

Probability and Statistics

Omid Safarzadeh

January 30, 2022

Table of contents

- 1 Basic Concepts of Random Samples
- 2 Sums of Random Variable from a Random Sample
- 3 Convergence Concepts
 - Almost Sure Convergence
 - Convergence in Probability
 - Convergence in Distribution
 - The Delta Method
 - Some More Large Sample Results
- 4 Reference

***Acknowledgement:** This slide is prepared based on Casella and Berger, 2002

Basic Concepts of Random Samples

Definition 1.1

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 $f(x)$. Alternatively, X_1, \dots, X_n are called **independent and identically distributed random variable with pdf or pmf $f(x)$** . This is commonly abbreviated to iid random variables.

Basic Concepts of Random Samples

- In many experiments there are $n > 1$ repeated observations made on the variable, where X_1 is the first observation, X_2 is the second observation etc.
- In that case, each X_i has a marginal distribution given by $f(x)$.
- In addition, the value of one observation has no effect on or relationship with any of the other observations.
- X_1, \dots, X_n are mutually independent.

Basic Concepts of Random Samples

- Then, the joint pdf or pmf is given by

$$f(x_1, \dots, x_n | \theta) = f(x_1 | \theta) f(x_2 | \theta) \dots f(x_n | \theta) = \prod_{i=1}^n f(x_i | \theta),$$

where we assume that the population pdf or pmf is a member of a parametric family, and θ is the vector of parameters.

- The random sampling model in Definition (1.1) is sometimes called **sampling from an infinite population**.
- Suppose we obtain the values of X_1, \dots, X_n sequentially.
- First, the experiment is performed and $X_1 = x_1$ is observed.
- Then, the experiment is repeated and $X_2 = x_2$ is observed.
- The assumption of independence implies that the probability distribution for X_2 is unaffected by the fact that $X_1 = x_1$ was observed first.

Sums of Random Variable from a Random Sample

- Suppose we have drawn a sample (X_1, \dots, X_n) from the population. We might want to obtain some **summary** statistics.
- This summary might be defined as a function

$$T(x_1, \dots, x_n),$$

which might be real or vector-valued.

- So,

$$Y = T(X_1, \dots, X_n),$$

is a random variable or a random vector.

- We can use similar techniques as those introduced for functions of random variables, to investigate the distributional properties of Y .
- Thanks to (X_1, \dots, X_n) possessing the iid property, the distribution of Y will be tractable.
- This distribution is usually derived from the distribution of the variables in the random sample. Hence, it is called the **sampling distribution** of Y .

Sums of Random Variable from a Random Sample

Definition 2.1

Let X_1, \dots, X_n be a random sample of size n from a population and let $T = (x_1, \dots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of (X_1, \dots, X_n) . Then, the random variable or random vector $Y = T(X_1, \dots, X_n)$ is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution of Y** .

Definition 2.2

The **sample mean** is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Definition 2.3

The **sample variance** is the statistic defined by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

The **sample standard deviation** is the statistic defined by

$$S = \sqrt{S^2}.$$

Sums of Random Variable from a Random Sample

Lemma 2.1

Let X_1, \dots, X_n be a random sample from a population and let $g(x)$ be a function such that $E[g(X_1)]$ and $\text{Var}(g(X_1))$ exist. Then

$$E\left[\sum_{i=1}^n g(X_i)\right] = nE[g(X_1)]$$

and

$$\text{Var}\left(\sum_{i=1}^n g(X_i)\right) = n\text{Var}(g(X_1)).$$

- **Proof:** This is straightforward. First,

$$E\left[\sum_{i=1}^n g(X_i)\right] = \sum_{i=1}^n E[g(X_i)] = \sum_{i=1}^n E[g(X_1)] = nE[g(X_1)],$$

since X_i are distributed identically. Note that independence is not required here.

Sums of Random Variable from a Random Sample

- Then,

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^n g(X_i)\right) &= \sum_{i=1}^n \text{Var}(g(X_i)) + \sum_{i \neq j} \Sigma \text{Cov}(g(X_i), g(X_j)) \\ &= \sum_{i=1}^n \text{Var}(g(X_i)) + 0 = \sum_{i=1}^n \text{Var}(g(X_1)) = n\text{Var}(g(X_1)), \end{aligned}$$

where we have used the fact that

$$\text{Cov}(g(X_i), g(X_j)) = 0 \quad \text{for all } i \neq j,$$

due to independence and that $\text{Var}(g(X_i))$ is the same for all X_i , due to their distribution being identical.

Sums of Random Variable from a Random Sample

Theorem 2.1

Let X_1, \dots, X_n be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then,

- ① $E[\bar{X}] = \mu,$
- ② $\text{Var}(\bar{X}) = \sigma^2 / n,$
- ③ $E[S^2] = \sigma^2.$

Convergence Concepts

- The underlying idea in this section is to understand what happens to sequence of random variables, or also summary statistics, when we let the sample size go to infinity.
- This is, of course, an idealistic concepts, if you like, because the sample size never goes to infinity. However, the idea is to attain a grasp of what happens when the sample size becomes **large enough**.
- This is important because, although important results are based on the case when the sample size approaches ∞ , they are actually relevant for finite sample size, which are **large enough**.
- How large is "large enough" is very much related to the data, its dependence structure, the econometric model at hand etc, so it is difficult to give a proper rule of thumb.

Convergence Concepts

- The tools you will learn in this section are the fundamental building blocks of what is known as "asymptotic theory" or "large sample theory". Although a bit abstract at first sight, these results are at the core of many proofs you will encounter in econometrics articles.
- The three important concepts we will consider in what follows are
 - 1 almost sure convergence,
 - 2 convergence in probability,
 - 3 convergence in distribution.
- In the first instance we will consider the case where a sequence of random variables X_1, \dots, X_n are independently and identically distribution (iid). This is one (and the simplest) of many possible dependence settings.

Almost Sure Convergence

- Remember that random variables are functions defined on the sample space, e.g. $X(\omega)$. Our interest will be on a sequence of random variables, indexed by sample size, i.e. $X_n(\omega)$.
- To motivate the following discussion, consider pointwise convergence:

$$\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \quad \text{for all } \omega \in \Omega,$$

where Ω is, as before, the sample space.

- Notice that, convergence occurs for all ω ! There is not much of probabilistic statement here.
- This is the strongest form of convergence we can have on the sample space. But it is not relevant for probabilistic statements.
- A slightly restricted version of pointwise convergence is **almost sure convergence**.

Almost Sure Convergence

- Remember that random variables are functions defined on the sample space, e.g. $X(\omega)$. Our interest will be on a sequence of random variables, indexed by sample size, i.e. $X_n(\omega)$.
- To motivate the following discussion, consider pointwise convergence:

$$\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \quad \text{for all } \omega \in \Omega,$$

where Ω is, as before, the sample space.

- Notice that, convergence occurs for all ω ! There is not much of probabilistic statement here.
- This is the strongest form of convergence we can have on the sample space. But it is not relevant for probabilistic statements.
- A slightly restricted version of pointwise convergence is **almost sure convergence**.

Almost Sure Convergence

Definition 3.1

Almost Sure Convergence: A sequence of random variables X_1, X_2, \dots defined on a probability space (Ω, \mathcal{F}, P) convergence almost surely to a random variable X if

$$\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega),$$

for each ω , except for $\omega \in E$, where $P(E) = 0$.

- The idea is this: pointwise convergence fails for some points in Ω . However, the number of such points is so small that we can safely assign zero measure (or zero probability) to the set of these points.
- Other ways of expressing this definition are,

$$P\left(\lim_{n \rightarrow \infty} |X_n - X| > \varepsilon\right) = 0 \quad \text{for every } \varepsilon > 0,$$

or

$$P(\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}) = 1.$$

Almost Sure Convergence

- This time, we have a difference. Convergence fails on a very small set E such that $P(E) = 0$.
- That $P(E) = 0$ is due to the set being so small that we can safely assign zero probability to the set.
- Remember that in earlier lectures we have stated that for a continuously distributed random variable, the probability of a single point is always equal to zero. This is similar, in spirit, to the situation at hand.
- This type of convergence is also called **convergence almost everywhere** and **convergence with probability 1**.
- The following notation is common:

$$X_n \xrightarrow{a.s.} X$$

$$X_n \xrightarrow{wp1} X$$

$$\lim_{n \rightarrow \infty} X_n = X \text{ a.s.}$$

- Note that, at the cost of notational sloppiness, the argument of $X_n(\omega)$ is usually dropped.
- Also, $X_n(\omega)$ need not converge to a function. It can also simply converge to some constant, say, a .

Convergence in Probability

- The next convergence type is **convergence in probability** . Its definition is similar to that of almost sure convergence but in essence it is a much weaker convergence concept.

Definition 3.2

Convergence in Probability: A sequence of random variables X_1, X_2, \dots **converges in probability** to a random variable X if, for every $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0.$$

Convergence in Probability

- An equivalent statement is that given $\varepsilon > 0$ and $\sigma > 0$ there exists an N , which depends on both σ and ε , such that $P(|X_n - X| > \varepsilon) < \sigma$ for all $n > N$. This merely is a restatement using the formal definition of limit.
- One could also write

$$\lim_{n \rightarrow \infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0,$$

$$X_n \xrightarrow{P} X,$$

or

$$p \lim_{n \rightarrow \infty} X_n = X.$$

Convergence in Probability

- It is not easy to see the difference between the two modes of convergence. However, let's give it a try!
- Almost sure convergence states that we have pointwise convergence for all $\omega \in \Omega$ except for a small, zero measure set E . Importantly, this set is independent of n .
- Convergence in probability states that as the sample size goes towards ∞ , the probability that X_n will deviate from X by more than ε decreases towards zero.
- However, for any sample size, there is a positive probability that X_n will deviate by more than ε . In other words, for some $\omega \in E_n \subset \Omega$, such that $P(E_n) > 0$, $|X_n - X|$ will be larger than ε .
- Importantly, there is nothing that restricts E_n to be the same for all n . Hence, the set on which X_n deviates from X may change as n increases.
- The good news is, deviation probability slowly goes to zero, hence $P(E_n)$ will eventually be zero.

Convergence in Probability

- Now, as before, let $Z_i, i = 1, \dots, n$, be some random variable and let

$$X_n = \frac{1}{n} \sum_{i=1}^n Z_i.$$

White (2001, p.24): "With almost sure convergence, the probability measures P takes into account the joint distribution of the entire sequence $\{Z_i\}$, but with convergence in probability, we only need concern ourselves sequentially with the joint distribution of the elements of $\{Z_i\}$ that actually appear in X_n , typically the first n .

Convergence in Probability

- Associated with convergence in probability is the Weak Law of Large Numbers (WLLN)

Theorem 3.1

Weak Law of Large Numbers: If X_1, X_2, \dots are iid random variables with common mean $\mu < \infty$ and variance $\sigma^2 < \infty$, then

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mu,$$

as $n \rightarrow \infty$.

Convergence in Probability

- **Proof:** The proof uses Chebychev's Inequality. Remember this says that if X is a random variable and if $g(x)$ is a non-negative function, then, for any $r > 0$,

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r}.$$

Now, consider

$$P(|\bar{X}_n - \mu| \geq \varepsilon) = P((\bar{X}_n - \mu)^2 \geq \varepsilon^2) \leq \frac{E[(\bar{X}_n - \mu)^2]}{\varepsilon^2} = \frac{\text{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2},$$

where we use $r = \varepsilon^2$ and $g(\bar{X}_n) = (\bar{X}_n - \mu)^2$.

- The above result implies that

$$P(|\bar{X}_n - \mu| < \varepsilon) = 1 - P(|\bar{X}_n - \mu| \geq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2},$$

and

$$\lim_{n \rightarrow \infty} 1 - \frac{\sigma^2}{n\varepsilon^2} = 1.$$

- Hence,

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1.$$

Convergence in Probability

- This is perhaps a good time to stop and reflect a little bit on these new concepts.
- What the Weak and Strong LLNs are saying is that under certain conditions, the **sample mean converges** to the **population mean** as $n \rightarrow \infty$. This is known as consistency: one would say that **the sample mean is a consistent estimator of the population mean**.
- In actual applications, this means that if the sample size is **large enough**, then the sample mean is close to the population mean. So n does not have to be that close to infinity. On the other hand, as mentioned at the beginning, **“how large”** the sample size should be in order to be considered a **“large enough”** sample is a different question in its own. We will not deal with this here.
- Sometimes, consistency is compared to unbiasedness.

Convergence in Probability

- An estimator $\hat{\beta}$ of a population value β is an unbiased estimator

$$E[\hat{\beta}] = \beta.$$

- What are the things we might want to estimate? One example would be parameters of a distribution family. For example, we might know that the data are distributed with $N(\mu, \sigma^2)$, but we may not know the particular values of μ and σ^2 . In this case, we would estimate these parameters.

Convergence in Probability

- The p lim operator has some nice properties that makes it much more convenient to deal with, compared to the expectation operator. In particular, let X_1, X_2, \dots and Y_1, Y_2, \dots be two random sequences and a_1, a_2, \dots be some non-stochastic sequence. Then, we have the following.

①

$$p \lim_{n \rightarrow \infty} \frac{X_n}{Y_n} = \frac{p \lim_{n \rightarrow \infty} X_n}{p \lim_{n \rightarrow \infty} Y_n},$$

while we usually have

$$E\left[\frac{X_n}{Y_n}\right] \neq \frac{E[X_n]}{E[Y_n]};$$

②

$$p \lim_{n \rightarrow \infty} (X_n + Y_n) = p \lim_{n \rightarrow \infty} X_n + p \lim_{n \rightarrow \infty} Y_n;$$

③

the p lim of a non-random sequence is equal to its limit:

$$p \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} a_n.$$

Convergence in Probability

- Note that, one concept does not usually imply another. In other words, a consistent estimator can be biased, while an unbiased estimator can be inconsistent.
- Suppose we are trying to estimate the population parameter β . Consider the following estimators.
 - 1 $\hat{\beta} = \beta + 20/n$: consistent but biased.
 - 2 $\hat{\beta} = X$ where $P(X = \beta + 100) = P(X = \beta - 100) = 0.5$: unbiased but inconsistent.

Convergence in Probability

- Returning to the discussion at hand, it is important to acknowledge that neither almost sure convergence nor convergence in probability (and nor any convergence type) says anything about the distribution of the sequence X_1, X_2, \dots . For example, it might be such that the distribution of X_i changes as i varies. This is fine.
- So far, we have only considered LLNs that work when the sequence is drawn from an iid population. If this assumption is violated, we can still probably have convergence of the sample mean to the population mean, **but we will have to find an appropriate LLN that works for the particular population distribution we have.**
- A useful result relating almost sure convergence and convergence in probability is that

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X.$$

Convergence in probability, however, does not imply almost sure convergence.

- Obviously, for some constant K

$$K \xrightarrow{a.s.} K \quad \text{and} \quad K \xrightarrow{P} K.$$

Convergence in Probability

- As far as economists and most of the econometricians are concerned, one would not care too much about whether convergence is achieved almost surely or in probability. As long as convergence is achieved, the rest is not important.
- However, in some cases it might be easier to prove the LLN for one of the two convergence types. This is no problem, as $\xrightarrow{a.s.} \text{ implies } \xrightarrow{P}$ anyway.
- In addition, convergence almost surely might be slower than convergence in probability in the sense that it might require a larger sample size before the sample mean is close enough to the population mean.

Convergence in Probability

- Before we move on to a different type of convergence, let us, for sake of completeness, introduce one more type of convergence.

Definition 3.3

L_p Convergence: Let $0 < p, \infty$, let X_1, X_2, \dots be a sequence of random variables with $E[|X_n|^p] < \infty$ and let X be a random variable with $E[|X|^p] < \infty$. Then, X_n converges in L_p to X if

$$\lim_{n \rightarrow \infty} E[|X_n - X|^p] = 0.$$

- Just as a reference, L_p convergence does not imply almost sure convergence and nor does almost sure convergence imply L_p convergence. However, L_p convergence implies convergence in probability.

Convergence in Distribution

Definition 3.4

Convergence in Distribution: A sequence of random variables X_1, X_2, \dots **converges in distribution** to a random variable X if

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

at every x where $F(x)$ is continuous.

- This is also called **convergence in law**. The following short hand notation is used to denote convergence in distribution:

$$X_n \xrightarrow{d} X,$$

$$X_n \xrightarrow{d} F_X,$$

$$X_n \xrightarrow{L} F_X.$$

- It is important to underline that it is not X_n that converges to a distribution. Instead, it is the distribution of X_n that converges to the distribution of X .

Convergence in Distribution

- As far as sequences of random vectors are concerned, a sequence of random vectors $X_n = (X_{1,n}, \dots, X_{d,n})$ converges in distribution to a random vector X if

$$\lim_{n \rightarrow \infty} F_{X_n}(x_1, \dots, x_d) = F_X(x_1, \dots, x_d),$$

at every $x = (x_1, \dots, x_d)$ where $F(x_1, \dots, x_d)$ is continuous.

- Importantly, convergence in probability implies convergence in distribution.

Theorem 3.2

If the sequence of random variables X_1, X_2, \dots converges in probability to a random variable X , the sequence also converges in distribution to X .

- Consequently, almost sure convergence implies convergence in distribution, as well.

Theorem 3.3

The sequence of random variables X_1, X_2, \dots converges in probability to a constant a **if and only if** the sequence also converges in distribution to a . Equivalently, the statement

$$P(|X_n - a| > \varepsilon) \rightarrow 0 \text{ for every } \varepsilon > 0$$

is equivalent to

$$F_{X_n}(x) = P(X_n \leq x) \rightarrow \begin{cases} 0 & \text{if } x < a \\ 1 & \text{if } x > a \end{cases}$$

Convergence in Distribution

- We now introduce one of the most useful theorems we have considered so far.

Theorem 3.4

Central Limit Theorem: Let X_1, X_2, \dots be a sequence of iid random variables with $E[X_i] = \mu < \infty$ and $0 < \text{Var}(X_i) = \sigma^2 < \infty$. Define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Let $G_n(x)$ denote the cdf of $\sqrt{n} \frac{(\bar{X}_n - \mu)}{\sigma}$. Then, for any x , $-\infty < x < \infty$,

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

In other words,

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

Convergence in Distribution

- This is a powerful result! We start with the iid and finite mean and variance assumptions. In return, the Central Limit Theorem (CLT) promises us that the distribution of a properly standardised version of the sample mean given by

$$\sqrt{n} \frac{(\bar{X}_n - \mu)}{\sigma}$$

will converge to the standard normal distribution as the sample size tends to infinity.

- As before, the sample size will never be equal to ∞ . BUT, for large enough samples, $\sqrt{n} \frac{(\bar{X}_n - \mu)}{\sigma}$ will be **approximately standard normal**. As n becomes larger, this approximate result will become more accurate.
- As with LLNs, it is possible to obtain CLTs for non-iid data. However, this will require one to make stronger assumptions regarding the moments of the sequence of random variables. The trade-off between dependence and moment assumptions is always there.
- Two useful results are given next.

Convergence in Distribution

Theorem 3.5

If X_n is a sequence of random vectors each with support χ , $g(x)$ is continuous on χ and

$$X_n \xrightarrow{d} X,$$

then

$$g(X_n) \xrightarrow{d} g(X)$$

Theorem 3.6

Slutsky's Theorem: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} k$, where k is a constant, then

① $Y_n X_n \xrightarrow{d} kX,$

② $X_n + Y_n \xrightarrow{d} X + k.$

Example 3.1

Suppose that

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

however the value of σ is unknown. What to do?

The Delta Method

- When talking about the CLT, our focus has been on the limiting distribution of some standardised random variable.
- There are many instances, however, when we are not specifically interested in the distribution of the standardised random variable itself, but rather of some function of it.
- The delta method comes in handy in such cases. This method utilises our knowledge on the limiting distribution of a random variable in order find the limiting distribution of a function of this random variable.
- In essence, this method is a combination of Slutsky's Theorem and Taylor's approximation.

Theorem 3.7

Delta Method: Let Y_n be a sequence of random variables that satisfies

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

For a given function $g(\cdot)$ and a specific value of θ , suppose that $g'(\theta)$ exist and $g'(\theta) \neq 0$. Then,

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

Some More Large Sample Results

- **Corollary (White (2001)):** Let X_1, \dots, X_n be a sequence of independent random variables such that $E[|X_i|^{1+\sigma}] < \infty$ for some $\sigma > 0$ and all i . Then, $\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mu_i \xrightarrow{a.s.} 0$.

Casella, G., & Berger, R. (2002). *Statistical inference*. Cengage Learning.
<https://books.google.fr/books?id=FAUVEAAAQBAJ>