Chapter 10. Convergence Concepts

10.1 Convergence in Distribution and the Central Limit Theorem

10.1.1 Statement of the Central Limit Theorem

One of the most important probabilistic results for statistics is the central limit theorem. It states that under very general conditions the distribution of the sample mean of a random sample approaches normality as the size of the sample size increases. This means that we can use intervals such as those described in Chapter 8 and other statistical methods based on the normal distribution *even when the individual random variables are not normal*. We begin with a formal statement of the central limit theorem.

Theorem 10.2. *(Central Limit Theorem,* CLT*) Let $X_1, X_2, \ldots$ be a countable sequence of i.i.d random variables$^{(\star\star)}$ with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, and $G_n(x)$ be the cdf of $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$. Then*

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

$^{(\star\star)}$ To prove the CLT we assume the *moment generating function* of X_i exists (see below).

Example Suppose X_i are i.i.d. Bernoulli(p) for $i = 1, 2, \ldots$ and let $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

We know $n\overline{X}_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$. Then $\mathrm{E}(n\overline{X}_n) = np$, $\mathrm{Var}(n\overline{X}_n) = np(1-p)$, and for large n,

$$\frac{n\overline{X}_n - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \overset{\text{approx}}{\sim} \mathrm{N}(0,1).$$

This result can be verified by applying the CLT to $X_1, X_2, \ldots$. In particular, note that $\mathrm{E}(X_i) = p$ and $\mathrm{Var}(X_i) = p(1-p)$, so that by the CLT, the CDF of

$$\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}}$$

converges to Φ, the CDF of a standard normal random variable. ■

10.2.1 Convergence in Distribution

The CLT describes convergence of CDFs. This is formalized as convergence in distribution.

Definition 10.2.1. *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 of continuity of F_X. We write $X_n \xrightarrow{\mathcal{D}} X$.

Note:

1. Convergence in distribution means that the probability of intervals converge.
2. The result of the CLT can be written succinctly as $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)$.
3. In some cases a sequence of random variables converges in distribution to a constant. This is formalized in the following definition.

Definition 10.2.2. *If $X_n \xrightarrow{\mathcal{D}} X$ and $\mathrm{P}(X = c) = 1$ for some c, we say the limiting distribution of X_n is* **degenerate at c** *and write $X_n \xrightarrow{\mathcal{D}} c$.*

In other words,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) = \begin{cases} 0, & x < c \\ 1, & x \geq c \end{cases}$$

(though the limit need not exist at $x = c$).

Outline: We discuss two methods to prove convergence in distribution, direct methods in this section and via moment generating functions in Section 10.2.2. We will then use moment generating functions to prove the CLT in Section 10.3.1.

Example Suppose $(X_1, \ldots, X_n)$ is a random sample from a Uniform$(0, \alpha)$ distribution, for some $\alpha > 0$. Let $M_n = \max(X_1, \ldots, X_n)$. Derive (a) the sampling distribution of M_n and (b) the asymptotic (large n) distribution of $U_n = n(\alpha - M_n)/\alpha$.

Solution: For part (a), $M_n \leq m$ if and only if $X_i \leq m$ for $i = 1, \ldots, n$.

Then

$$F_{M_n}(m) = \mathrm{P}(M_n \leq m) = \prod_{i=1}^{n} \mathrm{P}(X_i \leq m) = \left(\frac{m}{\alpha}\right)^n$$

for any $0 \leq m \leq \alpha$ (and of course the cdf of M_n is 0 when $m < 0$ and 1 when $m > 1$).

For part (b), first note that since M_n must be between 0 and α, $U_n = n\dfrac{\alpha - M_n}{\alpha}$ must be between 0 (occurs when $M_n = \alpha$) and n (occurs when $M_n = 0$).

Now we compute the cdf of U_n.

$$\begin{aligned} F_{U_n}(u) = \mathrm{P}(U_n \le u) &= \mathrm{P}\left(\frac{n(\alpha - M_n)}{\alpha} \le u\right) = \mathrm{P}\left(M_n \ge \alpha\left(1 - \frac{u}{n}\right)\right) \\ &= 1 - F_{M_n}\left(\alpha\left(1 - \frac{u}{n}\right)\right) = 1 - \left(1 - \frac{u}{n}\right)^n \end{aligned}$$

(along with $F_{U_n}(u) = 1$ for $u \ge n$, $F_{U_n}(u) = 0$ for $u < 0$).

Taking the limit as $n \to \infty$, for any $u \ge 0$,

$$\lim_{n\to\infty} F_{U_n}(u) = 1 - \lim_{n\to\infty}\left(1 - \frac{u}{n}\right)^n = 1 - e^{-u}.$$

That is, $U_n \xrightarrow{\mathcal{D}}$ Exponential(1). ■

10.2.2 Moment Generating Functions

Moment generating functions are very useful mathematical objects. We will use them to prove convergence in distribution in general and the CLT in particular.

Definition 10.2.3. *The* **moment generating function** *(MGF) of the random variable X is*

$$M_X(t) = \mathrm{E}\left(e^{tX}\right),$$

provided the expectation exists in some neighborhood of zero, i.e., the expectation exists for all $|t| < \epsilon$ for some $\epsilon > 0$.

Theorem 10.3. *If X has a MGF, then*

$$\mathrm{E}(X^n) = M_X^{(n)}(0) \equiv \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}.$$

Note For $n = 1$,

$$\begin{aligned} \frac{d}{dt} M_X(t) = \frac{d}{dt}\int e^{tx} f_X(x)dx &= \int \frac{d}{dt} e^{tx} f_X(x)dx \\ &= \int x e^{tx} f_X(x)dx = \mathrm{E}\left(Xe^{tX}\right). \end{aligned}$$

Evaluating at $t = 0$ gives $\mathrm{E}(X)$, likewise for $n = 2, 3, \ldots$.

Note Another way of seeing this property is by writing the exponential as a power series:

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}\left(1 + tX + \frac{1}{2!}(tX)^2 + \frac{1}{3!}(tX)^3 + \cdots\right)$$
$$= 1 + \mathrm{E}(X)t + \frac{\mathrm{E}(X^2)}{2!}t^2 + \frac{\mathrm{E}(X^3)}{3!}t^3 + \cdots.$$

The coefficient of t^n can be found by differentiating n times and then setting $t = 0$.

Example Suppose $X \sim \mathrm{N}(0,1)$. Derive the MGF and the first four moments of X.

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(x^2 - 2tx\right)\right)dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(x^2 - 2tx + t^2\right)\right)\exp\left(\frac{t^2}{2}\right)dx \qquad \text{(completing the square)} \\
&= e^{t^2/2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-t)^2}{2}\right)dx \qquad \text{(integrand is pdf of a } \mathrm{N}(t,1)\text{, which integrates to 1)} \\
&= e^{t^2/2}.
\end{aligned}$$

The first four moments are,

$$\begin{aligned}
E(X) &= M_X^{(1)}(0) = \frac{d}{dt}M_X(t)\Big|_{t=0} = te^{t^2/2}\Big|_{t=0} = 0 \\
E(X^2) &= M_X^{(2)}(0) = \frac{d^2}{dt^2}M_X(t)\Big|_{t=0} = e^{t^2/2} + t^2e^{t^2/2}\Big|_{t=0} = 1 \\
E(X^3) &= M_X^{(3)}(0) = \frac{d^3}{dt^3}M_X(t)\Big|_{t=0} = te^{t^2/2} + 2te^{t^2/2} + t^3e^{t^2/2}\Big|_{t=0} = 0 \\
E(X^4) &= M_X^{(4)}(0) = \frac{d^4}{dt^4}M_X(t)\Big|_{t=0} = \left(3 + 6t^2 + t^4\right)e^{t^2/2}\Big|_{t=0} = 3
\end{aligned}$$

■

Properties of MGFs

Theorem 1: $M_{aX+b}(t) = e^{bt}M_X(at)$.

Proof: $M_{aX+b}(t) = \mathrm{E}\Big(\exp[t(aX+b)]\Big) = \mathrm{E}\Big(\exp[atX]e^{bt}\Big) = \mathrm{E}\Big(\exp[atX]\Big)e^{bt} = e^{bt}M_X(at)$.

Example Suppose $Z \sim \mathrm{N}(0,1)$ and $X = \mu + \sigma Z$. Then $X \sim \mathrm{N}(\mu, \sigma^2)$.

$$M_X(t) = e^{\mu t}M_Z(\sigma t) = e^{\mu t}e^{\sigma^2t^2/2} = e^{\mu t + \sigma^2t^2/2}.$$

Theorem 2: Let $X_1, \ldots, X_n$ be a sequence of *independent* random variables with MGFs $M_{X_1}(t), \ldots, M_{X_n}(t)$, and let $Z = X_1 + \cdots + X_n$. Then

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

Proof: $M_Z(t) = \mathrm{E}\Big(\exp[t(X_1 + \cdots + X_n)]\Big) = \mathrm{E}\left(\prod_{i=1}^{n} e^{X_i t}\right) = \prod_{i=1}^{n} \mathrm{E}\left(e^{X_i t}\right) = \prod_{i=1}^{n} M_{X_i}(t).$

Example. Suppose $X_i \overset{\text{ind}}{\sim} \text{Gamma}(\alpha_i, \beta)$ for $i = 1, \ldots, n$ and that $S = \sum_{i=1}^{n} X_i$. Then

$$M_{X_i}(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha_i} \quad \text{so that} \quad M_S(t) = \left(\frac{\beta}{\beta - t}\right)^{\sum_{i=1}^{n} \alpha_i}$$

and

$$S \sim \text{Gamma}\left(\sum_{i=1}^{n} \alpha_i,\ \beta\right).$$

Theorem 3: Let $X_1, \ldots, X_n$ be iid random variables, each with MGF, $M_X(t)$. Then

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

Proof: Apply Theorems 1 and 2.

Theorem 4: If $M_X(t)$ exists (in a neighborhood of zero), then for $r = 0, 1, 2, \ldots$

i) $M_X^{(r)}(t)$ exists near zero, and

ii) $\mathrm{E}(|X^r|) < \infty$.

Theorem 5: (Characterization) If the MGFs of X and Y exist and $M_X(t) = M_Y(t)$ in a neighborhood of zero, then

$$F_X(u) = F_Y(u) \ \text{ for all } u,$$

i.e., MGFs characterize distributions of random variables.

Example. Refer to gamma example after Theorem 2.

Theorem 6: (Convergence of MGFs) Let $X_1, X_2, \ldots$ be a countable sequence of random variables with MGFs $M_{X_1}(t), M_{X_2}(t), \ldots$ so that

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t) \text{ for } t \text{ in a neighborhood of zero,}$$

where $M_X(t)$ is a MGF. Then there is a unique CDF, F_X for which

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x) \text{ for all } x \text{ where } F_X(x) \text{ is continuous,}$$

that is, $X_n \xrightarrow{\mathcal{D}} X$. The moments of $F_X(x)$ are determined by $M_X(t)$

Proofs of Theorems 4–6 follow from the properties of Laplace transforms (but are beyond the scope of this course).

Notes on MGFs:

(1) We can show convergence in distribution by showing convergence of MGF in a neighborhood of zero.

(2) Convergence of MGF is a sufficient, but not necessary condition for convergence in distribution. (The MGF may not exist in a neighborhood of zero.)

(3) A more general strategy involves characteristic functions, $\phi_X(t) = \mathrm{E}(e^{itX})$.

(4) Moments alone do not characterize distributions. That is, there exist X and Y such that $\mathrm{E}(X^r) = \mathrm{E}(Y^r)$ for $r = 0, 1, 2, \ldots$, but $F_X \neq F_Y$. (See example 2.3.10 in Casella & Berger, *Statistical Inference*, 2nd edition, 2008.)

(5) However, *if X and Y have finite support,* and all moments exist, then $F_X(u) = F_Y(u)$ for all u if and only if $\mathrm{E}(X^r) = \mathrm{E}(Y^r)$ for $r = 0, 1, \ldots$.

10.3.1 Proof of the Central Limit Theorem

Theorem 10.4. *Let $X_1, X_2, \ldots$ be a sequence of iid random variables, each with MGF, $M_X(t)$.*

Then

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

where $G_n(x)$ is the CDF of $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$. That is, $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)$.

Note: By assumption $M_X(t)$ exists in a neighborhood of zero, such that $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ exist.

Proof: Let $Z_n = \dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$. It is sufficient to show that

$$\lim_{n\to\infty} M_{Z_n}(t) = M_Z(t) = e^{t^2/2},$$

where $Z \sim \mathrm{N}(0,1)$.

Let $W_i = (X_i - \mu)/\sigma$ so that $\mathrm{E}(W_i) = 0$, $\mathrm{Var}(W_i) = \mathrm{E}(W_i^2) = 1$, and

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{W_1 + \cdots + W_n}{\sqrt{n}}.$$

Using Taylor's theorem, expanding around $t = 0$,

$$\begin{aligned} M_{W_i}(t) = \sum_{k=0}^{\infty} \frac{M_W^{(k)}(0)}{k!} t^k &= M_W^{(0)}(0) + M_W^{(1)}(0)t + \frac{1}{2} M_W^{(2)}(r)t^2 \quad \text{for some r with } |r| < t. \\ &= 1 + 0t + \frac{1}{2} M_W^{(2)}(r)t^2. \qquad \begin{array}{l}\text{(subscript } i\text{'s are absent since every } W_i \text{ has}\\ \text{the same MGF)}\end{array} \end{aligned}$$

Now we write the MGF of Z_n, which is the sum of the n independent RVs $W_i/\sqrt{n}$.

$$M_{Z_n}(t) = \prod_{i=1}^{n} M_{W_i}\left(\frac{t}{\sqrt{n}}\right) = \left[M_{W_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 + \frac{t^2}{2n} M_W^{(2)}(r_n)\right]^n \quad \text{for some } r_n \text{ with } |r_n| < \frac{t}{\sqrt{n}},$$

hence,

$$\log\left(M_{Z_n}(t)\right) = n \log\left(1 + \frac{t^2}{2n} M_W^{(2)}(r_n)\right). \tag{10.1}$$

We now take the limit as n goes to infinity for fixed t. As $n \to \infty$, $r_n \to 0$. Because $M_W^{(3)}(t)$ exists we know that $M_W^{(2)}(t)$ is continuous and so $M_W^{(2)}(r_n) \to M_W^{(2)}(0) = \mathrm{E}(W_i^2) = 1$.

We see in Eq. 10.1 that we have a factor that looks like $\log\left(1 + \frac{\text{const}}{n}\right)$ as n increases. We can again use a Taylor series to expand $\log(1+x)$ for small x: $\log(1+x) = x + \epsilon(x)x$, where $\lim_{x\to 0} \epsilon(x) = 0$.

Putting this together we have

$$\begin{aligned}\log\left(M_{Z_n}(t)\right) &= n\log\left(1+\frac{t^2}{2n}M_W^{(2)}(r_n)\right)\\ &= n\left(\frac{t^2}{2n}M_W^{(2)}(r_n)+\epsilon\left(\delta_n\right)\delta_n\right)\\ &= \frac{t^2}{2}M_W^{(2)}(r_n)+\epsilon\left(\delta_n\right)n\delta_n,\end{aligned}$$

where $\delta_n = \frac{t^2}{2n}M_W^{(2)}(r_n)$. Notice that as $n \to \infty$ we have $\delta_n \to 0$, thus $\epsilon(\delta_n) \to 0$, and $n\delta_n \to$ constant.

This shows that the MGF of Z_n converges to a limit as $n \to \infty$:

$$\lim_{n\to\infty}\log\left(M_{Z_n}(t)\right) = \frac{t^2}{2},$$

and so

$$\lim_{n\to\infty}M_{Z_n}(t) = \exp\left(\frac{t^2}{2}\right),$$

the MGF of a standard normal. Therefore $Z_n = \dfrac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)$.

10.5 Convergence in Probability and Inequalities

10.5.1 Convergence in Probability

Definition 10.5.1. *A sequence of random variables,* $X_1, X_2, \ldots$, **converges in probability** *to the random variable* X *if* $\forall \epsilon > 0$,

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

We write $X_n \xrightarrow{\mathcal{P}} X$.

Example Consider the random variable $X \sim \mathrm{N}(0,1)$ and the sequence of RVs $X_n = X + Y_n$, where $Y_n \sim \mathrm{N}(0, 1/n)$ and the Y_k are independent. It is clear that as n increases the probability that X_n is close to X will become large.

In particular, for any $\epsilon > 0$,

$$\mathrm{P}\left(|X_n - X| \geq \epsilon\right) = \mathrm{P}\left(|Y_n| \geq \epsilon\right) = \mathrm{P}\left(Y_n \geq \epsilon \text{ or } Y_n \leq -\epsilon\right) = 2\Phi\left(-\sqrt{n}\epsilon\right) \longrightarrow 0 \quad \text{as } n \to \infty,$$

and therefore $X_n \xrightarrow{\mathcal{P}} X$.

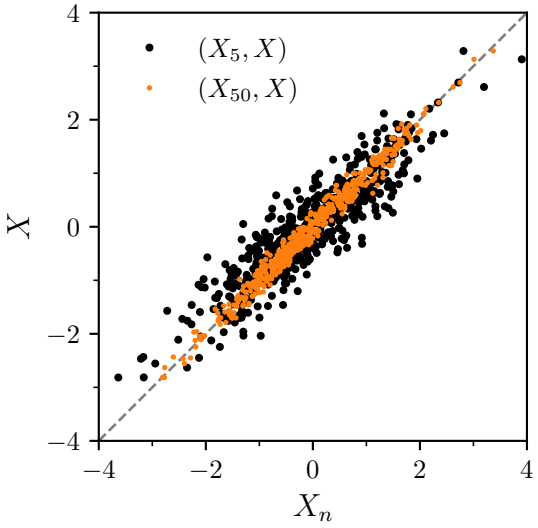


Figure 10.1: 500 samples of X, X_5, and X_{50}.

■

Example Convergence in distribution does not imply convergence in probability. Suppose $X \sim \mathrm{N}(0,1)$ and let

$$X_n = -X, \quad \text{for } n = 1, 2, 3, \ldots$$

Then $X_n \sim \mathrm{N}(0,1)$, and trivially, $X_n \xrightarrow{\mathcal{D}} X$. But

$$\mathrm{P}\left(|X_n - X| \geq \epsilon\right) = \mathrm{P}\left(|2X| \geq \epsilon\right) = \mathrm{P}\left(|X| \geq \frac{\epsilon}{2}\right),$$

which does not converge to zero as n grows. Thus X_n does not converge to X in probability.

■

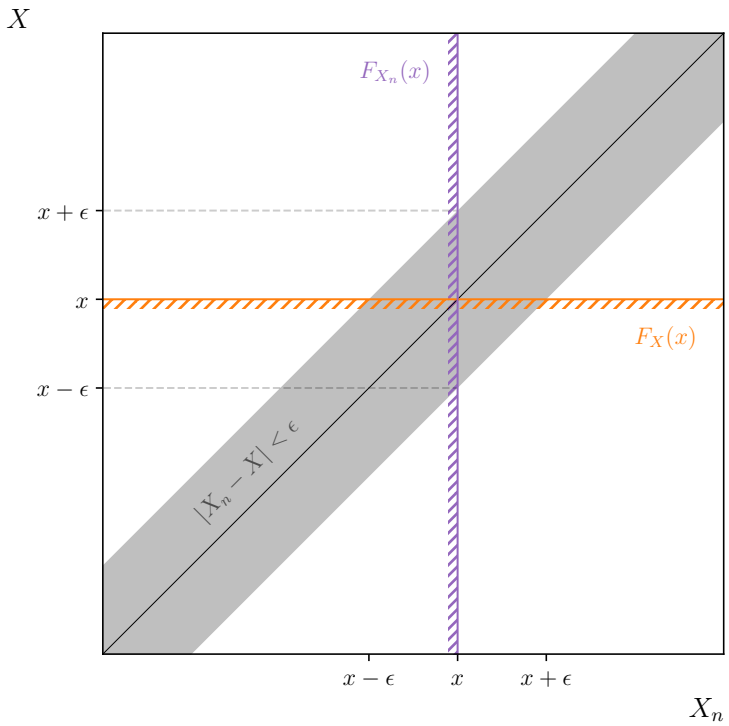
Convergence in distribution says that for large enough n, the cdf/pdf/pmf of X_n is very close to that of X. But it doesn't say anything about whether the value of X_n will be close to X for a single run of the experiment (i.e. when nature selects a single outcome from the sample space).

Convergence in probability is stronger than convergence in distribution. It says that for large enough n, *for each run of the experiment* there is a high probability that the two values, X_n and X, will be close together.

In fact, $X_n \xrightarrow{\mathcal{P}} X$ implies that $X_n \xrightarrow{\mathcal{D}} X$. This is formalized in the following theorem.

Theorem 10.6. *Convergence in probability implies convergence in distribution.*

Proof: We want to show that $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for any x at which $F_X(\cdot)$ is continuous. The figure below, showing the joint sample space of the RVs (X_n, X), motivates the proof.



$F_X(x) = \mathrm{P}(X \leq x)$ is the probability of the region below the orange line. Similarly, $F_{X_n}(x)$ is the probability of the region to the left of the purple line. For any ϵ the probability that

(X_n, X) is outside the gray region, corresponding to $|X_n - X| \geq \epsilon$, goes to 0 as $n \to \infty$. As we take $\epsilon \to 0$ we see that both $F_X(x)$ and $F_{X_n}(x)$ will both approach the probability of being in the lower-left quadrant $\{X_n \leq x \text{ and } X \leq x\}$, i.e. they are equal.

To make the proof precise, we partition the (X_n, X) plane into a few useful regions:

$$F_{X_n}(x) = \mathrm{P}(X_n \leq x) = \mathrm{P}(X_n \leq x \text{ and } X \leq x) + \mathrm{P}(X_n \leq x \text{ and } X > x),$$
$$F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(X_n \leq x \text{ and } X \leq x) + \mathrm{P}(X_n > x \text{ and } X \leq x).$$

This lets us write the difference $F_{X_n}(x) - F_X(x)$ in terms of two probabilities that will go to zero in the appropriate limits:

$$F_{X_n}(x) - F_X(x) = \mathrm{P}(X_n \leq x \text{ and } X > x) - \mathrm{P}(X_n > x \text{ and } X \leq x),$$
$$\Rightarrow |F_{X_n}(x) - F_X(x)| \leq \mathrm{P}(X_n \leq x \text{ and } X > x) + \mathrm{P}(X_n > x \text{ and } X \leq x). \tag{10.2}$$

For each of the regions on the righthand side we can partition it as the sum of a piece inside the gray region (where $|X_n - X| < \epsilon$) plus a piece outside (where $|X_n - X| \geq \epsilon$). For instance,

$$\begin{aligned}\mathrm{P}(X_n \leq x \text{ and } X > x) &= \mathrm{P}(X_n \leq x,\, X > x, \text{ and } |X_n - X| \geq \epsilon) \quad &+ \mathrm{P}(X_n \leq x,\, X > x, \text{ and } |X_n - X| < \epsilon) \\ &\leq \mathrm{P}(|X_n - X| \geq \epsilon) \quad &+ \mathrm{P}(x < X < x + \epsilon),\end{aligned}$$

where the inequalities follow from $A \subseteq B \Rightarrow \mathrm{P}(A) \leq \mathrm{P}(B)$ (with A and B being the regions of the (X_n, X) plane stated).

We can make the first term on the right arbitrarily close to zero by letting $n \to \infty$ (since $X_n \xrightarrow{\mathcal{P}} X$). We can make the second term on the right arbitrarily close to zero by taking $\epsilon \to 0$ because F_X is continuous at x. Specifically $\mathrm{P}(x < X < x + \epsilon) \leq \mathrm{P}(x < X \leq x + \epsilon) = F_X(x + \epsilon) - F_X(x) \to 0$ as $\epsilon \to 0$.

The above argument applies to both terms on the righthand side of Eq. 10.2. We have thus shown that $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all points of continuity of F_X, which is just the definition that X_n converges in distribution to X.

Notes:

(1) Although convergence in distribution does not *generally* imply convergence in probability, $X_n \xrightarrow{\mathcal{D}} c$ implies $X_n \xrightarrow{\mathcal{P}} c$. That is, a sequence of random variables convergence in distribution to a constant if and only if it convergence in probability to that constant.

(2) (*Slutsky's Theorem:*) If $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{\mathcal{D}} c$, then

$$X_n Y_n \xrightarrow{\mathcal{D}} cX \text{ and } X_n + Y_n \xrightarrow{\mathcal{D}} X + c.$$

10.6.1 The Law of Large Numbers and Chebyshev's Ineqaulity

Theorem 10.7. *(Weak Law of Large Numbers,* WLLN*) Suppose $X_1, X_2, \ldots$ is a sequence of iid random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. If $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\overline{X}_n \xrightarrow{\mathcal{P}} \mu$, i.e.,*

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \mathrm{P}\left(\left|\overline{X}_n - \mu\right| < \epsilon\right) = 1.$$

Note: Recall that we say the statistic T_n is a **consistent** estimator of θ if $T_n \xrightarrow{\mathcal{P}} \theta$. Thus the WLLN says that $\overline{X}_n$ is a consistent estimator of μ.

The proof of the Weak Law of Large Numbers relies on

Theorem 10.8. *(Chebyshev's Inequality) Let X be a random variable and $g(\cdot)$ be a non-negative function. Then for any $r > 0$,*

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

Proof: (of Chebyshev's inequality) If X is a continuous RV,

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx \quad \geq \int_{\{x\,:\,g(x)\geq r\}} g(x) f_X(x)dx \quad \geq r\int_{\{x\,:\,g(x)\geq r\}} f_X(x)dx = \mathrm{P}\Big(g(X) \geq r\Big).$$

If X is a discrete RV the proof goes the same way, just replace integrals with sums and the pdf with the pmf.

Proof: (of the WLLN)

We apply Chebyshev's inequality to the RV $\overline{X}$ and non-negative function $g(x) = (x-\mu)^2$. The expectation of $g(\overline{X})$ is just the variance of the sample mean, which is σ^2/n. Then,

$$\mathrm{P}\left(|\overline{X} - \mu| \geq \epsilon\right) = \mathrm{P}\left((\overline{X} - \mu)^2 \geq \epsilon^2\right) \leq \frac{E\left[(\overline{X} - \mu)^2\right]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{as } n \to \infty$$

This holds for any $\epsilon > 0$. Thus we have $\overline{X} \xrightarrow{\mathcal{P}} \mu$.

10.8.1 Jensen's Ineqaulity

Question: We know $\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$, generally. But what more can we say?

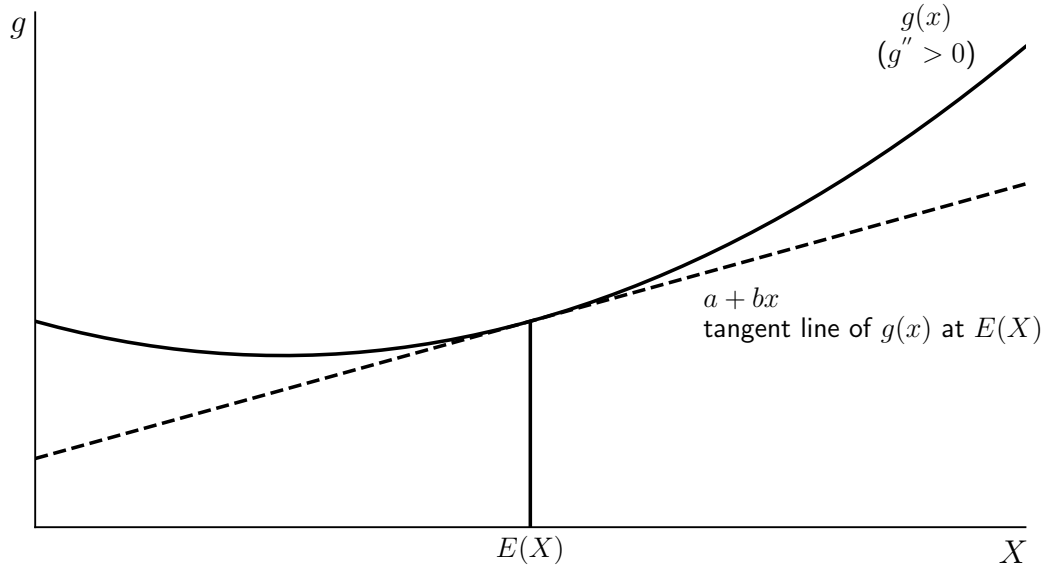
Theorem 10.9. *(Jensen's Inequality)*

i) If $g(x)$ is a convex function ($g'' \geq 0$), then $\mathrm{E}[g(X)] \geq g[E(X)]$.

ii) If $g(x)$ is a concave function ($g'' \leq 0$), then $\mathrm{E}[g(X)] \leq g[E(X)]$.

In either case, equality holds if and only if $\mathrm{P}\Big(g(X) = a + bX\Big) = 1$ for some a and b, i.e., if g is a linear function on the support of X.

Proof: (Graphical)



Consider the tangent line to $g(x)$ at $x = \mathrm{E}(X)$, whose equation is $a + bx$. If g is convex then the curve $g(x)$ lies above every tangent line, so $g(x) \geq a + bx$ for all x. Taking expectations of both sides:

$$g(x) \geq a + bx, \ \forall x \quad \Longrightarrow \quad \mathrm{E}[g(X)] \geq \mathrm{E}[a + bX] = a + b\mathrm{E}[X] = g\Big(\mathrm{E}(X)\Big),$$

and equality holds if and only if $\mathrm{P}\Big(g(X) = a + bX\Big) = 1$ (i.e. the set of x s.t. $g(x) \neq a + bx$ has probability 0). An analagous argument applies to concave $g(x)$.

Say $\mathrm{E}(h(X)) = 0$ for some non-negative function $h(x)$. Then, $0 = \mathrm{E}(h(X)) = \sum_{x_i \in \mathbb{X}} h(x_i) p_X(x_i)$. Since each term of this sum is non-negative, each term must be exactly 0.

Therefore,

$$\mathrm{E}(h(X)) = 0 \iff h(x_i) = 0 \text{ or } p_X(x_i) = 0, \ \forall x_i \iff \mathrm{P}(h(X) = 0) = \sum_{\{x_i : h(x_i) = 0\}} p_X(x_i) = 1 - \sum_{\{x_i : h(x_i) \neq 0\}} p_X(x_i) = 1.$$

Setting $h(x) = g(x) - (a + bx)$ gives the necessary and sufficient conditions for Jensen's inequality to be an exact equality. For continuous X the proof is along the same lines.

Example Suppose $X \sim \text{Binomial}(n, p)$ and consider the estimator $\hat{P} = \frac{1}{n}X$ of p.

$\text{E}(\hat{P}) = \text{E}\left(\frac{1}{n}X\right) = \frac{1}{n}\text{E}(X) = \frac{1}{n}np = p$, so $\hat{P}$ is an unbiased estimator of p.

Suppose we want an estimator of the odds ratio $\xi = \frac{p}{1-p} = g(p)$.

$$\frac{d}{dp}g(p) = (1-p)^{-2}$$

and

$$\frac{d^2}{dp^2} = 2(1-p)^{-3} > 0$$

when $p \in [0, 1]$.

Consider the estimator $\hat{\Xi} = \frac{\hat{P}}{1-\hat{P}}$ of ξ.

$$\text{E}(\hat{\Xi}) = \text{E}\left(g(\hat{P})\right) > g\left(\text{E}(\hat{P})\right) = \text{E}(p) = \xi,$$

where the inequality comes from Jensen's inequality for the convex function $g(p)$ (looking back at the proof of Jensen's inequality, we only need that g be convex on the range of X; in this case $\hat{P}$ is always in $[0, 1]$).

So $\text{E}\left(\hat{\Xi}\right) > \xi$ because $\frac{p}{1-p}$ is not linear on $[0, 1]$. In other words, $\hat{\Xi}$ is an *upwardly (positively) biased estimator* of ξ.

(In fact, in this case $\text{E}(\hat{\Xi}) = \infty$ because of the finite possibility of $X = n$.) ■

In general if we have an unbiased estimator $\hat{\theta}$ of parameter θ and want to estimate some function of the parameter $\phi = g(\theta)$ using the estimator $\hat{\phi} = g(\hat{\theta})$ it is important to realize that

$$\text{E}(\hat{\phi}) = \text{E}(g(\hat{\theta})) \neq g(\text{E}(\hat{\theta})) = g(\theta) = \phi,$$

i.e. unbiasedness is not necessarily invariant to transformation.