LARGE DEVIATION THEORY

Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of independent and identically distributed random variables each with common mean μ and $S_n = X_1 + X_2 + ... + X_n$. The weak law of large number and the central limit theorem were concerned with the behavior of $\frac{S_n}{n}$ for large *n*. The large deviation theory deals with the tail probability of the form $P\left(\frac{S_n}{n} \geq a\right)$.

The WLLN states

$$\frac{S_n}{n} \xrightarrow{P} \mu$$

i.e. $\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0$

Thus the WLLN asserts that $\frac{S_n}{n}$ converges to μ and does not deal with any deviation.

The central limit theorem says

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

i.e. we can find the probability $P\left(\mu - a/\sqrt{n} < \frac{S_n}{n} < \mu + a/\sqrt{n}\right)$

In this case, we can find the probability the deviation is of the order $1/\sqrt{n}$ which is not large. Thus the CLT is not applicable for finding the probability of the large deviation of the mean.

MGF and CGF

Recall that the moment generating function (MGF) $M_X(s)$ of a random variable *X* is defined by

$$M_X(s) = Ee^{sX} = \int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx$$

where s is a real variable. Unlike the characteristic function, the MGF may not exist for all random variables and all values of *s*. If $M_X(s)$ exists at $s = 0$, it may be conveniently used to generate the moments as:

$$EX^k = M_X^{(k)}(s)\Big|_{s=0} \quad ,k = 1, 2, ..$$

where $M_X^{(k)}(s)$ is the *k*-th order derivative of $M_X(s)$.

Cumulant generating function

The *cumulant generating function (CGF)* of a random variable *X* is defined as $K_X(s) = \log_e\left(M_X(s)\right) = \log_e\left(Ee^{sX}\right)$

where $M_X(s)$ is the moment generating function.

If $M_X(s)$ exists and is non-zero, then $K_X(s)$ also exists. We also note the followings:

(a) $K_X(0) = 0$ and $K_X{}'(0) = \mu_X$

As $M_X(0) = 1$, we get

$$K_X(0) = \ln M_X(0) = 0.$$

Similarly, if $M_X{}'(0)$ exists, then

$$K_X{}'(0) = \frac{M_X{}'(0)}{M_X(0)} = \mu_X$$

(b) $K_X(s)$ is a convex function of *s*.

We have

$$\begin{aligned} K_X{}''(s) &= \frac{2M_X(s)M_X{}''(s) - \left(M_X{}'(s)\right)^2}{M_X{}^2(s)} \\ &= \frac{M_X(s)M_X{}''(s) - \left(M_X{}'(s)\right)^2}{M_X{}^2(s)} \\ &= \frac{Ee^{sX}EX^2e^{sX} - \left(EXe^{sX}\right)^2}{M_X{}^2(s)} \end{aligned}$$

The Taylor series expansion of $K_X(s)$ about the origin gives

$$K_X(s) = \sum_{n=1}^{\infty} k_X(n)\frac{s^n}{n!}$$

where the *n*th coefficient k_n is called the *n*th *cumulant* of the random variable of *X*.

From the above expression, we get

$$k_X(n) = \frac{d^n}{ds^n} K_X(s)\Big|_{s=0}$$

$$= \frac{d^n}{ds^n} \log_e\left(M_X(s)\right)\Big|_{s=0}$$

We can derive the first four cumulants as

$$k_X(1) = \frac{dK_X(s)}{ds}\Big|_{s=0} = \mu_X$$

$$k_X(2) = \frac{d^2K_X(s)}{ds^2}\Big|_{s=0} = \sigma_X^2$$

$$k_X(3) = \frac{d^3K_X(s)}{ds^3}\Big|_{s=0} = EX^3 - \mu EX^2 + 2\mu^3 = E(X-\mu_X)^3$$

$$k_X(4) = \frac{d^4K_X(s)}{ds^4}\Big|_{s=0} = E(X-\mu_X)^4 - 3\sigma_X^4$$

and so on**.** Observe that the mean is the first cumulant and the variance is the second cumulant.

Cramer's Theorem

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence iid random variables with mean μ and the MGF $M_X(s)$ which is finite in a neighbourhood of s = 0. Then for any $a > \mu$,

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{S_n}{n} \geq a\right) = -l(a)$$

where

$$l(a) = \max_{s>0}\left(sa - \log M_X(s)\right) = s^*a - \log M_X(s^*)$$

and s^* corresponding point of maximum.

$l(a)$ is called a rate function in the sense that for large n,

$$P\left(\frac{S_n}{n} \geq a\right) \simeq e^{-nl(a)}$$

Proof:

We can easily find an upper bound for $P\left(\left\{\frac{S_n}{n} > a\right\}\right)$ by using the Chernoff bound. According to the Chernoff bound,
($P(X \geq a) = \min\limits_{s>0} e^{-as} M_X(s)$).

$$\therefore P\left(\left\{\frac{S_n}{n} > a\right\}\right) = P(S_n > na) \leq \min_{s>0} e^{-ans} M_{S_n}(s)$$

$$\begin{aligned}
M_{S_n}(s) &= Ee^{sS_n} \\
&= Ee^{s\sum_{i=1}^{n} X_i} \\
&= E\prod_{i=1}^{n} e^{sX_i} \\
&= \prod_{i=1}^{n} Ee^{sX_i} \\
&= (M_X(s))^n
\end{aligned}$$

$$\begin{aligned}
\therefore P\left(\frac{S_n}{n} > a\right) &\leq \min_{s>0} e^{-ans} (M_X(s))^n \\
&= \min_{s>0} e^{-ans + n\log_e M_X(s)} \\
&= \min_{s>0} e^{-ans + nK_X(s)}
\end{aligned}$$

where $K_X(s)$ is the cumulant generating function

$$P\left(\frac{S_n}{n} > a\right) \leq \min_{s>0} e^{-n(as - \log_e M_X(s))}$$

Minimization of $e^{-n(as-\log_e M_X(s))}$ is equivalent to maximization of $as - \log_e M_X(s)$. The point maximum of $as - \log_e M_X(s)$ is given by

$$\frac{d}{ds}\left(sa - \log M_X(s)\right)\Big]_{s=s^*} = 0.$$

Equivalently,

$$\frac{M'_X(s^*)}{M_X(s^*)} = a$$

Therefore,

$$\begin{aligned}
l(a) &= \max_{s>0}\left(sa - \log_e M_X(s)\right) \\
&= s^* a - \log_e M_X(s^*)
\end{aligned}.$$

$l(a)$ is known as the *Fenchel-Legendre transform* of $\log_e M_X(s)$ and s^* is the value of s corresponding to the maximum in the above expression.

Thus,

$$P\left(\frac{S_n}{n} > a\right) \le e^{-nl(a)}$$

$$\therefore \frac{1}{n}\log_e P\left(\frac{S_n}{n} > a\right) \le -l(a)$$

We have also to show that

$$\lim_{n\to\infty}\frac{1}{n}P\left(\frac{S_n}{n} > a\right) \ge -(l(a)+\varepsilon)$$

For this, consider a sequence of new random variables $Y_1, Y_2, ..., Y_n, ...$ obtained by mapping X_i to Y_i such that the common probability density function of Y_i s is given by

$$f_Y(y) = \frac{e^{s^*y}}{M_X(s^*)} f_X(y)$$

We can verify that $f_Y(y)$ is a valid probability density function. As $X_i s$ are iid random variables so also $Y_i s$.

The MGF of Y_i is given by

$$\begin{aligned}
M_Y(s) &= \int_{-\infty}^{\infty} e^{sy} f_Y(y)dy \\
&= \int_{-\infty}^{\infty} \frac{e^{sy}e^{s^*y}}{M_X(s^*)} f_X(y)dy \\
&= \frac{M_X(s+s^*)}{M_X(s^*)}
\end{aligned}$$

$$\therefore EY_i = M_Y^{/}(0) = \frac{M_X^{/}(s^*)}{M_X(s^*)} = a$$

Similarly,

$$VarY_i = M_Y^{//}(0) - \left(M_Y^{/}(0)\right)^2$$

$$= \frac{M_X^{//}(s^*)}{M_X(s^*)} - \left(\frac{M_X^{/}(s^*)}{M_X(s^*)}\right)^2$$

Define $\tilde{S}_n = \sum_{i=1}^{n} Y_i$. Clearly $E\tilde{S}_n = nEY_i = na$

Now

$$M_{\tilde{S}_n}(s) = Ee^{s\sum_{i=1}^{n} Y_i}$$

$$= \prod_{i=1}^{n} Ee^{sY_i}$$

$$= \left(Ee^{sY}\right)^n$$

which we get using the i.i.d. property.

$$\therefore M_{\tilde{S}_n}(s) = \left(M_Y(s)\right)^n$$

$$= \frac{\left(M_X(s+s^*)\right)^n}{(M_x(s^*))^n}$$

$$= \frac{M_{S_n}(s+s^*)}{(M_x(s^*))^n}$$

Noting the definition of $M_{\tilde{S}_n}(s)$, we get

$$\therefore \int_{-\infty}^{\infty} e^{su} f_{\tilde{S}_n}(u)du = \frac{\int_{-\infty}^{\infty} e^{(s+s^*)u} f_{S_n}(u)du}{(M_x(s^*))^n}$$

so that

$$f_{\tilde{S}_n}(u) = \frac{e^{s^*u} f_{S_n}(u)}{(M_x(s^*))^n}$$

Using the above relationship, the probability involving S_n can be studied in terms of the probabilities involving $\tilde{S}_n$

Suppose $\delta > 0$ is a small real number. We have

$$P\left(\frac{S_n}{n} > a\right) = P\left(S_n > na\right)$$
$$= \int_{na}^{\infty} f_{S_n}(u)du$$
$$= \int_{na}^{\infty} (M_X(s^*))^n e^{-s^*u} f_{\tilde{S}_n}(u)du$$
$$\geq e^{-s^*n(a+\delta)}(M_x(s^*))^n \int_{na}^{na+n\delta} f_{\tilde{S}_n}(u)du$$
$$= e^{-(s^*na - n\log_e(M_X(s^*)) - s^*n\delta} \int_{na}^{na+n\delta} f_{\tilde{S}_n}(u)du$$
$$= e^{-(s^*na - n\log_e(M_x(s^*))} e^{-s^*n\delta} P(na < \tilde{S}_n < na + n\delta)$$

We have

$$P(na \leq \tilde{S}_n \leq na + n\delta)$$
$$= P\left(0 \leq \frac{\tilde{S}_n - na}{\sqrt{n}\sigma_Y} \leq \frac{\sqrt{n}\delta}{\sigma_Y}\right)$$

Now applying the CLT,

$$\lim_{n\to\infty} P\left(0 \leq \frac{\tilde{S}_n - na}{\sqrt{n}\sigma_Y} < \infty\right)$$
$$= \frac{1}{2}$$

$$\therefore P\left(\frac{S_n}{n} > a\right) \geq e^{-(s^*na - n\log_e(M_x(s^*))} \frac{1}{2} e^{-n\delta s^*}$$

Note also that $\frac{1}{2} > \frac{1}{e}$. Therefore, as n becomes large and noting that we can take δ arbitrarily close to *a*, we get

$$\therefore P\left(\frac{S_n}{n} > a\right) \geq e^{-(s^*na - n\log_e(M_x(s^*))}\frac{1}{e}e^{-n\delta s^*}$$

$$= e^{-(s^*na - n\log_e(M_x(s^*))}e^{-n(\delta s^* + \frac{1}{n})}$$

$$= e^{-(s^*na - n\log_e(M_x(s^*))}e^{-n\varepsilon}$$

Taking logarithm of both sides and dividing by *n*, we get

$$\frac{1}{n}\log_e P\left(\frac{S_n}{n} > a\right) \geq -(s^*a - \log_e(M_x(s^*) + \varepsilon)$$

$$= -(l\ \ ^*(a) + \varepsilon)$$

Combining the lower bound and upper bounds,

$$\lim_{n\to\infty}\frac{1}{n}\log P\left(\frac{S_n}{n} \geq a\right) = -l^*(a)$$

Thus for or large n,

$$P\left(\frac{S_n}{n} \geq a\right) \cong e^{-nl^*(a)}$$

Example: Let X_i **s** are Bernoulli random variable.

$$X_i = 1 \quad \text{with probability '}p\text{'}$$

$$= 0 \quad \text{with probability '1-}p\text{'}$$

The moment generating function is given by

$$M_X(s) = Ee^{sX} = pe^s + (1-p)$$

$$\therefore \log M_X(s) = \log\left(1 - p + pe^s\right)$$

$$\therefore l(s) = as - \log\left(1 - p + pe^s\right)$$

$l(s)$ is maximum at $s = \log\dfrac{a(1-p)}{p(1-a)}$

Then, $l^*(a) = a\log\dfrac{a}{p} + (1-a)\log\left(\dfrac{1-a}{1-p}\right)$

In fact, $S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(p,n)$ so that we can find $P\left(\dfrac{S_n}{n} \geq a\right)$ from which we can check the tightness of the bounds. As a is a number between 0 and 1 so a has a probability interpretation.

Since $l^*(a) = a\log\dfrac{a}{p} + (1-a)\log\left(\dfrac{1-a}{1-p}\right)$, so we can call $l^*(a)$ is the relative entropy or Kullback Leibler distance between $\left(a, 1-a\right)$ and $(p, 1-p)$.