_{i=1}^n \boldsymbol{X}_i \to_d N(0, \mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right))$.*

**Proof:** Let $\boldsymbol{\lambda}$ be a $k$-vector of constants. Consider $Y_i = \boldsymbol{\lambda}'\boldsymbol{X}_i$. We have that $Y_1, \ldots, Y_n$ are iid. Further,

$$\begin{aligned}
\mathbb{E}Y_1 &= \boldsymbol{\lambda}'\mathbb{E}\boldsymbol{X}_1 \\
&= 0, \\
\operatorname{Var}(Y_1) &= \mathbb{E}Y_1^2 \\
&= \boldsymbol{\lambda}'\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\boldsymbol{\lambda}.
\end{aligned}$$

The variance of $Y_1$ is finite provided that all the elements of the variance-covariance matrix $\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)$ are finite. In order to show that, note that the $(r,s)$-th element of $\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)$ is given by $\mathbb{E}\left(X_{1,r}X_{1,s}\right)$. By the Cauchy-Schwartz inequality,

$$\mathbb{E}\left|X_{1,r}X_{1,s}\right| \leq \sqrt{\mathbb{E}X_{1,r}^2\mathbb{E}X_{1,s}^2},$$

which is finite for all $r = 1, \ldots, k$, $s = 1, \ldots, k$ due to the assumption that $\mathbb{E}X_{1,j}^2 < \infty$ for all $j = 1, \ldots, k$. Consequently, $\mathbb{E}Y_1^2 < \infty$, and it follows from the univariate CLT that

$$n^{-1/2}\sum_{i=1}^n Y_i \to_d N\left(0, \boldsymbol{\lambda}'\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\boldsymbol{\lambda}\right).$$

Let $\boldsymbol{W}$ be any $N\left(0, \mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\right)$ random vector. Since $\boldsymbol{\lambda}'\boldsymbol{W} \sim N\left(0, \boldsymbol{\lambda}'\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\boldsymbol{\lambda}\right)$, we have that

$$\begin{aligned}
\boldsymbol{\lambda}'\left(n^{-1/2}\sum_{i=1}^n \boldsymbol{X}_i\right) &= n^{-1/2}\sum_{i=1}^n Y_i \\
&\to_d \boldsymbol{\lambda}'\boldsymbol{W}.
\end{aligned}$$

Therefore, by the Cramer-Wold device we have that

$$n^{-1/2}\sum_{i=1}^n \boldsymbol{X}_i \to_d \boldsymbol{W} \sim N\left(0, \mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\right).$$

## Delta method

The delta method is used to derive the asymptotic distribution of the nonlinear functions of estimators. For example, in the case of iid random sample, we have that by the WLLN the average converges in probability to the expected value of an observation: $\overline{X}_n \to_p \mathbb{E}X_1 = \mu$. Further, it follows from the CMT that $h\left(\overline{X}_n\right) \to_p h(\mu)$. However, this does not allow us to approximate the distribution of $h\left(\overline{X}_n\right)$, since $h(\mu)$ is a constant (non-random). Note, that the CMT cannot be applied to general nonlinear $h\left(\overline{X}_n\right)$, since we have only a convergence in distribution result for $n^{1/2}\left(\overline{X}_n - \mu\right)$.

**Theorem 9.** *(**Delta method**) Let $\widehat{\boldsymbol{\theta}}_n$ be a random $k$-vector, and suppose that $n^{1/2}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \to_d \boldsymbol{Y}$ as $n \to \infty$, where $\boldsymbol{\theta}$ is a $k$-vector of constants ($\boldsymbol{\theta} = (\theta_1, ..., \theta_k)'$), and $\boldsymbol{Y}$ is a random $k$-vector. Let*

$\boldsymbol{h}:\mathbb{R}^k\to\mathbb{R}^m$ *be a function continuously differentiable on some open neighborhood of* $\boldsymbol{\theta}$. *Equivalently, we can denote* $\boldsymbol{h}=(h_1,...,h_m)'$, *where* $h_j:\mathbb{R}^k\to\mathbb{R}$, $j=1,...,m$. *Then,* $n^{1/2}\left(\boldsymbol{h}\left(\widehat{\boldsymbol{\theta}}_n\right)-\boldsymbol{h}(\boldsymbol{\theta})\right)\to_d \frac{\partial\boldsymbol{h}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}Y$, *where*

$$\frac{\partial\boldsymbol{h}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}=\begin{pmatrix}\frac{\partial h_1(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\\ \vdots\\ \frac{\partial h_m(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\end{pmatrix}=\begin{pmatrix}\frac{\partial h_1(\boldsymbol{\theta})}{\partial\theta_1} & \cdots & \frac{\partial h_1(\boldsymbol{\theta})}{\partial\theta_k}\\ \vdots & \ddots & \vdots\\ \frac{\partial h_m(\boldsymbol{\theta})}{\partial\theta_1} & \cdots & \frac{\partial h_m(\boldsymbol{\theta})}{\partial\theta_k}\end{pmatrix}.$$

**Proof:** First, note that $n^{1/2}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right)\to_d \boldsymbol{Y}$ implies that $\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\to_p \mathbf{0}$ or $\widehat{\boldsymbol{\theta}}_n\to_p\boldsymbol{\theta}$. Indeed, define $\tau_n=n^{-1/2}$. We have that $\tau_n\to 0$, and, consequently, $\tau_n\to_p 0$. By the Cramer Convergence Theorem (b),

$$\begin{aligned}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right) &= \tau_n n^{1/2}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right)\\ &\to_d \left(p\lim\tau_n\right)\boldsymbol{Y}\\ &= \mathbf{0}.\end{aligned}$$

Therefore, by property (iii) of convergence in distribution, $\widehat{\boldsymbol{\theta}}_n\to_p\boldsymbol{\theta}$.

Apply the mean value theorem to the function $\boldsymbol{h}\left(\widehat{\boldsymbol{\theta}}_n\right)$ element-by-element (see the Appendix) to obtain

$$\boldsymbol{h}\left(\widehat{\boldsymbol{\theta}}_n\right)=\boldsymbol{h}(\boldsymbol{\theta})+\frac{\partial\boldsymbol{h}(\boldsymbol{\theta}_n^*)}{\partial\boldsymbol{\theta}'}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right),\tag{1}$$

where $\boldsymbol{\theta}_n^*$ is a random variable that lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}$ (element-by-element), i.e. $\|\boldsymbol{\theta}_n^*-\boldsymbol{\theta}\|\le\left\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right\|$. Note that "$\boldsymbol{\theta}_n^*$" on different rows of the matrix $\frac{\partial\boldsymbol{h}(\boldsymbol{\theta}_n^*)}{\partial\boldsymbol{\theta}'}$ could be different. Since $\widehat{\boldsymbol{\theta}}_n\to_p\boldsymbol{\theta}$, it has to be that $\boldsymbol{\theta}_n^*\to_p\boldsymbol{\theta}$ as well:

$$\begin{aligned}\Pr\left(\|\boldsymbol{\theta}_n^*-\boldsymbol{\theta}\|\ge\varepsilon\right) &\le \Pr\left(\left\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right\|\ge\varepsilon\right)\\ &\to 0.\end{aligned}$$

Furthermore, by the CMT,

$$\frac{\partial\boldsymbol{h}(\boldsymbol{\theta}_n^*)}{\partial\boldsymbol{\theta}'}\to_p\frac{\partial\boldsymbol{h}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}.\tag{2}$$

Next, re-write (1) as follows:

$$n^{1/2}\left(\boldsymbol{h}\left(\widehat{\boldsymbol{\theta}}_n\right)-\boldsymbol{h}(\boldsymbol{\theta})\right)=\frac{\partial\boldsymbol{h}(\boldsymbol{\theta}_n^*)}{\partial\boldsymbol{\theta}'}n^{1/2}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right).$$

Then, it follows from the result in (2), assumption $n^{1/2}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right)\to_d\boldsymbol{Y}$ and Cramer Convergence Theorem (b) that

$$n^{1/2}\left(\boldsymbol{h}\left(\widehat{\boldsymbol{\theta}}_n\right)-\boldsymbol{h}(\boldsymbol{\theta})\right)\to_d\frac{\partial\boldsymbol{h}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\boldsymbol{Y}.$$

Consider again the example of the average of iid random variables with finite variance. We have that $n^{1/2}\left(\overline{X}_n-\mu\right)\to_d N\left(0,\sigma^2\right)$. Suppose that $\mu\ne 0$. Then, by the delta method,

$$n^{1/2}\left(\frac{1}{\overline{X}_n}-\frac{1}{\mu}\right)\quad\to_d\quad -\frac{1}{\mu^2}N\left(0,\sigma^2\right)$$

$$= \quad N\left(0, \frac{\sigma^2}{\mu^4}\right).$$

## Appendix A: Mean value theorem

**Theorem 10.** ***(One-Dimensional Mean-Value Theorem)*** *Let $f : [a,b] \to \mathbb{R}$ be continuous on $[a,b]$ and differentiable on $(a,b)$. Then there is $c \in (a,b)$ such that*

$$f(b) - f(a) = \frac{df(c)}{dx}(b-a).$$

Now, suppose we have $h : \Theta \to \mathbb{R}$, where $\Theta \subset \mathbb{R}^k$. Suppose further that $h$ is continuously differentiable on some open neighborhood of $\boldsymbol{\theta}_0$, say $N_0$, and let $\boldsymbol{u}$ be such that $\boldsymbol{\theta}_0 + t\boldsymbol{u} \in N_0$ for all $t \in [0,1]$. Define $f(t) = h(\boldsymbol{\theta}_0 + t\boldsymbol{u})$. The function $f$ is continuous and differentiable on $[0,1]$ interval, and by the one-dimensional mean-value theorem,

$$\begin{aligned} h(\boldsymbol{\theta}_0 + \boldsymbol{u}) - h(\boldsymbol{\theta}_0) &= f(1) - f(0) \\ &= \frac{df(t^*)}{dt} \text{ for some } t^* \in (0,1) \\ &= \frac{\partial h(\boldsymbol{\theta}_0 + t^*\boldsymbol{u})}{\partial \boldsymbol{\theta}'}\boldsymbol{u} \\ &= \frac{\partial h(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}\boldsymbol{u}, \end{aligned}$$

where

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_0 + \boldsymbol{t}^*\boldsymbol{u},$$

and

$$\begin{aligned} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\| &= \boldsymbol{t}^* \|\boldsymbol{u}\| \\ &< \|\boldsymbol{u}\|. \end{aligned}$$

(The argument follows closely that of Theorem 10 on page 106 of Magnus and Neudecker (2007): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd Edition.) We have established a mean-value theorem for real-valued functions of several variables:

**Theorem 11.** *Let $h : \Theta \to \mathbb{R}$, where $\Theta \subset \mathbb{R}^k$, be continuously differentiable on some open neighborhood $N_{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. If $\widehat{\boldsymbol{\theta}} \in N_{\boldsymbol{\theta}}$, then there is $\boldsymbol{\theta}^* \in N_{\boldsymbol{\theta}}$ such that*

$$h(\widehat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta}) = \frac{\partial h(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right),$$

*where $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\| \leq \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|$.*

If $\boldsymbol{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \ldots, h_m(\boldsymbol{\theta}))'$ is a vector valued function with $h_j : \Theta \to \mathbb{R}$ for all $j = 1, \ldots, m$, the above theorem can be applied element-by-element:

$$h(\widehat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta}) \quad = \quad \begin{pmatrix} h_1(\widehat{\boldsymbol{\theta}}) - h_1(\boldsymbol{\theta}) \\ \vdots \\ h_m(\widehat{\boldsymbol{\theta}}) - h_m(\boldsymbol{\theta}) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial \boldsymbol{h}(\boldsymbol{\theta}^{*,1})}{\partial \boldsymbol{\theta}'} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \\ \vdots \\ \frac{\partial \boldsymbol{h}(\boldsymbol{\theta}^{*,m})}{\partial \boldsymbol{\theta}'} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial \boldsymbol{h}(\boldsymbol{\theta}^{*,1})}{\partial \boldsymbol{\theta}'} \\ \vdots \\ \frac{\partial \boldsymbol{h}(\boldsymbol{\theta}^{*,m})}{\partial \boldsymbol{\theta}'} \end{pmatrix} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right),$$

where

$$\left\|\boldsymbol{\theta}^{*,j} - \boldsymbol{\theta}\right\| \leq \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|, \tag{3}$$

for all $j = 1, \ldots, m$. To simplify the notation, we can write

$$\begin{pmatrix} \frac{\partial \boldsymbol{h}(\boldsymbol{\theta}^{*,1})}{\partial \boldsymbol{\theta}'} \\ \vdots \\ \frac{\partial \boldsymbol{h}(\boldsymbol{\theta}^{*,m})}{\partial \boldsymbol{\theta}'} \end{pmatrix} = \frac{\partial \boldsymbol{h}\left(\boldsymbol{\theta}^{*}\right)}{\partial \boldsymbol{\theta}'},$$

indicating that $\boldsymbol{\theta}^*$ may be different across the rows of the matrix $\partial \boldsymbol{h}\left(\boldsymbol{\theta}^*\right)/\partial \boldsymbol{\theta}'$, and that, in each row, $\boldsymbol{\theta}^*$ satisfies (3).

## Appendix B: Proof of the CLT

The material discussed here is adopted from Hogg, McKean, and Craig (2005): *Introduction to Mathematical Statistics.* Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$, where $X_i$'s are iid with mean $\mu$ and variance $\sigma^2$. The moment generating function (MGF) of a $N(0,\sigma^2)$ distribution is given by $\exp(t^2\sigma^2/2)$. It suffices to show that the MGF of $n^{1/2}(\overline{X}_n - \mu)$ converges to $\exp(t^2\sigma^2/2)$.[1]

Let $m(t)$ denote the MGF of $X_1 - \mu$:

$$m(t) = \mathbb{E}\exp\left(t\left(X_1 - \mu\right)\right).$$

The MGF has the following properties:

$$\begin{aligned} m(0) &= 1, \\ m^{(1)}(0) &= \mathbb{E}(X_1 - \mu) = 0, \\ m^{(2)}(0) &= \mathbb{E}(X_1 - \mu)^2 = \sigma^2, \end{aligned}$$

where

$$m^{(s)}(0) = \left.\frac{d^s m(t)}{dt^s}\right|_{t=0}.$$

We have the following expansion of $m(t)$:

$$m(t) = m(0) + m^{(1)}(0)t + \frac{m^{(2)}(s)t^2}{2}$$

[1] If the MGF does not exist, one can replace it with the *characteristic function* of $X_1 - \mu$, which is defined as $\varphi(t) = \mathbb{E}\exp(it(X_1 - \mu))$, where $i = \sqrt{-1}$. Note that the characteristic function always exists, and the proof with the characteristic function is essentially the same as the proof that uses the MGF.

$$= 1 + \frac{m^{(2)}(s)t^2}{2}. \tag{4}$$

where $s$ is a mean value that lies between 0 and $t$.

Let $M_n(t)$ denote the MGF of $n^{1/2}(\overline{X}_n - \mu)$. We have:

$$\begin{aligned}
M_n(t) &= \mathbb{E} \exp\left( t \frac{1}{n^{1/2}} \sum_{i=1}^{n} (X_i - \mu) \right) \\
&= \mathbb{E} \prod_{i=1}^{n} \exp\left( \frac{t}{n^{1/2}} (X_i - \mu) \right) \\
&= \prod_{i=1}^{n} \mathbb{E} \exp\left( \frac{t}{n^{1/2}} (X_i - \mu) \right) \text{ (by independence)} \\
&= \left( \mathbb{E} \exp\left( \frac{t}{n^{1/2}} (X_1 - \mu) \right) \right)^n \text{ (because of "identical distributed")} \\
&= \left( m\left( \frac{t}{n^{1/2}} \right) \right)^n \text{ (by definition of } m(t)) \\
&= \left( 1 + \frac{m^{(2)}(s) \left( \frac{t}{n^{1/2}} \right)^2}{2} \right)^n \text{ (by (4))} \\
&= \left( 1 + \frac{m^{(2)}(s)\, t^2}{2n} \right)^n \\
&= \left( 1 + \frac{a_n}{n} \right)^n,
\end{aligned}$$

where $s$ lies between 0 and $t/n^{1/2}$ and therefore converges to zero as $n \to \infty$, and

$$\begin{aligned}
a_n &= \frac{m^{(2)}(s)\, t^2}{2} \\
&\to \frac{m^{(2)}(0) t^2}{2} \text{ (as } n \to \infty) \\
&= \frac{\sigma^2 t^2}{2}.
\end{aligned}$$

We will show next that

$$\log M_n(t) = n \log\left(1 + \frac{a_n}{n}\right) \to \lim_{n\to\infty} a_n = \frac{\sigma^2 t^2}{2}. \tag{5}$$

Note that the result in (5) implies that

$$M_n(t) = \exp\left(\log M_n(t)\right) \to \exp\left( \lim_{n\to\infty} \log M_n(t) \right) = \exp\left( \frac{\sigma^2 t^2}{2} \right).$$

To show (5), write

$$\begin{aligned}
\lim_{n\to\infty} n \log\left(1 + \frac{a_n}{n}\right) &= \lim_{n\to\infty} \frac{\log\left(1 + a_n/n\right)}{1/n} \\
&= \lim_{n\to\infty} a_n \frac{\log\left(1 + a_n/n\right)}{a_n/n}
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} a_n \lim_{\delta\to 0} \frac{\log(1+\delta)}{\delta} \text{ (by change of variable } \delta = a_n/n\text{)}\\
&= \frac{\sigma^2 t^2}{2} \lim_{\delta\to 0} \frac{\log(1+\delta)}{\delta}.
\end{aligned}
$$

Lastly, by l'Hôpital's rule,

$$\lim_{\delta\to 0} \frac{\log(1+\delta)}{\delta} = \lim_{\delta\to 0} \frac{1/(1+\delta)}{1} = 1.$$