# Lecture 1: Properties of A Random Sample

Econ 205A: Econometric Methods I

Ruoyao Shi

Fall 2018

## Contents



## 1 Basic Concepts of a Random Sample

### Random Sample and Statistic

- The random variables $X_1, \dots, X_n$ are called a **random sample** of size $n$ from the population $f(x)$ if $X_1, \dots, X_n$ are:
  - mutually independent random variables, and
  - the marginal pdf or pmf of each $X_i$ is the same function $f(x)$.
- We implicitly assume that the size of the population is much larger than the sample. Even though actual sampling is often without replacement, the impact on $f(x)$ is negligible (like with replacement).

- Sometimes a random sample is denoted as $\{X_i\}_{i=1}^n$.
- They are also called **independent and identically distributed** random variables with pdf or pmf $f(x)$, denoted as **i.i.d.** random variables.
- Let $X_1, \ldots, X_n$ be a random sample of size $n$ from a population, and let $T(x_1, \ldots, x_n)$ be a real-valued (or vector-valued) function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable (or random vector) $Y \equiv T(X_1, \ldots, X_n)$ is called a **statistic**.
  - **Example 1** *Sample mean* $\bar{X} \equiv n^{-1} \sum_{i=1}^n X_i$, *sample variance* $S^2 \equiv (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ *and order statistic* $X_{(1)} \equiv \min_{1 \le i \le n} X_i$ *are all statistics. We often use* $\bar{X}_n$ *and* $S_n^2$ *to emphasize the sample size.*
  - Population mean $\mu$, population variance $\sigma^2$ and population median are not statistics. They are parameters.
- The probability distribution of a statistic $Y$ is called the **sampling distribution** of $Y$.
- Statistics are functions of and only of the sample, not of the parameters. Hence a statistic is a random variable. The sampling distributions of statistics generally depend on parameters.
- Why are we interested in sampling distribution? Because they form the basis of statistical inference and hypothesis testing.

# 2 Sample Mean and Sample Variance

### Sample Mean and Sample Variance

- Now let's prove some results that we knew from undergraduate econometrics class (but probably didn't know why).
- **Theorem 1** *Let* $X_1, \ldots, X_n$ *be a random sample from a population with mean* $\mu$ *and variance* $\sigma^2 < \infty$*. Then:*

  *(a)* $\mathbb{E}(\bar{X}) = \mu$;

  *(b)* $var(\bar{X}) = \sigma^2/n$;

*(c)* $\mathbb{E}(S^2) = \sigma^2$.

- *Proof.* To prove (a),

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(n^{-1}\sum_{i=1}^{n} X_i\right) = n^{-1}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{n}{n}\mathbb{E}(X_1) = \mu.$$

To prove (b),

$$\begin{aligned} var(\bar{X}) &= var\left(n^{-1}\sum_{i=1}^{n} X_i\right) = n^{-2} var\left(\sum_{i=1}^{n} X_i\right) \\ &= n^{-2}\sum_{i=1}^{n} var(X_i) = \frac{n}{n^2} var(X_1) = \frac{\sigma^2}{n}. \end{aligned}$$

To prove (c),

$$\begin{aligned} \mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\ &= \frac{1}{n-1}\left(n\mathbb{E}(X_1^2) - n\mathbb{E}(\bar{X}^2)\right) = \frac{n}{n-1}\left((\mu^2 + \sigma^2) - (\mu^2 + \frac{\sigma^2}{n})\right) = \sigma^2. \end{aligned}$$

- **Lemma 1** ***(Jensen's) Inequality*** *If $X$ is a (non-constant with positive probability) random variable and $\varphi(\cdot)$ is a convex function, then $\varphi(\mathbb{E}(X)) < \mathbb{E}(\varphi(X))$. The direction of the inequality is reversed for concave function.*

- Only linear functions $\varphi(x) \equiv a + bx$ ($a, b$ are constants) are both concave and convex, then $\varphi(\mathbb{E}(X)) = a + b\mathbb{E}(X) = \mathbb{E}(a + bX) = \mathbb{E}(\varphi(X))$.

- If $X$ is a constant (with probability one) and $\varphi(\cdot)$ is a convex function, then $\varphi(\mathbb{E}(X)) = \mathbb{E}(\varphi(X))$; similarly, if $X$ is a constant (with probability one) and $\varphi(\cdot)$ is a concave function, then $\varphi(\mathbb{E}(X)) = \mathbb{E}(\varphi(X))$.

- Using the Jensen's inequality, it is easy to show that $\mathbb{E}(S) \leq \sigma$.

# 3 Sampling from Normal Distribution

### Sampling Distribution of Sample Mean

- **Lemma 2** *Let $X_1, \dots, X_n$ be a random sample from a population with mgf $M_X(t)$, then the mgf of the sample mean is*

$$M_{\bar{X}}(t) = [M_X(t/n)]^n.$$

- Then throughout this section, we assume that the random sample $\{X_i\}_{i=1}^n$ is drawn from a $\mathcal{N}(\mu, \sigma^2)$ population, and $\sigma^2 < \infty$. In this section, we treat the sample size $n$ as a fixed number. The properties discussed in this section are called **finite sample properties**.

- **Theorem 2** *$\bar{X}_n$ has a $\mathcal{N}(\mu, \sigma^2/n)$ distribution.*

- *Proof.* Recall that if two distributions have the same mgf, then their cdf must be the same at almost all points in the support. Also recall that the mgf of a $\mathcal{N}(\mu, \sigma^2)$ random variable is $M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$, then by Lemma 2, the mgf of the sample mean is

$$\begin{aligned} M_{\bar{X}}(t) &= \left[\exp\left(\mu\frac{t}{n} + \frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2\right)\right]^n = \exp\left[n\left(\mu\frac{t}{n} + \frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2\right)\right] \\ &= \exp\left(\mu t + \left(\frac{1}{2}\frac{\sigma^2}{n}\right)t^2\right). \end{aligned}$$

- If the mgf exists, then it is unique for any given random variable. If two random variables have the same mgf, then their distributions (cdf's) must be the same. So if you want to show a random variable follows a particular distribution, one way is to show that the mgf of the random variable is the same as that particular distribution.

### Sampling Distribution of Sample Variance

- **Theorem 3** *$\bar{X}$ and $S^2$ are independent random variables.*

- *Proof.* Recall that a linear combination of normal random variables is still a normal random variable. Hence both $\bar{X}$ and $X_i - \bar{X}$ $(i = 1, \ldots, n)$ are normal. There covariance is

$$\begin{aligned} cov(\bar{X}, X_i - \bar{X}) &= cov(\bar{X}, X_i) - cov(\bar{X}, \bar{X}) \\ &= cov\left(\frac{X_1 + \cdots + X_n}{n}, X_i\right) - var(\bar{X}) \\ &= 0 + \cdots + \frac{\sigma^2}{n} + \cdots + 0 - \frac{\sigma^2}{n} \\ &= 0. \end{aligned}$$

  Moreover, recall that uncorrelated normal random variables are independent, so we have proven that $\bar{X}$ and $X_i - \bar{X}$ $(i = 1, \ldots, n)$ are independent. Finally, since $S^2$ is a function of $X_i - \bar{X}$ $(i = 1, \ldots, n)$, then $\bar{X}$ and $S^2$ are independent.

- **Theorem 4** *$(n-1)S^2/\sigma^2$ has chi-squared distribution with $n-1$ degrees of freedom.*

- *Proof.* Recall that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. As a result, $\frac{(\bar{X}-\mu)^2}{\sigma^2/n} \sim \chi_1^2$. Similarly, since $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$ and $\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} \sim \chi_n^2$. Note that

$$\begin{aligned} \underbrace{\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2}}_{A} &= \frac{\sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\sum_{i=1}^n (X_i - \bar{X})}{\sigma^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2}{\sigma^2} \\ &= \underbrace{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}}_{B} + \underbrace{\frac{(\bar{X} - \mu)^2}{\sigma^2/n}}_{C}. \end{aligned}$$

  We have shown above that $A \sim \chi_n^2$ and $C \sim \chi_1^2$, so their mgf are $M_A(t) = (1-2t)^{-n/2}$ and $M_C(t) = (1-2t)^{-1/2}$. Moreover, note that $B$ is a function of $X_i - \bar{X}$ $(i = 1, \ldots, n)$ and $C$ is a function of $\bar{X}$. In the proof of Theorem 3 we showed that $\bar{X}$ and $X_i - \bar{X}$ are independent, so $B$ and $C$ are independent. This implies that

$$M_B(t) = M_A(t)/M_C(t) = (1-2t)^{-(n-1)/2},$$

which is the mgf of a chi-squared distribution with df $n-1$. Note that $B = (n-1)S^2/\sigma^2$ and this completes the proof.

**Sampling Distribution of t and F Statistics**

- The sampling distributions of sample mean $\bar{X}$ and the sample variance $S^2$ enable us to prove some familiar results from undergraduate econometrics class.

- **Definition 1** *If $W \sim \mathcal{N}(0,1)$, $U \sim \chi^2_r$, and $W$ and $U$ are independent, then*

$$T \equiv \frac{W}{\sqrt{U/r}}$$

  *has a* ***(Student) t distribution*** *with $r$ degrees of freedom.*

- **Definition 2** *If $U \sim \chi^2_{r_1}$ and $V \sim \chi^2_{r_2}$ and they are independent, then*

$$F \equiv \frac{U/r_1}{V/r_2}$$

  *has an* ***F distribution*** *with $(r_1, r_2)$ degrees of freedom .*

- Proof of the following theorems are left as exercises.

- **Theorem 5** ***(t statistic)*** *The random variable*

$$T \equiv \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

  *has a Student t distribution with $n-1$ degrees of freedom.*

- **Theorem 6** ***(F statistic)*** *We have two random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. Suppose their population variances are $\sigma^2_X$ and $\sigma^2_Y$, respectively. And let $S^2_n$ and $S^2_m$ be their respective sample variances. Then the random variable*

$$F \equiv \frac{\sigma^2_Y S^2_n}{\sigma^2_X S^2_m}$$

  *has an F distribution with $(n-1, m-1)$ degrees of freedom.*

# 4 Sampling Distributions Without Normality

### Why Do We Need Asymptotic Distributions

- In the previous section, we learned the sampling distributions of certain statistics under the assumption that the random sample is drawn from normally distributed population.
- Those results hold regardless of the sample size $n$. They are called **finite sample sampling distributions**.
- If the population does not have normal distribution, then in general the statistics computed from finite samples do not have simple known distributions.
- In such general cases, we need to invoke **asymptotic (sampling) distributions**, which are the *limits* of (unknown) finite sample sampling distributions of the statistics as the sample size $n$ approaches infinity.
- Asymptotic distributions are approximations of finite sample distributions.
- In order to study the asymptotic distributions, we need to take a not so short digression on different convergence concepts.
- Asymptotic properties (mean, variance, sampling distribution, etc.) are often referred to as **large sample properties** as well.

### Notation and Definitions

- Let $\{X_n\}$ for $n = 1, 2, 3, \ldots$ denote a sequence of random variables on a common probability space $(S, \mathcal{F}, P)$ and let $X$ denote another random variable defined on the same probability space.
- Let $F_n$ denote the distribution function (cdf) of $X_n$ and let $F$ denote the distribution function (cdf) of $X$.

- **Definition 3** *Let $\{a_n\}$ be a sequence of non-stochastic real numbers. If there exists a real number $a$ such that for any $\epsilon > 0$, there exists a finite integer $N(\epsilon)$ such that for all $n \geq N(\epsilon)$, we have $|a_n - a| < \epsilon$, then $a$ is called the **limit** of the sequence $\{a_n\}$. We write $a_n \to a$ as $n \to \infty$, or $\lim_{n\to\infty} a_n = a$.*

- **Definition 4** *A function $g(\cdot)$ is **continuous at** $a$ if and only if for any $\epsilon > 0$, there exists a positive real number $\delta(\epsilon)$ such that $|a_n - a| < \delta(\epsilon)$ implies $|g(a_n) - g(a)| < \epsilon$.*

- Equivalently, the function $g(\cdot)$ is continuous at $a$ if and only if for any sequence $\{a_n\}$, $a_n \to a$ as $n \to \infty$ implies $g(a_n) \to g(a)$ as $n \to \infty$.

- **Definition 5** *A function $g(\cdot)$ is called a **continuous function** if it is continuous at any value in its domain.*

**Convergence in Probability**

- **Definition 6** *A sequence of random variable $X_1, X_2, \ldots$ **converges in probability** to a random variable $X$ if, for any $\epsilon > 0$,*

  $$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0, \text{ or equivalently, } \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

  *This is denoted as $X_n \xrightarrow{p.} X$, or $\text{plim}_{n\to\infty} X_n = X$. We call $X$ to be the **probability limit** of $X_n$.*

- Convergence in probability implies that the "tail probability", i.e. probability of extreme events vanishes.

- If $X$ is a degenerated random variable, i.e. it is a constant (we denote it by $\mu$), then we say that the sequence of random variables $X_1, X_2, \ldots$ converges in probability to a constant $\mu$. This is denoted as $X_n \xrightarrow{p.} \mu$ or $\text{plim}_{n\to\infty} X_n = \mu$.

- **Lemma 3** ***(Chebyshev's inequality)*** *Let $X$ be a random variable with finite mean $\mu$ and finite variance $\sigma^2$. Then for any $\epsilon > 0$, we have*

  $$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

- **Theorem 7** ***(WLLN)*** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu$ and variance $\sigma^2 < \infty$. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0;$$

  *that is, $\bar{X}_n$ converges in probability to $\mu$.*

- *Proof.* The proof is a straightforward application of Chebyshev's inequality. Recall that $\mathbb{E}(\bar{X}_n) = \mu$ and $var(\bar{X}_n) = \sigma^2/n$. We have, for any $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty.$$

  Since probability is always non-negative, we have $\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$ by the squeeze theorem.

- Weak Law of Large Numbers (WLLN) states that under quite mild conditions, the sample mean approaches the population mean as $n \to \infty$.

- There is a more general version of WLLN, where we only need to assume that $\mu$ is finite.

- **Definition 7** *Let $X_1, \ldots, X_n$ be a random sample from a population $f(x; \theta)$, where $\theta$ is the parameter of the pdf (or pmf). Let $T_n$ be a statistic computed using the sample of size $n$. $T_n$ is called a* ***consistent estimator*** *of $\theta$ if $T_n \xrightarrow{p.} \theta$.*

- WLLN states that under certain conditions, sample mean is a consistent estimator of population mean.

- In the above, we used Chebyshev's inequality to prove WLLN. This argument suggests a general approach of showing that a statistic $T_n$ is a consistent estimator of parameter $\theta$. We only need three steps:

  (i) Show that $\mathbb{E}(T_n) = \theta$;

  (ii) Show that $var(T_n) \to 0$ as $n \to \infty$;

  (iii) Invoke Chebyshev's inequality to show that for any $\epsilon > 0$,

$$P(|T_n - \theta| \geq \epsilon) \leq \frac{var(T_n)}{\epsilon^2} \to 0 \text{ as } n \to \infty.$$

- **Example 2** ***(Consistency of*** $S_n^2$***)*** *Now we follow this approach to show the consistency of sample variance. Suppose* $X_1, X_2, \dots$ *is a sequence of i.i.d. random variable with mean* $\mu$ *and variance* $\sigma^2 < \infty$*. By Theorem 1 (c), we have* $\mathbb{E}(S_n^2) = \sigma^2$*. Then as long as* $var(S_n^2) \to 0$ *as* $n \to \infty$*, by Chebyshev's inequality we have*

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{var(S_n^2)}{\epsilon^2} \to 0 \text{ as } n \to \infty.$$

- **Theorem 8** *Suppose that* $X_1, X_2, \dots$ *converges in probability to a random variable* $X$ *and* $g(\cdot)$ *is a continuous function. Then* $g(X_1), g(X_2), \dots$ *converges in probability to* $g(X)$*.*

- This theorem is a special case of the **continuous mapping theorem (CMT)**.

- **Example 3** *Suppose* $X_n \xrightarrow{p.} X$ *and* $a$ *is a constant. Then* $aX_n \xrightarrow{p.} aX$*.*

- **Example 4** *Suppose* $X_n \xrightarrow{p.} c$ *(*$c$ *is a constant) and the function* $g(\cdot)$ *is continuous at* $c$*. Then* $g(X_n) \xrightarrow{p.} g(c)$*.*

- **Theorem 9** *Suppose* $X_n \xrightarrow{p.} X$ *and* $Y_n \xrightarrow{p.} Y$*. Then* $X_n + Y_n \xrightarrow{p.} X + Y$*.*

- **Theorem 10** *Suppose* $X_n \xrightarrow{p.} X$ *and* $Y_n \xrightarrow{p.} Y$*. Then* $X_nY_n \xrightarrow{p.} XY$*.*

- **Example 5** *If* $X_1, X_2, \dots$ *are i.i.d. uniform* $(0, \theta)$ *where* $\theta$ *is unknown, and let* $X_{(n)} \equiv \max_{1 \leq i \leq n} X_i$*. Then* $X_{(n)}$ *is consistent for* $\theta$*. To see this,*

$$\begin{aligned}
P(|X_{(n)} - \theta| \geq \epsilon) &= P(X_{(n)} \geq \theta + \epsilon) + P(X_{(n)} \leq \theta - \epsilon) \\
&= 0 + P(X_{(n)} \leq \theta - \epsilon) \\
&= P\left(\max_{1 \leq i \leq n} X_i \leq \theta - \epsilon\right) \\
&= P(X_1 \leq \theta - \epsilon, \dots, X_n \leq \theta - \epsilon) \\
&= \Pi_{i=1}^n P(X_i \leq \theta - \epsilon) \\
&= \left(1 - \frac{\epsilon}{\theta}\right)^n \\
&\to 0 \text{ as } n \to \infty \text{ for any } 0 < \epsilon \leq \theta.
\end{aligned}$$

*And it equals zero for* $\epsilon > \theta$*.*

**Convergence in r-th Mean**

- **Definition 8** *Let $X_1, X_2, \ldots$ be a sequence of random variables. If there exists another random variable $X$ such that $\mathbb{E}(|X_n - X|^r) \to 0$ as $n \to \infty$ for some real number $r \geq 1$. Then we say $X_n$* ***converges in the r-th mean*** *to $X$, denoted as $X_n \xrightarrow{L^r} X$.*

- Two special cases: when $r = 1$, we say $X_n$ converges in mean to $X$ (mean convergence); when $r = 2$, we say $X_n$ converges in mean square to $X$ (mean square convergence). Mean square convergens can be denoted as $\xrightarrow{L^2}$ or $\xrightarrow{m.s.}$.

- **Theorem 11** *If $X_n \xrightarrow{L^r} X$, then $X_n \xrightarrow{p.} X$.*

- *Proof.* By the Markov's inequality, we have

$$P(|X_n - X| \geq \epsilon) \leq \frac{\mathbb{E}(|X_n - X|^r)}{\epsilon^r} \to 0$$

  as $n \to \infty$, and the result follows.

- The mean squared error of an estimator can be decomposed into a term related to the variance and a term related to the bias (squared).

$$\begin{aligned}
\mathbb{E}[(X_n - X)^2] =& \mathbb{E}[(X_n - \mathbb{E}(X_n) + \mathbb{E}(X_n) - X)^2] \\
=& \mathbb{E}[(X_n - \mathbb{E}(X_n))^2] + \mathbb{E}[(\mathbb{E}(X_n) - X)^2] \\
&+ \underbrace{2\mathbb{E}[(X_n - \mathbb{E}(X_n))(\mathbb{E}(X_n) - X)]}_{= 0 \text{ by iterated law of large numbers}} \\
=& \underbrace{\mathbb{E}[(X_n - \mathbb{E}(X_n))^2]}_{\text{variance}} + \mathbb{E}[(\underbrace{\mathbb{E}(X_n) - X}_{\text{bias}})^2].
\end{aligned}$$

- This decomposition implies that a consistent estimator does not have to be unbiased, as long as its bias vanishes as the sample size increases. So the general approach of showing that a statistic $T_n$ is a consistent estimator of parameter $\theta$ can be modified to:

  (i) Show that $\mathbb{E}(T_n) = \theta + g(n)$, where $g(\cdot)$ is a function of the sample size and $g(n) \to 0$ as $n \to \infty$;

  (ii) Show that $var(T_n) \to 0$ as $n \to \infty$;

(iii) Invoke Markov's inequality to show that for any $\epsilon > 0$,

$$P(|T_n - \theta| \geq \epsilon) \leq \frac{\mathbb{E}[(T_n - \theta)^2]}{\epsilon^2} = \frac{var(T_n) + [g(n)]^2}{\epsilon^2} \to 0 \text{ as } n \to \infty.$$

**Almost Sure Convergence**

- **Definition 9** *A sequence of random variables $X_1, X_2, \ldots$* ***converges almost surely*** *to a random variable $X$ if, for any $\epsilon > 0$,*

$$P\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1, \textit{ or equivalently, } P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

  *This is denoted as $X_n \xrightarrow{a.s.} X$.*

- Almost sure convergence is also known as **convergence with probability 1**.

- Recall that a random variable is a function on the sample space, i.e., $X_i$: $S \to \mathbb{R}$. For random variables $X_1, X_2, \ldots$, every point $s \in S$ generate a particular sequence of real numbers $x_1, x_2, \ldots$. Almost sure convergence is similar to pointwise convergence of a sequence of functions, with one exception: the convergence is allowed to fail on a zero probability subset of $S$.

- Almost sure convergence is stronger than convergence in probability, in the sense that $X_n \overset{a.s.}{\to} X$ implies $X_n \overset{p.}{\to} X$, but not vice versa. However, if a sequence converges in probability, it is possible to find a subsequence that converges almost surely.

- Examples of sequence that $\overset{p.}{\to}$ but not $\overset{a.s.}{\to}$ will be given in TA sessions.

- **Theorem 12** ***(SLLN)*** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu$ and variance $\sigma^2 < \infty$. Then for any $\epsilon > 0$,*

$$P\left(\lim_{n\to\infty} |\bar{X} - \mu| < \epsilon\right) = 1;$$

  *that is, $\bar{X}$ converges almost surely to $\mu$.*

- Continuous mapping theorem (i.e. Theorems 8, 9, 10) also applies to almost sure convergence.

**Convergence in Distribution**

- **Definition 10** *A sequence of random variables* $X_1, X_2, \ldots$ ***converges in distribution*** *to a random variable* $X$ *if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

*at all points* $x$ *where* $F_X(x)$ *is continuous. This is denoted as* $X_n \xrightarrow{d.} X$.

- **Example 6** *(Continuation of Example 5) Previously we have shown that for any* $0 < \epsilon \leq \theta$,

$$P(X_{(n)} \leq \theta - \epsilon) = \left(1 - \frac{\epsilon}{\theta}\right)^n.$$

*If we take* $\epsilon = t/n$ *(note that for fixed* $t$*,* $t/n < \theta$ *for large enough* $n$*), then we have*

$$P\left(X_{(n)} \leq \theta\left(1 - \frac{t}{n\theta}\right)\right) = \left(1 - \frac{t}{n\theta}\right)^n \to e^{-t/\theta},$$

*which, upon rearranging, yields*

$$P\left(n\left(\theta - X_{(n)}\right) \leq t\right) \to 1 - e^{-t/\theta};$$

*that is, the random variable* $n(\theta - X_{(n)})$ *converges in distribution to an exponential(*$\theta$*) random variable. We say that the asymptotic distribution of the estimator* $X_{(n)}$ *is exponential(*$\theta$*), and its rate of convergence is* $n$.

- Although we talk of a sequence of random variables converging in distribution, it is really the *cdf*s that converge, not the random variables themselves. In this fundamental way convergence in distribution is quite different from convergence in probability or almost sure convergence.

- **Theorem 13** *If the sequence of random variables* $X_1, X_2, \ldots$ *converges in probability to a random variable* $X$*, the sequence also converges in distribution to* $X$.

- **Theorem 14** *The sequence of random variables $X_1, X_2, \ldots$ converges in probability to a constant $c$ if and only if the sequence also converges in distribution to $c$; that is, the statement*

$$P(|X_n - c| > \epsilon) \to 0 \text{ for any } \epsilon > 0$$

*is equivalent to*

$$P(X_n \leq x) \to \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x \geq c \end{cases}$$

- The asymptotic distribution of sample mean is summarized in the most important theorems in statistics, the Central Limit Theorems (CLT). The following is the simplest version of CLT (there are many others).

- **Theorem 15** *Let $X_1, X_2, \ldots,$ be a sequence of i.i.d. random variables whose mgfs exist in a neighborhood of 0 (that is, $M_{X_i}(t)$ exists for $|t| < h$ for some positive $h$). Let $\mathbb{E}(X_i) = \mu$ and $var(X_i) = \sigma^2 > 0$. (Both $\mu$ and $\sigma^2$ are finite since the mgf exists.) Define $\bar{X}_n \equiv n^{-1}\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(X_n - \mu)/\sigma$. Then for any $x$ such that $-\infty < x < \infty$,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy;$$

*that is, the asymptotic distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is standard normal.*

- An equivalent way of stating the result of this CLT is: define

$$Z_n \equiv \frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{var(\bar{X}_n)}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d.} Z \sim \mathcal{N}(0, 1).$$

- To prove this CLT, we need one important property of mgf: if two distributions have the same mgf, then they are identical at almost all points. Equivalently, if for all values of $t$ in a neighborhood of 0,

$$M_X(t) = M_Y(t),$$

then

$$F_X(x) = F_Y(x)$$

for all values of $x$; that is, $X$ and $Y$ have the same distribution.

- *Proof.* We will show that for $|t| < h$, the mgf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges to $e^{t^2/2}$, the mgf of a $\mathcal{N}(0,1)$ random variable. Define $Y_i \equiv (X_i - \mu)/\sigma$, and let $M_Y(t)$ the common mgf of $Y_i$s, which exists for $|t| < \sigma h$. Note that $Y_i$ has mean 0 and variance 1, then by the property of mgf, we have

$$M_Y^{(1)}(0) \equiv \frac{d}{dt} M_Y(t)|_{t=0} = 0,$$
$$M_Y^{(2)}(0) \equiv \frac{d^2}{dt^2} M_Y(t)|_{t=0} = 1.$$

Since

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i,$$

we have, by Lemmas 2 and 4,

$$\begin{aligned} M_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) &= M_{\sum_{i=1}^n Y_i/\sqrt{n}}(t) \\ &= M_{\sum_{i=1}^n Y_i}(t/\sqrt{n}) \\ &= \left[M_Y(t/\sqrt{n})\right]^n. \end{aligned}$$

Expanding $M_Y(t/\sqrt{n})$ in a Taylor series around 0 gives

$$\begin{aligned} M_Y(t/\sqrt{n}) &= 1 + M_Y^{(1)}(0)\frac{t/\sqrt{n}}{1!} + M_Y^{(2)}(0)\frac{(t/\sqrt{n})^2}{2!} + R_Y(t/\sqrt{n}) \\ &= 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y(t/\sqrt{n}) \end{aligned}$$

By the definition of the Taylor remainder term, it is easy to see that for fixed $t$, $\lim_{n\to\infty} R_Y(t/\sqrt{n}) \cdot (\sqrt{n})^2 = 0$. As a result,

$$\lim_{n\to\infty} \left[M_Y(t/\sqrt{n})\right]^n = \lim_{n\to\infty} \left[1 + \frac{(t/\sqrt{n})^2}{2!}\right]^n = \lim_{n\to\infty} \left[1 + \frac{t^2}{2n}\right]^n = e^{t^2/2}.$$

This completes the proof.

- **Lemma 4** *For any constants $a$ and $b$, the mgf of the random variable $aX + b$ is given by*

$$M_{aX+b}(t) = e^{bt} M_X(at).$$

- Even though we prove the previous CLT using mgf, the CLT is valid in much more generality. In particular, all of the assumptions about mgfs are not needed, similar proof goes through with characteristic functions.

- **Theorem 16** ***(Lindeberg-Levy CLT)*** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean $\mu$ and variance $0 < \sigma^2 < \infty$. Define $\bar{X}_n \equiv n^{-1} \sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any $x$, $-\infty < x < \infty$,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy;$$

  *that is, the asymptotic distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is standard normal.*

- The rate of convergence given by the CLT is $\sqrt{n}$ (root-n), the usual parametric rate.

- Continuous mapping theorem (Theorem 8 only) also applies to convergence in distribution.

- **Example 7** *If $Z_n \xrightarrow{d.} Z \sim \mathcal{N}(0,1)$, then the continuous mapping theorem implies that $Z_n^2 \xrightarrow{d.} Z^2 \sim \chi_1^2$.*

- However, $X_n \xrightarrow{d.} X$ and $Y_n \xrightarrow{d.} Y$ do not necessarily imply that $X_n + Y_n \xrightarrow{d.} X + Y$ or $X_n Y_n \xrightarrow{d.} XY$. (Can you come up with a counterexample?)

**Slutsky's Theorem**

- Slutsky's theorem is an important tool when deriving the asymptotic distributions for statistics.

- **Theorem 17** *If $X_n \xrightarrow{d.} X$ and $Y_n \xrightarrow{p.} c$, where $c$ is a constant. Then:*

  *(a)* $Y_n X_n \xrightarrow{d.} cX$;

  *(b)* $X_n + Y_n \xrightarrow{d.} X + c$.

- **Example 8** ***(t statistic)*** *Suppose $X_1, X_2, \ldots$ is a random sequence of i.i.d random variables with mean $\mu$ and variance $\sigma^2$ such that $0 < \sigma^2 < \infty$, and suppose that $var(S_n^2) \to 0$ as $n \to \infty$. Then by CLT we have*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d.} \mathcal{N}(0, 1);$$

*and by the result in Example 2 and the continuous mapping theorem we have*

$$\frac{\sigma}{S_n} \xrightarrow{p.} 1.$$

*Then by Slutsky's theorem, we can conclude that*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{d.} \mathcal{N}(0, 1).$$

- **Example 9** ***(OLS estimator in linear regression)*** *Consider the linear regression model without intercept term:*

$$Y_i = \beta X_i + \epsilon_i.$$

*The OLS estimator of $\beta$ is*

$$\hat{\beta} \equiv \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} \Longrightarrow \sqrt{n}(\hat{\beta} - \beta) = \frac{n^{-1/2} \sum_{i=1}^n X_i \epsilon_i}{n^{-1} \sum_{i=1}^n X_i^2}.$$

*By WLLN, $n^{-1} \sum_{i=1}^n X_i^2 \xrightarrow{p.} \mathbb{E}(X^2)$; and then by CMT, $[n^{-1} \sum_{i=1}^n X_i^2]^{-1} \xrightarrow{p.} [\mathbb{E}(X^2)]^{-1}$. By CLT, $n^{-1/2} \sum_{i=1}^n X_i \epsilon_i \xrightarrow{d.} \mathcal{N}(0, V)$, where $V \equiv var(X_i \epsilon_i) = \sigma_u^2 \mathbb{E}(X^2)$ under zero conditional mean and homoskedastic conditions. Combining these two results using the Slutsky's theorem, we get the familiar result that $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d.} \mathcal{N}(0, \sigma_n^2 [\mathbb{E}(X^2)]^{-1})$.*

### Cramér-Wold Theorem

- We have learned the tools that allow us to show convergence in distribution of a sequence of random variables. Let $\{\mathbf{b}_n\}$ be a sequence of $k \times 1$ random vectors. In general, convergence in distribution of each element in the vectors does not imply the *joint* convergence in distribution of the vectors. But, we have the following result that gives the condition for the joint convergence to hold.

- **Theorem 18** ***(Cramér-Wold Theorem)*** *Let $\{\mathbf{b}_n\}$ be a sequence of $k \times 1$ random vectors and suppose that for any real $k \times 1$ vector $\lambda$ such that $\lambda'\lambda = 1$ and $\lambda'\mathbf{b}_n \xrightarrow{d.} \lambda'\mathbf{Z}$, where $\mathbf{Z}$ is a $k \times 1$ vector with joint distribution function $F$. Then the limit distribution function of $\mathbf{b}_n$ exists and equals $F$.*

- Cramér-Wold theorem implies that $\mathbf{b}_n$ is asymptotically jointly normal if any linear combination of its elements is asymptotically normal.

- Consider a special case where $\lambda = (1, 0, \ldots, 0)'$, then $\lambda'\mathbf{b}_n \xrightarrow{d.} \lambda'\mathbf{Z}$ is just the convergence of the first element of $\mathbf{b}_n$ to that of $\mathbf{Z}$. Similar results hold for $\lambda = (0, 1, 0, \ldots, 0)'$ and so on. Therefore, convergence of the elements is a necessary but not sufficient condition of the convergence of the vectors.

**Delta Method**

- CLT tells us that if a statistic takes the form of a sample mean, then under mild conditions and proper normalization its asymptotic distribution is normal. However, sometimes we are not interested in the sample means themselves, but some functions of the sample means.

- **Theorem 19** ***(Delta Method)*** *Let $Y_n$ be a sequence of random variables that satisfies $\sqrt{N}(Y_n - \theta) \xrightarrow{d.} \mathcal{N}(0, \sigma^2)$.[1] For a given function $g$ and a specific value of $\theta$, suppose that $g'(\theta)$ exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{d.} \mathcal{N}(0, \sigma^2[g'(\theta)]^2).$$

- *Proof.* The first order Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ implies

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta) + remainder,$$

  where the remainder term takes the form of $c_n\sqrt{n}(Y_n - \theta)$, where $c_n \xrightarrow{p.} 0$. By Slutsky's theorem, the remainder term $\xrightarrow{p.} 0$. Then again by Slutsky's theorem, we have

$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{d.} g'(\theta) \cdot \mathcal{N}(0, \sigma^2) \stackrel{d.}{=} \mathcal{N}(0, g'(\theta)^2\sigma^2).$$

  ($\stackrel{d.}{=}$ means that the left hand side and the right hand side have the same distribution.)

[1] In many but not all cases, this results from some CLT.

- **Example 10** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean $\mu \neq 0$ and variance $0 < \sigma^2 < \infty$. Consider*

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\mu}\right) = \sqrt{n}[g(\bar{X}_n) - g(\mu)],$$

*where the function $g(x) = 1/x$. Note that $g'(x) = -1/x^2$, then using the delta method,*

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\mu}\right) \xrightarrow{d.} \mathcal{N}(0, \sigma^2/\mu^4).$$

*Standardizing the statistic, we get*

$$Z_n \equiv \frac{\sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\mu}\right)}{\sigma/\mu^2} \xrightarrow{d.} \mathcal{N}(0, 1);$$

*and by Slutsky's theorem,*

$$t_n \equiv \frac{\sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\mu}\right)}{S_n/(\bar{X}_n)^2} \xrightarrow{d.} \mathcal{N}(0, 1).$$

- If $g'(\theta) = 0$, then we often need the second-order delta method.

- **Theorem 20** ***(Second-Oder Delta Method)*** *Let $Y_n$ be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \xrightarrow{d.} \mathcal{N}(0, \sigma^2)$. For a given function $g$ and a specific value of $\theta$, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not $0$. Then*

$$n[g(Y_n) - g(\theta)] \xrightarrow{d.} \frac{\sigma^2 g''(\theta)}{2}\chi_1^2.$$

- *Proof.* Here we only provide a sketch of proof. When $g'(\theta) = 0$, we take one more term in the Taylor expansion to get

$$\begin{aligned} g(Y_n) - g(\theta) &= g'(\theta)(Y_n - \theta) + \frac{g''(\theta)}{2}(Y_n - \theta)^2 + remainder \\ &= \frac{g''(\theta)}{2}(Y_n - \theta)^2 + remainder. \end{aligned}$$

Recall that the square of a $\mathcal{N}(0,1)$ variable has $\chi_1^2$ distribution, which implies that

$$\frac{n(Y_n-\theta)^2}{\sigma^2} \xrightarrow{d.} \chi_1^2.$$

Then the result of the theorem can be established by applying Slutsky's theorem.

- **Theorem 21** ***(Multivariate Delta Method)*** *Let* $\{\mathbf{X}_n\}$ *be a sequence of* $k \times 1$ *random vectors such that* $\sqrt{n}(\mathbf{X}_n - \theta) \xrightarrow{d.} \mathcal{N}(0, \Sigma)$, *where* $\Sigma$ *is a* $k \times k$ *positive semi-definite matrix. Suppose that the function* $g$: $\mathbb{R}^k \to \mathbb{R}$ *is differentiable at* $\theta$ *and* $\nabla g(\theta)$ *denotes the corresponding gradient vector. Then*

$$\sqrt{n}[g(\mathbf{X}_n) - g(\theta)] \xrightarrow{d.} \mathcal{N}(0, \nabla g(\theta)^T \Sigma \nabla g(\theta)).$$

### Bounded in Probability

- **Definition 11** *A sequence of random variables* $\{X_n\}$ *is called to be* ***bounded in probability*** *if for any* $\epsilon > 0$ *there exists a constant* $B_\epsilon > 0$ *and an integer* $N_\epsilon$ *such that for all* $n \geq N_\epsilon$, *we have*

$$P(|X_n| \leq B_\epsilon) \geq 1 - \epsilon$$

- **Theorem 22** *Let* $\{X_n\}$ *be a sequence of random variables and let* $X$ *be a random variable. If* $X_n \xrightarrow{d.} X$, *then* $\{X_n\}$ *is bounded in probability.*

- **Theorem 23** *Let* $\{X_n\}$ *be a sequence of random variables bounded in probability and let* $\{Y_n\}$ *be a sequence of random variables which converge to* $0$ *in probability. Then* $X_n Y_n \xrightarrow{p.} 0$.

### Big O and Small o Notation

- In this subsection, we introduce the big O small o notation. It is very useful in econometrics.

- **Definition 12** ***(Big O Notation)*** *For a sequence of random variables and a sequence* $a_n$ *(random variables or positive constants), the notation*

$$X_n = O_p(a_n)$$

*means that the sequence* $X_n/a_n$ *is bounded in probability. We say that* $X_n$ *is at most of order in magnitude of* $a_n$.

- **Definition 13** ***(Small o Notation)*** *The notation*

$$X_n = o_p(a_n)$$

*means that the sequence $X_n/a_n$ converges to zero in probability. We say that $X_n$ is of smaller order in magnitude than $a_n$.*

- **Example 11** *$X_n = o_p(1)$ is equivalent to $X_n \xrightarrow{p.} 0$ (when $a_n = 1$).*

- **Theorem 24** *If $X_n = o_p(1)$, then $X_n = O_p(1)$.*

- **Example 12** *If $X_n = O_p(a_n)$ and $a_n \to 0$ (or $a_n = o_p(1)$), then $X_n = o_p(1)$.*

- CLT implies that $X_n - \mu = O_p(n^{-1/2})$; the delta method implies that $g(Y_n) - g(\theta) = O_p(n^{-1/2})$ if $g'(\theta) \neq 0$.

- The second-order delta method implies that $g(Y_n) - g(\theta) = O_p(n^{-1})$ if $g'(\theta) = 0$ and $g''(\theta) \neq 0$.

- **Example 13** *Using this notation, the result of Theorems 22 and 23 can be written succinctly as:*

  *(a) If $X_n \xrightarrow{d.} X$, then $X_n = O_p(1)$.*

  *(b) $O_p(1) \cdot o_p(1) = o_p(1)$.*

- The following results hold:

  (a) $O_p(1) + O_p(1) = O_p(1)$, $O_p(1) \cdot O_p(1) = O_p(1)$;

  (b) $o_p(1) + o_p(1) = o_p(1)$, $o_p(1) \cdot o_p(1) = o_p(1)$;

  (c) $O_p(1) + o_p(1) = O_p(1)$, $O_p(1) \cdot o_p(1) = o_p(1)$;

  (d) $O_p(a_n) + O_p(b_n) = O_p(\max\{a_n, b_n\})$, $O_p(a_n) \cdot O_p(b_n) = O_p(a_n \cdot b_n)$;

  (e) $o_p(a_n) + o_p(b_n) = o_p(\max\{a_n, b_n\})$, $o_p(a_n) \cdot o_p(b_n) = o_p(a_n \cdot b_n)$;

  (f) $O_p(a_n) + o_p(b_n) =?$ (it depends on the exact order in magnitude of each), $O_p(a_n) \cdot o_p(b_n) = o_p(a_n \cdot b_n)$;

## 5 Exercises

1. Prove Lemma 2.
2. Find $var(S^2)$ for a normal random sample.
3. Prove Theorems 5 and 6.
4. Prove the result in Example 4 using the definition of convergence in probability, that is, without invoking the continuous mapping theorem.
5. Prove Theorem 9 . Hint: what is the relationship between $P(|(X_n + Y_n) - (X + Y)| \geq \epsilon)$, $P(|X_n - X| \geq \epsilon/2)$ and $P(|Y_n - Y| \geq \epsilon/2)$?
6. Prove Theorem 13 for the case where $X_n$ and $X$ are continuous random variables. Follow these steps:
   (a) Given $t$ and $\epsilon$, show that $P(X \leq t - \epsilon) \leq P(X_n \leq t) + P(|X_n - X| \geq \epsilon)$. This gives a lower bound on $P(X_n \leq t)$;
   (b) Use a similar strategy to get an upper bound on $P(X_n \leq t)$;
   (c) By the squeeze theorem, deduce that $P(X_n \leq t) \to P(X \leq t)$.
7. Generate 1000 random samples $\{X_i\}_{i=1}^n$ of size $n$ from each of the following distributions. Try different sample sizes $n = 1, 5, 10, 25, 50, 100$.
   (a) Bernoulli distribution with $p = 0.78$;
   (b) Uniform distribution on $[0, 2]$;
   (c) Exponential distribution with mean 0.5;
   (d) Distribution with the pdf
   $$f_X(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{x} \exp[-(\ln x)^2/2];$$
   (e) t distribution with df 2;
   (f) Cauchy distribution.

   Find $\mu$ and $\sigma^2$ for each distribution. Plot the histogram of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ for each $n$, and compare the histograms with pdf of the standard normal distribution.

8. Let $X_1, X_2, \ldots$ be a random sample from Bernoulli($p$) distribution, and let $Y_n \equiv n^{-1} \sum_{i=1}^n X_i$.

    (a) Show that $\sqrt{n}(Y_n - p) \xrightarrow{d.} \mathcal{N}(0, p(1-p))$;

    (b) Show that for $p \neq 1/2$, the estimate of variance $Y_n(1-Y_n)$ satisfies $\sqrt{n}[Y_n(1-Y_n) - p(1-p)] \xrightarrow{d.} \mathcal{N}(0, (1-2p)^2 p(1-p))$;

    (c) Show that for $p = 1/2$, $n[Y_n(1-Y_n) - 1/4] \xrightarrow{d.} -\chi_1^2/4$.

9. Prove Theorem 24.