CIS6930/4930

Random Number Generation, Moment Generating Functions, Characteristic Functions, Central Limit Theorem

Prof. Ye Xia

Generate Samples from a Distribution

The first step is to generate random integers chosen uniformly at random from a range. There are very good algorithms and implementations on computers. After scaling, we have a random number generator for the uniform distribution on $(0, 1)$.

For a uniform random variable U on $(0, 1)$, $P(U \leq u) = u$, where $u \in (0, 1)$.

Drawing samples from other distributions (e.g., exponential) can be done by 'inverting' the uniform distribution.

Suppose that we want to draw samples from the distribution function $F(x)$. Let U be a sample drawn from the uniform distribution on $(0, 1)$. Define

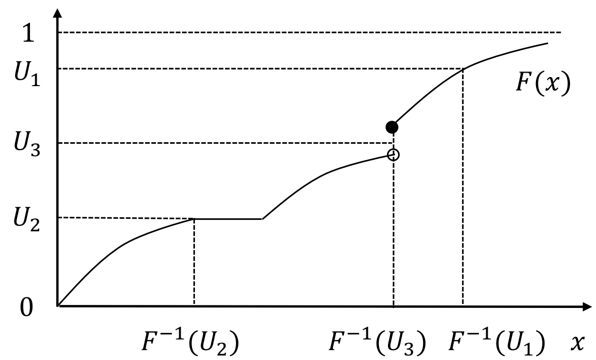
$$F^{-1}(U) = \inf\{x : F(x) \geq U\} = \sup\{x : F(x) < U\}.$$

Let $X = F^{-1}(U)$. Then, X has the distribution $F(x)$. Why?

For convenience, let $S = \{x : F(x) \geq U\}$.

Claim: $F(F^{-1}(U)) \geq U$ (see Figure), and therefore, $F^{-1}(U) \in S$ and

$F^{-1}(U) = \min S$.



Comment: An infimum of a set $S \subseteq \mathbb{R}$, denoted by $\inf S$, is also called the **greatest lower bound** of the set S. That is, it is a lower bound of S that cannot be improved. Suppose $v = \inf S$. Then, (i) $x \geq v$ for every $x \in S$; and (ii) for any $\epsilon > 0$, there exists $x \in S$ such that $x < v + \epsilon$.

An infimum is like a minimum except that the infimum z may not belong to the set S, whereas the minimum must be an element of S. If the set S is bounded from below, it will always have an infimum, but not necessarily a minimum. For instance, for the set $(0, 1)$, the minimum does not exist, but the infimum exists and it is 0.

Proof of Claim: $F^{-1}(U)$ is the infimum of the set $S = \{x : F(x) \geq U\}$. This means (i) $x \geq F^{-1}(U)$ for every $x \in S$; and (ii) for any $\epsilon > 0$, there exists $x \in S$ such that $x < F^{-1}(U) + \epsilon$.

Now, consider a sequence $1/n$, $n = 1, 2, \ldots$. For each n, there exists $x_n \in S$ (implying $F(x_n) \geq U$ and $x_n \geq F^{-1}(U)$) such that $x_n < F^{-1}(U) + 1/n$.

As $n \to \infty$, $x_n \downarrow F^{-1}(U)$.

Since F is right continuous, $F(x_n) \to F(F^{-1}(U))$ as $n \to \infty$. Now, for each n, $F(x_n) \geq U$. Taking $n \to \infty$, we have $F(F^{-1}(U)) \geq U$.

By the definition of the set S, we have $F^{-1}(U) \in S$. Hence, $F^{-1}(U)$ is in fact the minimum of S.

Corollary: $X \leq x \iff U \leq F(x)$.

Suppose $X \leq x$. Since F is non-decreasing, $F(X) \leq F(x)$. But,

$F(X) = F(F^{-1}(U)) \geq U$ by the claim above. Therefore, $U \leq F(X) \leq F(x)$.
Conversely, if $U \leq F(x)$, then $x \in S$ and therefore $x \geq X = \min S$.

Finally, we have

$$P(X \leq x) = P(U \leq F(x)) = F(x),$$

that is, X has the distribution function $F(x)$.

Moment Generating Functions

Given a random variable X with the distribution function F, the moment generating function of X, denoted by $M(t)$ (or $M_X(t)$), is

$$M(t) = E[e^{tX}] = \int_{\mathbb{R}} e^{tx} dF(x) = \begin{cases} \sum_x e^{tx} P(X = x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} e^{tx} f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

In the above, f is the pdf of X when X is a continuous random variable.

Simple fact: $M(0) = 1$ and $M(t) > 0$ for all t where it is defined.

We are mostly interested in $M(t)$ on some interval containing 0. If there is one such interval for which $M(t) < \infty$, then

- (i) $E[X] = M'(0)$, $E[X^k] = M^{(k)}(0)$ for all integer $k \geq 1$.
- (ii) $M(t) = \sum_{k=0}^{\infty} \frac{E[X^k]}{k!} t^k$ (Taylor's Theorem).

- (iii) If X and Y are independent random variables, then

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Note: $M^{(k)}$ denotes the k-th derivative of M.

Part (ii) shows what a generating function does. If there is a sequence of numbers $a_0, a_1, \ldots$, a generating function of the form $G(s) = \sum_{k=0}^{\infty} a_k s^k$ encodes the numbers into a function.

We see that $M(t)$ encodes all the moments of X.

To see part (i),

$$M'(t) = \frac{d}{dt}E[e^{tX}] = E[\frac{d}{dt}e^{tX}] = E[Xe^{tX}].$$

Then, $M'(0) = E[X]$.

$$M''(t) = \frac{d}{dt}M'(t) = \frac{d}{dt}E[Xe^{tX}] = E[\frac{d}{dt}(Xe^{tX})] = E[X^2e^{tX}].$$

Then, $M''(0) = E[X^2]$.

The interchange of expectation with derivative requires justification. The condition that $M(t) < \infty$ on some interval containing 0 is enough to justify the interchange (Casella and Berger, p70); it also ensures $M^{(k)}(0) = E[X^k]$ is finite for every positive integer k.

Part (i) implies the Taylor expansion in part (ii).

Consequence: Recall the Tayler expansion for e^{tX} and for $M(t)$:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$$

$$M(t) = M(0) + \frac{M'(0)}{1!}t + \frac{M''(0)}{2!}t^2 + \frac{M'''(0)}{3!}t^3 + \cdots$$

Part (i) implies

$$M(t) = E[e^{tX}] = 1 + tE[X] + \frac{t^2E[X^2]}{2!} + \frac{t^3E[X^3]}{3!} + \cdots$$

We see that the interchange of the expectation and sum is correct in this case.

To see part (iii):

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t).$$

Part (iii) shows why a moment generating function is useful. We often apply it to the sum of independent random variables.

Also, it is sometimes easier to compute the moments by first compute the moment generating function.

Examples

Example: X is an exponential random variable with parameter λ.

$$\begin{aligned} M(t) =& E[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t} \int_0^\infty (\lambda - t) e^{-(\lambda - t)x} dx \\ =& \frac{\lambda}{\lambda - t}, \quad \text{for } t < \lambda. \end{aligned}$$

$$M'(t) = \frac{\lambda}{(\lambda - t)^2}, \quad M''(t) = \frac{2\lambda}{(\lambda - t)^3}.$$

Hence,

$$E[X] = M'(0) = \frac{1}{\lambda}, \quad E[X^2] = M''(0) = \frac{2}{\lambda^2}.$$

Thus, the variance of X is

$$\text{var}(X) = E[X^2] - (E[X])^2 = \frac{1}{\lambda^2}.$$

Example: Suppose X is a Poisson random variable with mean λ. It can be shown that $M_X(t) = \exp(\lambda(e^t - 1))$.

If Y is another Poisson random variable with mean μ, and X and Y are independent of each other, then $M_Y(t) = \exp(\mu(e^t - 1))$, and $M_{X+Y}(t) = \exp((\lambda + \mu)(e^t - 1))$.

We see that $X + Y$ is also a Poisson random variable with mean $\lambda + \mu$.

Fact: If $a, b \in \mathbb{R}$ and $Y = aX + b$, then

$$M_Y(t) = E[e^{tY}] = E[e^{t(aX+b)}] = E[e^{atX+bt}] = e^{bt}E[e^{atX}] = e^{bt}M_X(at).$$

Example: Suppose X has the $\mathcal{N}(0, 1)$ distribution. Then, it can be shown that $M_X(t) = e^{t^2/2}$.

Suppose Y has the $\mathcal{N}(\mu, \sigma^2)$ distribution. Then, $Y \stackrel{\mathrm{d}}{=} \sigma X + \mu$, i.e., the two sides have the same distribution. Why? Because the right hand side has the distribution $\mathcal{N}(\mu, \sigma^2)$. Therefore, $M_Y(t) = e^{\mu t}M_X(\sigma t) = \exp(\frac{\sigma^2 t^2}{2} + \mu t)$.

Example (Sum of Independent Gaussian RVs): Suppose X and Y are independent Gaussian random variables with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2). Then, $X + Y$ is Gaussian with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

We have $M_X(t) = \exp(\frac{\sigma_1^2 t^2}{2} + \mu_1 t)$ and $M_Y(t) = \exp(\frac{\sigma_2^2 t^2}{2} + \mu_2 t)$. Since $M_{X+Y}(t) = M_X(t) M_Y(t)$, we have

$$M_{X+Y}(t) = \exp\left(\frac{(\sigma_1^2 + \sigma_2^2)t^2}{2} + (\mu_1 + \mu_2)t\right).$$

The right hand side is the moment generating function of a Gaussian random variable with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. In this case, the moment generating function uniquely determines the distribution function (see later). Hence, $X + Y$ has the distribution $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Corollary: Any linear combination of independent Gaussian random variables is a Gaussian random variable.

Joint Moment Generating Functions

Let $X = (X_1, \ldots, X_n)$ be a random vector, i.e., a collection of random variables defined on a common probability space. The joint moment generating function $M_X(t_1, \ldots, t_n)$, where each $t_i \in \mathbb{R}$, is defined by

$$M_X(t_1, \ldots, t_n) = E[e^{t_1 X_1 + \cdots + t_n X_n}].$$

Example (Multivariate Gaussian/Normal Distribution: Let $Z_1, \ldots, Z_n$ be a set of n independent standard Gaussian random variables with distribution $\mathcal{N}(0, 1)$. Let $a_{ij} \in \mathbb{R}$ and $\mu_i \in \mathbb{R}$ be constants, where $1 \leq i \leq m$ and $1 \leq j \leq n$. Let

$$\begin{aligned}
X_1 &= a_{11} Z_1 + \cdots + a_{1n} Z_n + \mu_1 \\
X_2 &= a_{21} Z_1 + \cdots + a_{2n} Z_n + \mu_2 \\
&\vdots \\
X_m &= a_{m1} Z_1 + \cdots + a_{mn} Z_n + \mu_m.
\end{aligned}$$

Then, the variables $X_1, \ldots, X_m$ are said to have a multivariate Gaussian/nomral distribution.

Each X_i is a sum of independent Gaussian random variables. By the earlier corollary, each

X_i is a Gaussian random variable with $E[X_i] = \mu_i$ and $\mathrm{var}(X_i) = \sum_{j=1}^n a_{ij}^2$.

We will calculate the joint moment generating function. First, $\sum_{i=1}^m t_i X_i$ is a Gaussian random variable, because it is also a sum of independent Gaussian random variables (the Z_i's). We only need to compute its mean and variance.

$$E[\sum_{i=1}^m t_i X_i] = \sum_{i=1}^m t_i \mu_i.$$

For the variance of $\sum_{i=1}^m t_i X_i$, we need a few facts about covariance. For generic random variables U, V, W and Z,

$$\mathrm{cov}(Z, Z) = \mathrm{var}(Z),\ \mathrm{cov}(W + a, Z + b) = \mathrm{cov}(W, Z) = \mathrm{cov}(Z, W)$$

The next one is bilinearity (which implies multi-linearity):

$$\mathrm{cov}(aU+bV, cW+dZ) = ac\,\mathrm{cov}(U, W)+ad\,\mathrm{cov}(U, Z)+bc\,\mathrm{cov}(V, W)+bd\,\mathrm{cov}(V, Z).$$

Then, we have

$$\begin{aligned}
\mathrm{var}(\sum_{i=1}^{m} t_i X_i) =& \mathrm{cov}(\sum_{i=1}^{m} t_i X_i, \sum_{i=1}^{m} t_i X_i) \\
=& \sum_{i=1}^{m}\sum_{j=1}^{m} t_i t_j \mathrm{cov}(X_i, X_j) \qquad \text{(used multi-linearity)} \\
=& \sum_{i=1}^{m} t_i^2 \mathrm{var}(X_i) + 2\sum_{i<j} t_i t_j \mathrm{cov}(X_i, X_j).
\end{aligned}$$

We can go further. For any pair of i and j, $\mathrm{cov}(X_i, X_j)$ is,

$$\begin{aligned}
\mathrm{cov}(X_i, X_j) =& \mathrm{cov}(\sum_{k=1}^{n} a_{ik} Z_k + \mu_i, \sum_{k=1}^{n} a_{jk} Z_k + \mu_j) \\
=& \mathrm{cov}(\sum_{k=1}^{n} a_{ik} Z_k, \sum_{k=1}^{n} a_{jk} Z_k) \\
=& \sum_{k=1}^{n} a_{ik} a_{jk} \mathrm{var}(Z_k) = \sum_{k=1}^{n} a_{ik} a_{jk}.
\end{aligned}$$

In the second to the last step, we used the fact that when $k \neq l$, Z_k and Z_l are independent, and therefore, $\mathrm{cov}(Z_k, Z_l) = 0$.

Therefore,

$$\mathrm{var}(\sum_{i=1}^{m} t_i X_i) = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n} t_i t_j a_{ik} a_{jk}.$$

For a random variable Z with the distribution $\mathcal{N}(\mu, \sigma^2)$, the moment generating function is $M_Z(s) = E[e^{sZ}] = \exp(\frac{\sigma^2 s^2}{2} + \mu s)$. Then,

$$E[e^Z] = M_Z(1) = \exp(\frac{\sigma^2}{2} + \mu). \tag{1}$$

Note: The random variable e^Z has a log-normal distribution.

The joint moment generating function for $X = (X_1, \ldots, X_m)$ is $M_X(t_1, \ldots, t_m) = E[e^{\sum_{i=1}^{m} t_i X_i}]$.

Since $\sum_{i=1}^{m} t_i X_i$ is a Gaussian random variable, we only need to substitute its mean and variance into the expression in (1).

$$\begin{aligned} M_X(t_1, \ldots, t_m) &= \exp\left(\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} t_i t_j \mathrm{cov}(X_i, X_j) + \sum_{i=1}^{m} t_i \mu_i\right) \\ &= \exp\left(\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n} t_i t_j a_{ik} a_{jk} + \sum_{i=1}^{m} t_i \mu_i\right). \end{aligned}$$

Under the right conditions, the joint moment generating function determines the distribution function uniquely. We will later state such a result for the single-variable case.

Then, the first equality implies that the joint distribution depends only on the means and covariances, i.e., $E[X_i]$ and $\mathrm{cov}(X_i, X_j)$ for $i, j = 1, \ldots, m$.

Important Conclusion: To specify a multivariate Gaussian, it is enough to specify the means and the covariance matrix $(\mathrm{cov}(X_i, X_j))_{1 \le i,j \le m}$.

Suppose the multivariate Gaussian $X = (X_1, \ldots, X_n)$ have the mean vector μ and covariance matrix Σ and suppose Σ is invertible. The joint pdf of X is

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right),$$

where $|\Sigma|$ is the determinant of the matrix Σ.

Conversely, if a random vector $X = (X_1, \ldots, X_n)$ has a joint pdf of the above form and Σ is a positive definite matrix, then X is multivariate Gaussian, μ is its mean vector and Σ is the covariance matrix.

More on Moment Generating Functions

Theorem: Suppose $M(t)$ exists (and hence, finite) at $\pm t$. Then, $M(s) < \infty$ for all $s \in (-t, t)$, and $M^{(k)}(0) = E[X^k] < \infty$ for every positive integer k.

Proof: Recall we have the background probability space $(\Omega, \mathcal{F}, P)$. The random variables $1_{(X>0)}$ and $1_{(X\leq 0)}$ are measurable functions on Ω. Note that $1_{(X>0)} + 1_{(X\leq 0)}$ is the constant function whose value is always equal to 1.

For any $s \in (-t, t)$,

$$\begin{aligned} M(s) =& E[e^{sX}] = E[e^{sX}1_{(X>0)} + e^{sX}1_{(X\leq 0)}] \\ =& E[e^{sX}1_{(X>0)}] + E[e^{sX}1_{(X\leq 0)}] \\ \leq& E[e^{tX}1_{(X>0)}] + E[e^{-tX}1_{(X\leq 0)}] \\ \leq& E[e^{tX}] + E[e^{-tX}] < \infty. \end{aligned}$$

Next, given $k \geq 1$, choose s such that $0 < ks < t$. Because $s|X| < e^{s|X|}$,

$s^k|X|^k \le e^{ks|X|} \le e^{ksX} + e^{-ksX}$. Then,

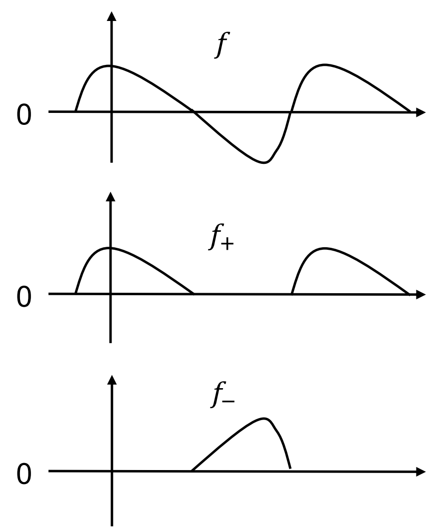
$$E[|X|^k] \le s^{-k}E[e^{ksX} + e^{-ksX}] = s^{-k}(M(ks) + M(-ks)) < \infty.$$

Note: $E[|X|^k] < \infty$ implies $E[X^k]$ is finite.

In general, for any function f, let

$$f_+ = \max(f, 0), \quad f_- = \max(-f, 0).$$

f_+ is called the positive part of f and f_- is called the negative part of f. Note that $f_- \ge 0$.



Then,

$$|f| = f_+ + f_-, \quad f = f_+ - f_-.$$

We have $\int |f| d\mu = \int f_+ d\mu + \int f_- d\mu$.

If $\int |f| d\mu < \infty$, then $\int f_+ d\mu < \infty$ and $\int f_- d\mu < \infty$. Therefore,

$$\int f d\mu = \int f_+ d\mu - \int f_- d\mu < \infty.$$

Another fact:

$$\begin{aligned} E[|X|^k] =& \int_0^\infty P(|X|^k > x) dx \\ =& \int_0^\infty P(|X|^k > y^k) k y^{k-1} dy \qquad \text{(let } x = y^k\text{)} \\ =& k \int_0^\infty P(|X| > y) y^{k-1} dy \end{aligned}$$

Therefore, to have have $E[|X|^k] < \infty$, the tail probability $P(|X| > y)$ must decay faster than $1/y^k$. More precisely, if there exists C such that

$P(|X| > y) \geq 1/y^k$ for all $y > C$, then $E[|X|^k]$ cannot be finite. On the other hand, if there exists C and $\delta > 0$ such that $P(|X| > y) < 1/y^{k+\delta}$ all $y > C$, then $E[|X|^k] < \infty$.

Can the moments determine the distribution? Not always. It can happen that two different distribution functions have the same moments of any order.

Does a moment generating function uniquely characterize the distribution function? Not always. But, the following results hold.

Theorem: Let X and Y be random variables with distribution functions F_X and F_Y, of which all the moments exist.

(i) If X and Y are bounded, then $F_X(u) = F_Y(u)$ for all u if and only if $E[X^n] = E[Y^n]$ for all integers $n \geq 0$.

(ii) If M_X and M_Y exist (i.e., well-defined and finite) and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.

Note: Under the condition of moment generating functions exist in a neighborhood V of 0, part (ii) implies

$$M_X(t) = M_Y(t), \forall t \in V \Leftrightarrow F_X(u) = F_Y(u), \forall u \Leftrightarrow E[X^n] = E[Y^n], \forall n \geq 0.$$

To see $E[X^n] = E[Y^n], \forall n \geq 0 \Rightarrow M_X(t) = M_Y(t)$, use Taylor's expansion of MGF.

Part (ii) implies part (i), since the moment generating function of a bounded random variable exists for all t.

The proof of (ii) relies on the theory of Laplace Transform.

Characteristic Functions

One of the problems with the moment generating function is that it may not exist on any neighborhood of 0.

The characteristic function of a random variable X is the function $\phi : \mathbb{R} \to \mathbb{C}$ defined by

$$\phi_X(t) = E[e^{itX}] = \int e^{itx} dF(x).$$

Here, $i = \sqrt{-1}$.

Recall that $e^{itx} = \cos(tx) + i \sin(tx)$. We have

$$\phi_X(t) = E[\cos(tX)] + iE[\sin(tX)].$$

Since $0 \leq |\cos(tX)|, |\sin(tX)| \leq 1$, $\phi_X(t)$ is well-defined for all t. In fact, $|\phi_X(t)| \leq 1$ for all t.

When X is a discrete random variable, there is little problem in calculating its characteristic function: $\phi_X(t) = \sum_i e^{itx_i} P(X = x_i)$.

When X is a continuous random variable with probability density function f,

$$\phi_X(t) = \int_{\mathbb{R}} e^{itx} f(x) dx. \tag{2}$$

This is a complex integral. In general, one cannot calculate it by treating i as a real number, although such malpractice sometimes yields correct integration results. The correct techniques are from complex analysis or by integrating $\cos(tx)$ and $\sin(tx)$ separately.

(2) shows that $\phi_X(t)$ and $f(x)$ are Fourier dual of each other. They can be constructed from each other. There is a large body of theory and techniques for Fourier transform.

Note that, in form, $\phi_X(t) = M_X(it)$. This identity is incorrect in general, although it is correct for some distributions, e.g., exponential and Gaussian. Again, fundamentally, i cannot be treated as a real number in evaluating the integral.

In short, $\phi_X(t)$ has nicer properties than $M_X(t)$, although its calculation/analysis involves complex numbers.

Characteristic functions and moment generating functions share some similar properties.

Fact: If X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

Fact: If $a, b \in \mathbb{R}$ and $Y = aX + b$, then

$$\phi_Y(t) = e^{itb}\phi_X(at).$$

Theorem: (i) Suppose $\phi^{(k)}(0)$ exists. Then, $E|X^k| < \infty$ if k is even, and $E|X^{k-1}| < \infty$ if k is odd.

(ii) If $E|X^k| < \infty$, then

$$\phi(t) = \sum_{j=0}^{k} \frac{E[X^j]}{j!}(it)^j + o(t^k).$$

So, $\phi^{(k)}(0) = i^k E[X^k]$.

Note: The Taylor expansion is for t near 0. The term $o(t^k)$ approaches 0 faster than t^k when $t \to 0$.

Theorem (Inversion): If X is continuous with density function f and characteristic function ϕ, then

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt,$$

at every point x at which f is differentiable.

There is an analogous inversion theorem for a general random variable, which leads to

Corollary: Random variables X and Y have the same characteristic function if and only if they have the same distribution function.

Definition: We say that a sequence of distribution functions $F_1, F_2, \ldots$ converges to the distribution function F, written as $F_n \to F$, if $F(x) = \lim_{n\to\infty} F_n(x)$ at each point x where F is continuous.

Continuity Theorem: Suppose $F_1, F_2, \ldots$ is a sequence of distribution functions with corresponding characteristic functions $\phi_1, \phi_2, \ldots$.

(a) If $F_n \to F$ for some distribution function F with characteristic function ϕ, then $\phi_n(t) \to \phi(t)$ for all t.

(b) Conversely, if $\phi(t) = \lim_{n\to\infty} \phi_n(t)$ exists for all t and is continuous at $t = 0$, then ϕ is the characteristic function of some distribution function F and $F_n \to F$.

One can also define joint characteristic functions for multiple random variables.

Examples

Example: X is a standard Gaussian random variable with distribution $\mathcal{N}(0, 1)$. We know that $M_X(t) = e^{t^2/2}$. If we substitute t by it, we get

$$\phi_X(t) = e^{-t^2/2}.$$

The substitution requires a justification from complex analysis. In this case, the result is correct.

Example: Next, suppose Y has the distribution $\mathcal{N}(\mu, \sigma^2)$. Since $Y = \sigma X + \mu$, by the earlier fact, we have

$$\phi_Y(t) = \exp(i\mu t - \frac{1}{2}\sigma^2 t^2).$$

Example - Multivariate Gaussian Random Variables: Later.

Limit Theorems

Definition: If $X_1, X_2, \ldots$ is a collection of random variables with distributions $F_1, F_2, \ldots$. Also, let X be a random variable with distribution F. We say that X_n **converges in distribution** to X, written $X_n \xrightarrow{\text{d}} X$, if $F_n \to F$.

Theorem: Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite mean μ. Let S_n be the partial sum:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then, we have

$$\frac{1}{n} S_n \xrightarrow{\text{d}} \mu, \quad \text{as } n \to \infty.$$

Proof: Let $\phi_X(t)$ be the characteristic function of each X_i. Then, the characteristic function of S_n is $(\phi_X(t))^n$. By the earlier fact, the characteristic function of S_n/n, denoted by $\phi_n(t)$, is

$$\phi_n(t) = (\phi_X(t/n))^n.$$

Since $E[X_i] = \mu < \infty$, we have

$$\phi_X(t) = 1 + i\mu t + o(t).$$

Then,

$$\phi_n(t) = \left(1 + \frac{i\mu t}{n} + o\big(\frac{t}{n}\big)\right)^n \to e^{i\mu t}, \quad \text{as } n \to \infty.$$

(The above uses the fact that $(1 + \frac{1}{n} + o(\frac{1}{n}))^n \to e$ as $n \to \infty$. It can be shown by taking the log and applying L'Hopital's rule.)

The constant random variable $Y = \mu$ has the characteristic function $e^{i\mu t}$. By the continuity theorem, we see that the distribution function of $\frac{1}{n}S_n$ converges to that of the Y.

Note: This is not the weak or strong law of large numbers, although for IID random variables, all three types of convergence hold.

It is weaker than weak law (convergence in probability).

Central Limit Theorem

For large n, S_n is about as big as $n\mu$. This will be even more obvious with the Strong Law of Large Numbers, $\frac{1}{n}S_n \to \mu$ with probability 1. It says for almost all $\omega \in \Omega$, $\left|\frac{S_n(\omega)-n\mu}{n}\right| \to 0$ as $n \to \infty$, which implies $S_n = n\mu + o(n)$ with probability 1.

We are interested in the term $o(n)$. Note that the $o(n)$ term is $S_n - n\mu$, i.e., how much S_n deviates from its mean $n\mu$.

If each X_i has finite variance σ^2, the $o(n)$ term is on the order of $n^{1/2}$, but in the distribution sense. This is easy to see because

$$\mathrm{var}(S_n - n\mu) = \mathrm{var}(S_n) = \sum_{i=1}^{n} \mathrm{var}(X_i) = n\sigma^2.$$

Hence, the standard deviation is $\sqrt{n}\sigma$. The standard deviation is a distributional property (as apposed to a sample-path property), i.e., it depends on the distribution function.

On a related note, $\mathrm{var}(\frac{S_n - n\mu}{n}) = \frac{\sigma^2}{n}$. The standard deviation is equal to $\frac{\sigma}{\sqrt{n}}$. The standard deviation is a measure of the fluctuation from the mean. Since the fluctuation of $\frac{S_n - n\mu}{n}$ approaches 0 as $n \to \infty$, this makes the laws of large numbers believable.

Since the fluctuation of $S_n - n\mu$ is roughly $\sqrt{n}\sigma$, we can consider the scaling $(S_n - n\mu)/\sqrt{n}$. The random variable $(S_n - n\mu)/\sqrt{n}$ has mean 0 and variance σ^2. The central limit theorem tells us more: $(S_n - n\mu)/\sqrt{n}$ approaches a normal distribution as $n \to \infty$.

Central Limit Theorem: Let $X_1, X_2, \ldots$ be a sequence of IID random variables with finite means μ and finite non-zero variances σ^2, and let $S_n = X_1 + X_2 + \cdots + X_n$. Then,

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{\mathrm{d}} X \text{ as } n \to \infty, \text{ where } X \text{ is distributed as } \mathcal{N}(0,1).$$

Proof: Let $Y_i = (X_i - \mu)/\sigma$, and let ϕ_Y be the characteristic function of Y_i.

Since $E[Y_i] = 0$ and $E[Y_i^2] = \text{var}(Y_i) = 1$, we have

$$\phi_Y(t) = 1 - \frac{1}{2}t^2 + o(t^2).$$

Let

$$U_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

Let ψ_n the be characteristic function of U_n for each n. Then,

$$\begin{aligned} \psi_n(t) =& (\phi_Y(t/\sqrt{n}))^n \\ =& \left(1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\right)^n \to e^{-t^2/2} \quad \text{as } n \to \infty. \end{aligned}$$

This is the characteristic function of the $\mathcal{N}(0,1)$ distribution. By the continuity theorem, we see that $U_n \xrightarrow{\text{d}} X$.

Law of the Iterated Logarithm.

- The strong law of large number says that for almost all $\omega \in \Omega$, $\left|\frac{S_n(\omega)}{n}\right| \to \mu$, or equivalently, $\frac{|S_n(\omega)-n\mu|}{n} \to 0$, as $n \to \infty$.
- The CLT says the distribution of $\frac{S_n-n\mu}{\sqrt{n}\sigma}$ converges to $\mathcal{N}(0,1)$.
- However, it can be shown that for almost all $\omega \in \Omega$, $\limsup_{n\to\infty} \frac{|S_n(\omega)-n\mu|}{\sqrt{n}\sigma} = \infty$, which implies that $\frac{|S_n(\omega)-n\mu|}{\sqrt{n}\sigma}$ exceeds any $K > 0$ for infinity many n. The sample path of $|S_n(\omega) - n\mu|$ is not in the order of $\sqrt{n}\sigma$. There exists a subsequence of $\{|S_n(\omega) - n\mu|\}_n$ that grows faster than $\sqrt{n}\sigma$.
- **Law of the Iterated Logarithm:** For almost all $\omega \in \Omega$,

$$\limsup_{n\to\infty} \frac{|S_n(\omega) - n\mu|}{\sqrt{2n\log\log n}\sigma} = 1.$$

It implies the following: If $c < 1$, then for almost all $\omega \in \Omega$,

$$|S_n(\omega) - n\mu| \geq c\sigma\sqrt{2n\log\log n} \quad \text{for infinitely many } n.$$

If $c > 1$, then for almost all $\omega \in \Omega$,

$$|S_n(\omega) - n\mu| \geq c\sigma\sqrt{2n \log\log n} \quad \text{for only finite many } n,$$

or equivalently,

$$|S_n(\omega) - n\mu| < c\sigma\sqrt{2n \log\log n} \quad \text{for all sufficiently large } n.$$

- When $c < 1$, $c\sigma\sqrt{2n \log\log n}$ is like an asymptotic lower bound $|S_n(\omega) - n\mu|$ (except that it is a subsequential lower bound); when $c > 1$, $c\sigma\sqrt{2n \log\log n}$ is like an asymptotic upper bound $|S_n(\omega) - n\mu|$.
- $|S_n(\omega) - n\mu|$ grows no faster than $c\sigma\sqrt{2n \log\log n}$ for any $c > 1$. But, for any $c < 1$, there exists a subsequence of $\{|S_n(\omega) - n\mu|\}_n$ that grows no slower than $c\sigma\sqrt{2n \log\log n}$.
- The above in turn implies the convergence speed of the law of large numbers. $\left|\frac{S_n(\omega)}{n} - \mu\right|$ converges to 0 at a speed no faster than $c\sigma\sqrt{2\log\log n/n}$ for any $c < 1$; it converges to 0 at a speed no slower than $c\sigma\sqrt{2\log\log n/n}$ for any $c > 1$.