

# Mathematical Statistics I

## Chapter 5: Limit theorems

Jesse Wheeler

### Contents

<b>1 Convergence Concepts</b>	<b>1</b>
<b>2 Test Section</b>	<b>8</b>

## 1 Convergence Concepts

### Introduction

- This material comes primarily from Rice (2007, Chapter 5), but will be supplemented with material from Casella and Berger (2024, Chapter 5).
- In this chapter, we are interested in the convergence of sequences of random variables.
- In particular, we are interested in the convergence of the sample mean,  $\bar{X}_n = (X_1 + X_2 + \dots + X_n)/n$ , as the number of samples  $n$  grows.
- Because  $\bar{X}_n$  is itself a random variable, we have to carefully define what it means for the convergence of a random variable.
- In this class, we are mainly concerned with three types of convergence.
- Because convergence of random variables is a tricky topic, we will treat them in varying amounts of detail.

### Convergence in Probability

- The first type of convergence is one of the weaker types, and is usually easy(ish) to verify.

#### Definition: Convergence in Probability

A sequence of random variables  $X_1, X_2, \dots$  converges in probability to a random variable  $X$  if, for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \epsilon) = 0$$

or, equivalently,

$$\lim_{n \rightarrow \infty} P(|X_n - X| < \epsilon) = 1.$$

- We often use the shorthand  $X_n \xrightarrow{P} X$  to denote “ $X_n$  converges in probability to  $X$  as  $n$  goes to infinity”.

- Note that the  $X_i$  in the definition above do *not* need to be independent and identically distributed.
- The distribution of  $X_n$  changes as the subscript changes, and each of the convergence concepts we will discuss will describe different ways in which the distribution of  $X_n$  converges to some limiting distribution as the subscript becomes large.
- A special case is when the limiting random variable  $X$  is a constant.

**Example: The (Weak) Law of Large Numbers**

Let  $X_1, X_2, \dots$  be iid random variables with  $E[X_i] = \mu$  and  $\text{Var}(X_i) = \sigma^2$ . Define  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ . Then  $\bar{X}_n \xrightarrow{P} \mu$ .

*Proof.* The proof is a straightforward application of Chebychev's Inequality.

- We want to show that

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

- For every  $\epsilon > 0$ , Chebychev's inequality gives us:

$$\begin{aligned} P(|\bar{X}_n - \mu| \geq \epsilon) &= P((\bar{X}_n - \mu)^2 \geq \epsilon^2) \\ &\leq \frac{E[(\bar{X}_n - \mu)^2]}{\epsilon^2} \\ &= \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}. \end{aligned}$$

- Thus, taking the limit, we have  $\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$

□

- The WLLN is very elegant; under general conditions, the sample mean of independent random variables approaches the population mean as  $n \rightarrow \infty$ .
- This is also used for proportions, as proportions are just means of indicator random variables.
- The WLLN can also be extended to show that the results hold even if the variance is infinite, the only condition needed is that the expectation is finite. However, the proof in this case is beyond the scope of this course.
- When a sequence of the “same” sample quantity approaches a constant, we say that the sample quantity is *consistent*.
- A natural extension of the definition of the convergence of probability, is convergence of functions of random variables:  $h(X_1), h(X_2), \dots$

**Theorem: Convergence in probability for continuous functions**

Let  $X_1, X_2, \dots$  be a sequence of random variables that converges in probability to a random variable  $X$ , and let  $h$  be a continuous function.

Then,  $h(X_1), h(X_2), \dots$  converges in probability to  $h(X)$ .

## Almost sure convergence

- Our next convergence concept is stronger than convergence in probability.

### Definition: Almost Sure Convergence

A sequence of random variables  $X_1, X_2, \dots$  converge *almost surely* to a random variable  $X$  if, for every  $\epsilon > 0$ ,

$$P\left(\lim_{n \rightarrow \infty} |X_n - X| < \epsilon\right) = 1,$$

or

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

- Almost sure convergence is often written as  $X_n \xrightarrow{a.s.} X$ .
- It appears similar to convergence in probability, but they are in fact very different. In particular, almost sure convergence is a stronger concept.
- One way to think about this difference is that the probability gives a weight to individual sets.
- For convergence in probability, the set where  $|X_n - X| > \epsilon$  can have positive probability, but that probability converges to zero for large  $n$ .
- For almost sure convergence, the set where  $|X_n - X| > \epsilon$  has probability zero. This doesn't imply that the set  $|X_n - X| > \epsilon$  is empty, but it has zero probability.
- Almost sure convergence is very similar to pointwise convergence of a sequence of functions. This is no accident, as random variables *are* functions:

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

- In the equivalent definition above, we see we must have point-wise convergence *almost-everywhere*, except for the possibility that for some set  $N \subset \Omega$  such that  $P(N) = 0$ , we allow  $s \in N$  to not converge:  $\lim_{n \rightarrow \infty} X_n(s) \neq X(s)$ .

*Example: Convergence in prob, not a.s.*

Let the sample space  $\Omega = [0, 1]$ , and assign the uniform probability on this interval. Define the sequence of random variables  $X_i$  as:  $X_1(s) = s + 1_{[0,1]}(s)$ ,  $X_2(s) = s + 1_{[0,\frac{1}{2}]}(s)$ ,  $X_3(s) = s + 1_{[\frac{1}{2},1]}(s)$ ,  $X_4(s) = s + 1_{[0,\frac{1}{3}]}(s)$ ,  $X_5(s) = s + 1_{[\frac{1}{3},\frac{2}{3}]}(s)$ ,  $X_6(s) = s + 1_{[\frac{2}{3},1]}(s)$ , ..., and then define  $X(s) = s$ . We can see that  $X_n \xrightarrow{P} X$ . However,  $X_n$  does not converge almost surely, because there is *no* values  $s \in \Omega$  that satisfy  $X_n(s) \rightarrow X(s)$ . For every  $\omega$ , the value of  $X_n(s)$  alternates between  $s$  and  $s + 1$  infinitely often.

### Theorem: almost sure convergence implies convergence in probability

If  $X_1, X_2, \dots$  are a sequence of random variables such that  $X_n \xrightarrow{a.s.} X$ , for some random variable  $X$ , then  $X_n \xrightarrow{P} X$ .

- The converse of the statement above is false. That is, convergence in probability does not imply almost sure convergence.
- A proof of the theorem above, as well as additional treatment of the connection between almost sure convergence and convergence in probability is found in Resnick (2019, Chapter 6).
- Note: As stated, the weak-law of large numbers (WLLN) can actually be shown to hold a.s., in which case we call it the strong-law of large numbers (SLLN).

## Convergence in Distribution

- The final form of convergence we will consider in this course is convergence in distribution.

### Definition: 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 all points  $x$  where  $F_X(x)$  is continuous.

- One way to think about convergence in distribution is that it's really a statement about the long-run behavior of a sequence of random variables, as it's a statement about the CDFs.
- This is different from the other types of convergence, which are concerned with the random variable itself.
- A quick recap of how the different types of convergence are related:
  - a.s. convergence  $\implies$  convergence in prob  $\implies$  convergence in Distribution.
- In a few special scenarios, we can talk about more connections between the types of convergence.
- One such example is convergence in probability to a constant. Casella and Berger (Theorem 5.5.13 of 2024) shows that  $X_n \xrightarrow{P} a$  for some constant  $a$  if and only if  $X_n \xrightarrow{d} a$ .

## The Central Limit Theorem

- Next we are going to introduce the Central Limit Theorem (CLT).
- The CLT is easily one of the most important theorems across all scientific disciplines, and arguably the most important result to modern science.

“I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the [CLT]. The law would have been personified by the Greeks and deified, if they had known of it...” - Sir Francis Galton

- The theory for the CLT was developed over a period of roughly 100 years, done by some of the greatest mathematicians of the 19th and 20th centuries.
- The theorem states that, under very weak conditions, the sum of any sequence of iid random variables (with finite mean and variance) converges to a normal distribution.
- Here, we are going to work towards a proof of a simple (weak) version of the theorem.

### Theorem: Continuity Theorem

Let  $X_n$  be a sequence of random variables with cdf  $F_n(x)$ , and let  $X$  be a random variable with cdf  $F(x)$ . Furthermore, let  $M_n(t)$  be the moment generating function of  $X_n$ , and  $M(t)$  the moment generating function of  $X$ .

If  $M_n(t) \rightarrow M(t)$  for all  $t$  in an open interval containing zero, then  $F_n(X) \rightarrow F(x)$  at all continuity points of  $F$ . That is,  $X_n \xrightarrow{d} X$ .

- Now, we do a brief reminder about Taylor Series and Taylor's Theorem

**Theorem: Taylor Series**

If a function  $f(x)$  has derivatives of order  $k$ , that is,  $\frac{d^k}{dx^k} f(x)$  exists, then for any constant  $a$ , the *Taylor Polynomial* of order  $k$ , centered about  $a$ , is

$$f(x) = \sum_{n=0}^k \frac{f^{(n)}(a)}{n!} (x-a)^n + R_k(x),$$

where  $R_k(x) = h_k(x)(x-a)^k$ , for some  $h_k$  such that  $\lim_{x \rightarrow a} h_k(x) = 0$ .

- In particular, it means that we can use a  $k$  order polynomial to approximate a differentiable function, and the remainder term  $R_k(x)$  goes to zero at a rate smaller than the rate that  $(x-a)^k$  goes to zero.

**Theorem: The (classic) Central Limit Theorem**

Let  $X_1, X_2, \dots$  be a sequence of independent and identical random variables with mean  $E[X_i] = \mu$  and variance  $\text{Var}(X_i) = \sigma^2 < \infty$ . Assume that the mgf of  $X_i$  exists and is defined in a neighborhood of zero, and denote the cdf and mgf as  $F$  and  $M$ , respectively. Then,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

*Proof.* The idea for this proof is to show that, for some neighborhood around  $t = 0$ , the mgf of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$  converges to  $e^{t^2/2}$ , which is the mgf of a  $N(0, 1)$  random variable. For simplicity, we will assume that the mgf  $M(t)$  exists around a symmetric neighborhood around zero (it could be larger), i.e., it exists if  $|t| < h$  for some  $h > 0$ .

Let  $Z_i = (X_i - \mu)/\sigma$ . Using the properties of mgf, the mgf of  $Z_i$ , denoted  $Z_i$  exists for  $|t| < \sigma t$ . The exact form of  $M_Z(t)$  can be found using the theorem on the mgf of linear transformations of random variables, but is not needed. Now note that the target RV  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$  can be written as:

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

Thus, using the properties of the mgf, we have

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

We now do a second order Taylor-Expansion of  $M_Z(t/\sqrt{n})$  about 0:

$$M_Z(t/\sqrt{n}) = M_Z(0) + M_Z^{(1)}(0) \frac{(t/\sqrt{n})^1}{1!} + M_Z^{(2)}(0) \frac{(t/\sqrt{n})^2}{2!} + R_2(t/\sqrt{n})$$

In particular, we note that  $M_Z(0) = E[e^{0X}] = E[1] = 1$ , and by how the mgf and  $Z$  are defined, we have  $M'_Z(0) = E[Z] = 0$ , and  $M''_Z(0) = E[Z^2] = \text{Var}(Z) = 1$  (since the mean is zero). Therefore

$$\begin{aligned} M_Z(t/\sqrt{n}) &= 1 + \frac{(t/\sqrt{n})^2}{2!} + R_2(t/\sqrt{n}) \\ &= 1 + \frac{t^2}{2n} + R_2(t/\sqrt{n}) \end{aligned}$$

Now by Taylor's theorem, there exists some function  $h_2(x)$  such that

$$h_2(x) = \frac{R_2(t/\sqrt{n})}{(t/\sqrt{n})^2},$$

where

$$\lim_{t \rightarrow 0} \frac{R_2(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0.$$

or equivalently,

$$\lim_{n \rightarrow \infty} \frac{R_2(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0.$$

Since  $t$  is fixed, (and because the below equality holds when  $t = 0$ ), this implies:

$$\lim_{n \rightarrow \infty} \frac{R_2(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{n \rightarrow \infty} nR_2(t/\sqrt{n}) = 0.$$

Finally, returning to the mgf of  $\sqrt{n}(\bar{X}_n - \mu)$ , we take the limit as  $n \rightarrow \infty$ :

$$\begin{aligned} \lim_{n \rightarrow \infty} M_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) &= \lim_{n \rightarrow \infty} \left( M_Z(t/\sqrt{n}) \right)^n \\ &= \lim_{n \rightarrow \infty} \left( 1 + \frac{t^2}{2n} + R_2(t/\sqrt{n}) \right)^n \\ &= \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{n} \left[ \frac{t^2}{2} + nR_2(t/\sqrt{n}) \right] \right)^n \\ &= \lim_{n \rightarrow \infty} \left( 1 + \frac{a_n}{n} \right)^n, \end{aligned}$$

where  $a_n = \frac{t^2}{2} + nR_2(t/\sqrt{n})$ . Here, we note that  $a_n \rightarrow \frac{t^2}{2}$  due to the convergence of  $nR_2(t/\sqrt{n}) \rightarrow 0$ , and we can apply the theorem (e.g., Lemma 2.3.14 of Casella and Berger, 2024) that if  $a_n \rightarrow a$ , then

$$\lim_{n \rightarrow \infty} \left( 1 + \frac{a_n}{n} \right)^n = e^a.$$

Therefore, we get

$$\lim_{n \rightarrow \infty} M_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) = e^{t^2/2}.$$

Thus, by the continuity theorem,

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

□

- One practical implication of the CLT is that, for large  $n$ , we can approximate

$$\bar{X}_n \xrightarrow{d} N(\mu, \sigma^2/n),$$

if  $X_i$  are independent and identically distributed with finite mean and variance.

- In practice,  $n \approx 30$  has been found to lead to good approximations, but it depends heavily on the distribution of  $X_i$ .
- A further investigation of the CLT proof shows that the convergence towards the normal distribution happens at a rate of  $1/\sqrt{n}$ .
- If we used a Taylor-series approximation with one additional order, we could derive a more accurate approximation under additional conditions known as the Edgeworth Expansion (See Theorem 19.3 of Keener, 2010). These are less commonly used in practice, because you need finite third moments.

*Example: Binomial-Normal Approximation*

Let  $X \sim \text{Binomial}(n, p)$ . For any  $k \in \{0, 1, \dots\}$ , approximate  $P(X \leq k)$ .

- Note that if  $X$  is binomial distributed, it has the same distribution as the sum of Bernoulli random variables.
- Let  $X_1, X_2, \dots, X_n$  be Bernoulli( $p$ ) random variables, and then  $X \stackrel{d}{=} \sum_i X_i$ .
- Recall  $E[X_i] = p$ ,  $\text{Var}(X_i) = p(1-p)$ .
- By considering  $\bar{X}_n = \frac{1}{n} \sum_i X_i$ , we can use the CLT to approximate:

$$\sqrt{n}(\bar{X}_n - p) \stackrel{d}{\approx} N(0, p(1-p)).$$

- Therefore,

$$\bar{X}_n \stackrel{d}{\approx} N(p, p(1-p)/n),$$

and

$$X \stackrel{d}{=} \sum_i X_i = n\bar{X}_n \stackrel{d}{\approx} N(np, np(1-p)).$$

Using this approximation, we have

$$P(X \leq k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right).$$

- One thing to note is we can make a fairly simple improvement to this approximation by doing a continuity correction.
- Specifically,  $X$  only takes on real values, so if we approximate it with a continuous distribution  $X \stackrel{d}{\approx} Y$ , we generally get a more accurate approximation if we use  $P(X \leq k) \approx P(Y \leq k + 0.5)$ .

## Slutsky's Theorem

- The following theorem is useful for our notes and supporting other ideas we will cover. However, we won't discuss the proof because it relies on other convergence concepts we don't cover in this class

### Theorem: Slutsky's Theorem

If  $X_n \xrightarrow{d} X$  and  $Y_n \xrightarrow{p} a$  for some random variable  $X$  and constant  $a$ , then

1.  $Y_n X_n \xrightarrow{d} aX$ .
2.  $X_n + Y_n \xrightarrow{d} X + a$ .

*Example: CLT with estimated variance*

(HW problem)? Suppose that  $X_i$  are iid  $N(\mu, \sigma^2)$  random variables. By the CLT, we have

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

The problem with this theorem in practice is that we assume  $\sigma$  is known, which is often not practical. If  $S_n^2$  is our estimate of the variance, and  $S_n^2 \xrightarrow{p} \sigma^2$ , then it can be shown that  $\sigma/S_n \xrightarrow{p} 1$ . Thus, by Slutsky's Theorem:

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

## **Delta-Method**

- TODO

## **2 Test Section**

**Test Frame**

## Acknowledgments

- Compiled on November 4, 2025 using R version 4.5.1.
- Licensed under the [Creative Commons Attribution-NonCommercial license](#).  Please share and remix non-commercially, mentioning its origin.
- We acknowledge [students and instructors for previous versions of this course / slides](#).

## References

- Casella G, Berger R (2024). *Statistical inference*. Chapman and Hall/CRC. [1](#), [4](#), [5](#)
- Keener RW (2010). *Theoretical statistics: Topics for a core course*. Springer Science & Business Media. [5](#)
- Resnick S (2019). *A probability path*. Springer. [3](#)
- Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. [1](#)