THE LIMIT THEOREMS OF THE THEORY OF PROBABILITY

Contents

1 The central problem of probability theory

The central problem of probability theory is the study the asymptotic behavior for $n \to \infty$ of sums

$$S_n = Y_1 + Y_2 + \ldots + Y_n \,, \tag{1}$$

where $Y_1, Y_2, \ldots, Y_n, \ldots$ is a sequence of random variables. The normalized sum

$$\bar{Y}_n = \frac{1}{n} S_n \tag{2}$$

is usually called the ***empirical mean*** of the sequence of random variables.

2 Some preliminaries

2.1 Equivalence of random variables

There are several different senses in which random variables can be considered to be equivalent.

Equality Two random variables X and Y are equal ($=$) if they are equal as functions on $(\Omega, \mathcal{F})$.

Almost sure equality Two random variables X and Y are equal almost surely if, and only if, the probability that they are different is zero, $\mathbb{P}\{X \neq Y\} = 0$. In this case we shall write

$$X \overset{\text{a.s.}}{=} Y\,. \tag{3}$$

Equality in distribution The random variables X and Y are equal in distribution if they have the same distribution functions, i.e.,

$$\rho^X = \rho^Y$$

In this case we shall write

$$X \overset{\mathcal{D}}{=} Y\,. \tag{4}$$

Equality up to a location-scale transformation The random variable Z is equal in distribution to Y up to a linear transformation, i.e., $Z \overset{\mathcal{D}}{=} a + bY$, with $a \in \mathbb{R}$ and $b \in (0, \infty)$; a is called the ***location parameter*** and b the ***scale parameter***. Then

$$\rho_Z(y) = \frac{1}{b}\rho_Y\left(\frac{y-a}{b}\right) \tag{5}$$

In Fourier space

$$\hat{\rho}_Z(k) = e^{ika}\hat{\rho}_Y(bk)\,. \tag{6}$$

Z and Y are said to be of the ***same type*** and this is denoted by

$$Z \overset{\mathcal{D}}{\approx} Y\,. \tag{7}$$

To be of the same type is an ***equivalence relation*** among random variables.

2.2 Some important inequalities

Jensen's inequality For any convex function f

$$f\left(\mathbb{E}\{Y\}\right) \leq \mathbb{E}\{f(Y)\}\,. \tag{8}$$

Markov's inequality For a non-negative random variable Y, that is with $Y(\omega) > 0$ for ***almost all*** ω, [1] the following inequality holds

$$\mathbb{P}\{Y \geq a\} \leq \frac{1}{a}\mathbb{E}(Y)\,, \quad a > 0\,. \tag{9}$$

(9) follows from $a\mathbb{1}_{\{Y \geq a\}} \leq Y$ and $\mathbb{E}\{\mathbb{1}_{\{Y \geq a\}}\} = \mathbb{P}\{Y \geq a\}$.

Chebyshev inequality By applying Markov's inequality with $a = \epsilon^2$, one obtains

$$\mathbb{P}\left\{|Y - \mathbb{E}[Y]|^2 \geq \epsilon^2\right\} \leq \frac{1}{\epsilon^2}\mathbb{E}\{|Y - \mathbb{E}[Y]|^2\} = \frac{1}{\epsilon^2}\mathrm{Var}\,(Y)\,,$$

whence Chebyshev inequality

$$\mathbb{P}\left\{|Y - \mathbb{E}[Y]| \geq \epsilon\right\} \leq \frac{1}{\epsilon^2}\mathrm{Var}\,(Y) \tag{10}$$

[1] That is $\mathbb{P}\{Y \geq 0\} = 1$

2.3 Convergence of random variables

In many situations, one deals with an infinite sequence $\mathsf{Y}_1, \ldots, \mathsf{Y}_n, \ldots$ of random variables and is interested in their asymptotic behavior, that is, in the existence of a limit Y of the sequence.

Almost sure convergence

$$\mathbb{P}\left\{\lim_{n\to\infty} |\mathsf{Y}_n - \mathsf{Y}| = 0\right\} = 1\,, \quad \forall \epsilon > 0 \tag{11}$$

denoted by

$$\mathsf{Y}_n \xrightarrow{\text{a.s.}} \mathsf{Y} \tag{12}$$

It is also called convergence with probability one.

Convergence in probability

$$\lim_{n\to\infty} \mathbb{P}\left\{|\mathsf{Y}_n - \mathsf{Y}| \geq \epsilon\right\} = 0\,, \quad \forall \epsilon > 0 \tag{13}$$

denoted by

$$\mathsf{Y}_n \xrightarrow{\mathbb{P}} \mathsf{Y} \tag{14}$$

It is the convergence in measure.

Mean square convergence

$\mathbb{E}(\mathsf{Y}_n) < \infty$, $n = 1, 2, \ldots$, $\mathbb{E}(\mathsf{Y}) < \infty$ and

$$\lim_{n\to\infty} \mathbb{E}(|\mathsf{Y}_n - \mathsf{Y}|^2) = 0 \tag{15}$$

This is the convergence in $L^2(\Omega, \mathbb{P})$, the Hillbert space of square-integrable random variables. It is denoted by

$$\mathsf{Y}_n \xrightarrow{L^2} \mathsf{Y} \tag{16}$$

Convergence in distribution or in law

$$\lim_{n\to\infty} \int f(y)\rho^{\mathsf{Y}_n}(y)dy = \int f(y)\rho^{\mathsf{Y}}(y)dy \tag{17}$$

for all continuos and bounded functions f. It is denoted by

$$\mathsf{Y}_n \xrightarrow{\mathscr{D}} \mathsf{Y} \quad \text{or} \quad \mathsf{Y}_n \xrightarrow{\mathscr{L}} \mathsf{Y}\,. \tag{18}$$

- A theorem named after the Lévy, establishes that the pointwise convergence of the characteristic functions

  $$\hat{\rho}^{\mathsf{Y}_n}(k) = \mathbb{E}\{e^{ik\mathsf{Y}_n}\} \to \hat{\rho}^{\mathsf{Y}}(k) = \mathbb{E}\{e^{ik\mathsf{Y}}\}$$

  guarantees the convergence in distribution.

Relations between the convergences

almost sure convergence
↘
convergence in probability → convergence in law
↗
mean square convergence

That mean square convergence implies convergence in probability follows immediately from Markov's inequality, eq. (9),

$$\mathbb{P}\left\{|\mathsf{Y}_n - \mathsf{Y}| \geq \epsilon\right\} = \mathbb{P}\left\{|\mathsf{Y}_n - \mathsf{Y}|^2 \geq \epsilon^2\right\} \leq \frac{1}{\epsilon^2}\mathbb{E}(|\mathsf{Y}_n - \mathsf{Y}|^2) \tag{19}$$

- Convergence in law implies convergence in probability only in the case in which the limit is a degenerate random variable (its distribution is concentrated on a point).

3 I.i.d. r.v.

$\mathsf{Y}_1, \mathsf{Y}_2, \ldots, \mathsf{Y}_n, \ldots$, i.i.d. r.v. with $\mathsf{Y} \overset{\mathcal{D}}{=} \mathsf{Y}_i$ where Y has distribution $\rho(y)$, c.f. $\hat{\rho}(k)$, mean m, and variance σ^2. Then

$$\rho_{\mathsf{S}_n} = \rho^{\star n}(y)\,, \quad \rho_{\bar{\mathsf{Y}}_n} = n\rho^{\star n}(ny) \tag{20}$$

where $\rho^{\star n}$ is the n-th power of the convolution product.

$$\mathbb{E}\{\mathsf{S}_n\} = nm\,, \qquad \mathrm{Var}\,(\mathsf{S}_n) = n\sigma^2\,, \qquad \hat{\rho}_{\mathsf{S}_n} = \hat{\rho}(k)^n \tag{21}$$

$$\mathbb{E}\{\bar{\mathsf{Y}}_n\} = m\,, \qquad \mathrm{Var}\,(\bar{\mathsf{Y}}_n) = \frac{\sigma^2}{n} \qquad \hat{\rho}_{\bar{\mathsf{Y}}_n} = \hat{\rho}(k/n)^n\,, \tag{22}$$

Binomial distribution Let $\mathsf{S}_n = \mathsf{B}_1 + \ldots + \mathsf{B}_n$, where each variable of the sum is Bernoulli(p) distributed. Then

$$\mathbb{P}\left\{\mathsf{S}_n = r\right\} = \binom{n}{r} p^r (1-p)^{n-r}\,, \quad \text{with} \quad \binom{n}{r} = \frac{n!}{r!(n-r)!}\,. \tag{23}$$

This is the ***binomial distribution*** $b(n, r, p)$.

$$\mathbb{E}\{\mathsf{S}_n\} = np\,, \qquad \mathrm{Var}\,(\mathsf{S}_n) = npq\,, \quad q \equiv (1-p)\,. \tag{24}$$

Sum of n normal distributed variables Let $\mathsf{S}_n = \mathsf{Y}_1 + \ldots + \mathsf{Y}_n$, where each variable of the sum is $N(m, \sigma)$ distributed. Then

$$\hat{\rho}_{\mathsf{S}_n}(k) = \hat{\rho}(k)^n = e^{imnk} e^{-\frac{1}{2}\sigma^2 nk^2} \quad \Rightarrow \quad \rho(y) = N(nm, \sqrt{n}\sigma)$$
$$\rho_{\bar{\mathsf{Y}}_n}(k) = \hat{\rho}(k/n)^n = e^{imk} e^{-\frac{1}{2}\sigma^2 k^2/n} \quad \Rightarrow \quad \rho(y) = N(m, \sigma/\sqrt{n})$$

4 The law of large numbers

Proposition 1 (weak LLN)

Let $\mathsf{Y}_1, \mathsf{Y}_2, \ldots, \mathsf{Y}_n, \ldots$ be a sequence of i.i.d. random variables with $\mathbb{E}[\mathsf{Y}] < \infty$,

$\mathrm{Var}(Y) < \infty$ and empirical mean $\bar{Y}_n$ given by eq. (2). Let $Y \overset{\mathcal{D}}{=} Y_i$. Then

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\overline{Y}_n - \mathbb{E}[Y]\right| \geq \epsilon\right\} = 0\,, \quad \forall \epsilon > 0\,. \tag{25}$$

Proof We have $\mathbb{E}\{\bar{Y}_n\} = \frac{1}{n} n \mathbb{E}\{Y\} = \mathbb{E}[Y]$ and $\mathbb{E}\{|\bar{Y}_n - \mathbb{E}[Y]|^2\} = \mathrm{Var}\left(\bar{Y}_n\right) = \frac{1}{n^2} n \mathrm{Var}(Y) = \frac{\mathrm{Var}(Y)}{n}$. Therefore, $\bar{Y}_n$ converges to $\mathbb{E}[Y]$ in mean square, and thus, by eq. (19), also in probability. □

Comments

(a) The first LLN appeared in the first book on probability, ***Ars Conjectandi***, written by Jacob Bernoulli between 1684 and 1689 and published posthumous in 1713. Bernoulli considered the sequence of normalized sums $\bar{Y}_n$ where the variables Y_i take the value one with probability p and the value zero with probability $1 - p$.

(b) The proof above relies only on $\mathrm{Var}(Y_1 + \ldots + Y_n) = n\mathrm{Var}(Y)$. Thus the assumption of independence can be replaced with the weaker condition that the variables are ***uncorrelated***.

Two r. v. X e Y are said ***uncorrelated*** if $\mathrm{Cov}(X, Y) = 0$. This is weaker than independence. Correlation is only a measure of linear dependence. It does not necessarily imply anything about other kinds of dependence. Consider for example $X \sim N(0, \sigma)$ and $Y = X^2$. The two variables are clearly dependent, however,

$$\mathrm{Cov}(X, Y) = \mathbb{E}\{XY\} - \mathbb{E}\{X\}\mathbb{E}\{Y\} = \mathbb{E}\{X^3\} - 0 = 0\,.$$

The random variables $Y_1, \ldots, Y_n$ are said uncorrelated if their are pairwise uncorrelated.

(c) The assumption of finite variance of the variables of the sequence is ***not necessary*** for the validity of the LLN. It is sufficient that the variables have finite mean. The proof is however more complex.

(d) Under the only assumption of finite mean, Kolmogorov proved the so called ***strong*** LLN:

$$\overline{Y}_n \xrightarrow{\text{a.s.}} \mathbb{E}[Y]\,, \tag{26}$$

which is a convergence stronger than the convergence in probability.

(e) The assumption of identical distribution of the variables is also not necessary. The convergence of eq. (26) still holds, provided that $\mathbb{E}[Y]$ is replaced by

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{Y_k\} \quad \text{(and } \sum_{n=1}^{\infty} \mathrm{Var}\left(Y_n\right)/n^2 < \infty\text{.)}$$

(f) While almost sure convergence is mathematically relevant, in the applications convergence in probability is sufficient.

(g) The meaning of a limit theorem, insofar as its applications to the real world are concerned, is that it guarantees that for sufficiently large n the quantity depending on n is close to the limit value. In the real world n can be large, but never infinite, so what eq. (25) actually means is that the set of exceptions to the "correct" behavior, $E = \left\{\omega \in \Omega : \left|\overline{Y}_n - \mathbb{E}[Y]\right| \geq \epsilon\right\}$ has small "size," in the sense provided by the probability measure $\mathbb{P}$. How small it is depends on ϵ and n and the LLN is silent about this. The LLN is an asymptotic statement and does not say how close is $\overline{Y}_n$ to $\mathbb{E}[Y]$. This information is provided by the central limit theorem.

5 More on the LLN

5.1 Ergodicity

Let $f : \mathbb{R} \to \mathbb{R}$ be a measurable function. If $Y_1, Y_2, \ldots, Y_n, \ldots$ are i.i.d. random variables, also $f(Y_1), f(Y_2), \ldots, f(Y_n), \ldots$ will be. Then by the LLN

$$\frac{1}{n}\sum_{i=1}^{n} f(Y_i) \xrightarrow{\mathbb{P}} \mathbb{E}\{f(Y)\} \tag{27}$$

In other words,

Proposition 2 (Ergodic theorem)

The empirical mean of f converges in probability to the theoretical or sample mean of f.

- The ergodic theorem is a very general form of LLN. It holds under (much) weaker conditions: It is require that
  - (i) There exist a map ϕ such that $Y_2 = \phi(Y_1), \ldots Y_{n+1} = \phi(Y_n), \ldots$ (very very general)
  - (ii) $\mathbb{P}$ is invariant under (the dynamics generated by) ϕ ($\Rightarrow$ the r.v. Y_i are i.d.). (very general)
  - (iii) ***Ergodicity***: all invariant functions are constants, i.e.,

$$f \circ \phi = f \quad \Rightarrow \quad f = \text{const. a.e.}$$

- The ergodic theorem is more difficult to prove. von Neumann (1932) proved

$$\frac{1}{n}\sum_{i=1}^{n} f(Y_i) \xrightarrow{L^2} \mathbb{E}\{f(Y)\} \tag{28}$$

  Birkhoff (1931) proved

$$\frac{1}{n}\sum_{i=1}^{n} f(Y_i) \xrightarrow{\text{a.s.}} \mathbb{E}\{f(Y)\} \tag{29}$$

- Continuos version

$$\frac{1}{T}\int_0^T f(Y_t)dt \xrightarrow{\text{a.s.}} \mathbb{E}\{f(Y_0)\} \tag{30}$$

- It is believed that ergodicity is generic, but it is difficult to prove for concrete systems.

Markov chains Consider a discrete time process $Y_1, Y_2, \ldots$ taking values in a finite set $\Sigma = \{y_1, y_2, \ldots, y_N\}$. Let $t = n$ be the present, so that Y_n is the state of the process in the present and Y_{n+1} its state in the immediate future; the past of the process is $Y_1, Y_2, \ldots Y_{n-1}$. ***Markov property***:

$$\mathbb{P}\{Y_{n+1}|Y_1, Y_2, \ldots Y_n\} = \mathbb{P}\{Y_{n+1}|Y_n\} . \tag{31}$$

Then

$$\begin{aligned}\mathbb{P}\{\mathsf{Y}_{n+1},\mathsf{Y}_n,\ldots,\mathsf{Y}_2,\mathsf{Y}_1\} &= \mathbb{P}\{\mathsf{Y}_{n+1}|\mathsf{Y}_1,\mathsf{Y}_2,\ldots\mathsf{Y}_n\}\,\mathbb{P}\{\mathsf{Y}_1,\mathsf{Y}_2,\ldots\mathsf{Y}_n\}\\ &= \mathbb{P}\{\mathsf{Y}_{n+1}|\mathsf{Y}_n\}\,\mathbb{P}\{\mathsf{Y}_n|\mathsf{Y}_1,\mathsf{Y}_2,\ldots\mathsf{Y}_{n-1}\}\,\mathbb{P}\{\mathsf{Y}_1,\mathsf{Y}_2,\ldots\mathsf{Y}_{n-1}\}\\ &= \mathbb{P}\{\mathsf{Y}_{n+1}|\mathsf{Y}_n\}\,\mathbb{P}\{\mathsf{Y}_n|\mathsf{Y}_{n-1}\}\,\mathbb{P}\{\mathsf{Y}_1,\mathsf{Y}_2,\ldots\mathsf{Y}_{n-1}\}\\ &= \ldots\\ &= \mathbb{P}\{\mathsf{Y}_{n+1}|\mathsf{Y}_n\}\,\mathbb{P}\{\mathsf{Y}_n|\mathsf{Y}_{n-1}\}\cdots\mathbb{P}\{\mathsf{Y}_2|\mathsf{Y}_1\}\,\mathbb{P}\{\mathsf{Y}_1\}\,.\end{aligned}$$

Thus , given the family of $N\times N$ matrices $P(n)=\left[p^{ij}(n)\right]$, $n=0,1,2,\ldots$, with components

$$p^{ij}(n)=\mathbb{P}\{\mathsf{Y}_{n+1}=y_j|\mathsf{Y}_n=y_i\}\quad\text{for}\quad i,j=1,2,\ldots,N\,, \tag{32}$$

and the initial distribution of Y_1, all the distributions of the process are determined. The process so defined is called a ***Markov chain*** and the matrix $P(n)$ is called its ***transition matrix***.

Obviously, $0\le p^{ij}(n)\le 1$ and, since Y_{n+1} can only attain states in Σ,

$$\sum_{j=1}^{N}p^{ij}(n)=1$$

for each $i=1,2,\ldots,N$ and $n=1,2,3,\ldots$. A probability vector on Σ is a vector $p=(p_1,p_2,\ldots p_N)\in\mathbb{R}^N$, with $0\le p_i\le 1$, for $i=1,2,\ldots,N$, and $\sum_i p_i=1$. Thus if $p(n)$ is the probability vector corresponding to the distribution of the random variable Y_n, that is $p^{ij}(n)=\mathbb{P}\{\mathsf{Y}_{n+1}=Y_j|\mathsf{Y}_n=Y_i\}$ for $i=1,2,\ldots,N$, then the probability vector $p(n+1)$ corresponding $\mathbb{P}\{\mathsf{Y}_{n+1}\}$ is related to it through the vector equation

$$p(n+1)=p(n)P(n)\,. \tag{33}$$

Hence, if we know the initial distribution probability vector $p(1)$, then we have

$$p(n)=p(1)P(1)P(2)\cdots P(n-1) \tag{34}$$

for $n=2,3,\ldots$ by applying eq. (33) recursively. In the case that the transition matrices are all the same, that is $P(n)=P$ for $n=1,2,3,\ldots$, the Markov chain is said ***homogeneous***. For a homogeneous Markov chain eq. (34) takes the form

$$p(n)=p(1)P^{n-1}\,. \tag{35}$$

A Markov chain is called a ***regular chain*** if some power of the transition matrix P has only positive elements.

Proposition 3 (Ergodic Markov chain)

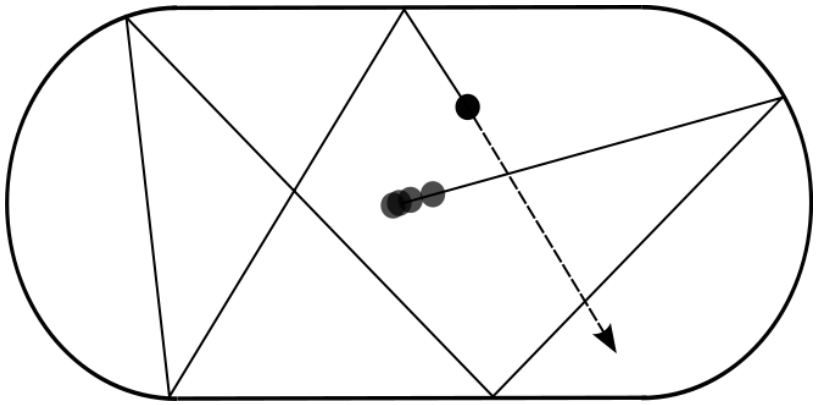
A homogeneous Markov chain is ergodic if it is possible to go from every state to every state (not necessarily in one move) and the chain is regular.

For an ergodic Markov chain, there is a unique probability vector w such that $wP=w$ and w is strictly positive. Any column vector f such that $Pf=f$ is a constant vector.

Hamiltonian systems Motion $\mathsf{X}_t = (Q_t, P_t) = T_t(\mathsf{X}_0)$ (T_t the Hamiltonian flow). Micro-canonical measure $\mathbb{P}_E$ on the shell Γ_E of constant energy.

Equivalent definition of ergodicity: if $T_t(A) = A$, then $\mathbb{P}_E\{A\} = 0$ or 1.

Physical meaning: Energy is the only constant of motion



The Bunimovich stadium. Proof of ergodicity 1979.

5.2 Empirical probability $\to$ theoretical probability

Let $f = \mathbb{1}_\Delta$. Then the l.h.s. of eq. (27) is

$$\frac{\text{numbers of terms of the sequence whose values fall in } \Delta}{n}$$
$$= \text{empirical probability that the first } n \text{ elements of the sequence are in } \Delta, \tag{36}$$

while the r.h.s. of eq. (27) is $\mathbb{E}\{\mathbb{1}_\Delta\} = \mathbb{P}\{\mathsf{Y} \in \Delta\}$. So,

$$\text{empirical probability} \longrightarrow \text{theoretical probability}\,. \tag{37}$$

This means that most sequences of data of the observed values $Y_1, Y_2, \dots$ of the random variables $\mathsf{Y}_1, \mathsf{Y}_2, \dots$ allow to reconstruct (with a small error) the statistical properties of the random variables, and the error (which, as we shall see next, is of order $1/\sqrt{n}$) becomes negligible when the number of data is sufficiently large.

N.B. Theoretical probability is an idealization, whose value lies in the approximation it provides for sums of a large but finite number of random variables.

Bernoulli shifts

The simplest non trivial dynamics possible is the following one. The phase-space is the interval $[0,1]$ and the time evolution of a point in it is one in which the time is discrete with the point evolving by iterating the map

$$\sigma : [0,1] \to [0,1]\,, \quad \sigma(x) = 2x \mod 1 \tag{38}$$

If we start with a value x_0 the map generates a sequence of iterates x_0, $x_1 = \sigma(x_0)$, $x_2 = \sigma(x_1) = \sigma(\sigma(x_0))$. In order to investigate the properties of this sequence we write x_0 in binary representation:

$$x_0 = \sum_{n=1}^{\infty} a_n 2^{-n} = 0, a_1 a_2 a_3 \cdots$$

where a_n has the values zero or unity. For $x_0 < 1/2$, we have $a_1 = 0$, and $x_0 > 1/2$ implies $a_1 = 1$. Therefore, the first iterate $\sigma(x_0)$ can be written as

$$\sigma(x_0) = \begin{cases} 2x_0 & \text{if } a_1 = 0 \\ 2x_0 - 1 & \text{if } a_1 = 1 \end{cases} = 0, a_2 a_3 a_4 \cdots .$$

i. e., the action of σ on the binary representation of x is to delete the first digit and shift the remaining sequence to the left. This is called the Bernoulli shift.

The Bernoulli property of $\sigma(x)$ demonstrates the sensitive dependence of the time evolution on the initial conditions. Even if two points x and x' differ only after their n-th digit a_n, this difference becomes amplified under the action of $\sigma(x)$, and their n-th iterates $\sigma^{(n)}(x)$ and $\sigma^{(n)}(x')$ already differ in the first digit because $\sigma^{(n)}(x) = 0, a_n \cdots$.

Moreover, the sequence of iterates $\sigma^{(n)}(x)$ has the same random properties as successive tosses of a coin. To see this, we attach to $\sigma^{(n)}(x)$ the symbol R or L depending on whether the iterate is contained in the right or left part of the unit interval. If we now prescribe an arbitrary sequence $RLLR\ldots$, e. g., by tossing a coin, we can always find an x_0 for which the series of iterates $x_0, \sigma(x_0), \sigma(\sigma(x_0))\ldots$ generates this sequence. This follows because $\sigma^{(n)}(x) = 0, a_n a_{n+1} \cdots$ corresponds to R or L if and only if $a_n = 1$ or $a_n = 0$; i.e.,the sequence $RLLR\ldots$ is isomorphic to the binary representation of x_0.

LLN (ergodic theorem) for

$$f_n^R(x_0) = \frac{1}{n} \sum_{k=1}^{n} a_k(x_0)$$

with respect to the Lebesgue measure on $[0, 1]$.

6 The Central limit theorem

Proposition 4 (CLT)

Let $\mathsf{Y}_1, \mathsf{Y}_2, \ldots, \mathsf{Y}_n, \ldots$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[\mathsf{Y}] = m < \infty$, variance $\mathrm{Var}(\mathsf{Y}) = \sigma^2 < \infty$. Let S_n be the sum of these random variables up to n. Then

$$\mathsf{Z}_n = \frac{\mathsf{S}_n - \mathbb{E}\{\mathsf{S}_n\}}{\sigma\sqrt{n}} \xrightarrow{\mathscr{D}} \mathsf{Z} \sim N(0,1) \quad \text{for} \quad n \to \infty\,, \tag{39}$$

i.e., Z_n converge in distribution to a normal distributed random variable with mean zero and unit variance.

Proof We have

$$\mathsf{Z}_n = \frac{\mathsf{S}_n - nm}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \left[\frac{\mathsf{Y}_1 - m}{\sigma} + \ldots + \frac{\mathsf{Y}_n - m}{\sigma} \right] .$$

$\frac{\mathsf{Y}_i - m}{\sigma} \overset{\mathscr{D}}{=} \mathsf{X}$ is a variable with zero mean and unit variance. Thus

$$\hat{\rho}_{\mathsf{Z}_n}(k) = \hat{\rho}_{\mathsf{X}}(k/\sqrt{n})^n$$

Then, (see previous lecture)

$$\hat{\rho}(k) = \left[1 - \frac{k^2}{2n} + o\left(\frac{k^2}{n}\right)\right]^n \to e^{-k^2/2}, \quad n \to \infty\,.$$

The limit is just the characteristic function of the standard normal distribution $N(0,1)$, whence eq. (39) (by appealing to the aforementioned Lévy theorem, according to which the convergence of characteristic functions implies convergence in distribution). □

Comments

- ***Universality***. The asymptotic distribution of the properly rescaled mean of the n independent random variables of vanishing expectation value is a Gaussian distribution, independent of the initial distribution (in some class); in particular, it depends only on one parameter σ. This independence provides a first example of universality. It shows a collective property of an infinite number of random uncorrelated variables.

- ***Renormalization***. Only a special affine function, with coefficients depending explicitly on n, of the sum S_n of the n initial random variables admits a non-trivial limiting (Gaussian) distribution. The affine transformation

$$\mathsf{Z}_n = \frac{\mathsf{S}_n - nm}{\sqrt{n}} = \frac{\mathsf{S}_n}{\sqrt{n}} - m\sqrt{n}$$

  is a first example of a renormalization, here a Gaussian renormalization.

- ***History***.The central limit theorem as stated above is known as Lindeberg—Lévy CLT. But already at the beginning of the XIX century Laplace used the method of the characteristic function to prove the theorem for sums of Bernoulli's distributed random variables. At the end of the XIX century the situation was more or less the following: the theorem was not proved for distributions with infinite support; there were no explicit conditions, in terms of the moments, under which the theorem would hold; the rate of convergence for the theorem was not studied. These problems were eventually solved by Russian mathematicians, between 1870 and 1910. Three probabilistic mathematicians are normally credited for this, namely Chebyshev, Markov and Liapounov—the so-called "St. Petersburg School."

- ***Lyapunov variant of the central limit theorem.*** In this variant of the theorem the random variables Y_i have to be independent, but not necessarily identically distributed. The theorem also requires that random variables$|\mathsf{Y}_i|$ have moments of some order $2+\delta$, and that the rate of growth of these moments is limited by the Lyapunov condition

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\big[\,|X_i - \mu_i|^{2+\delta}\,\big] = 0$$

  for some $\delta > 0$. Then

$$\frac{1}{s_n}\sum_{i=1}^{n}(\mathsf{Y}_i - \mathbb{E}\{\mathsf{Y}_i\}) \xrightarrow{\mathscr{D}} N(0,\,1)\,.$$

  Lyapunov proved this theorem in 1901. In 1920, Lindeberg showed that the Lyapunov condition could be replaced with a weaker one.

- ***Speed of convergence***. If we rewrite Z_n in eq. (39) as

$$\mathsf{Z}_n = \sqrt{n}\left(\bar{\mathsf{Y}}_n - \mathbb{E}[\mathsf{Y}]\right) \xrightarrow{\mathcal{D}} N(0,\sigma) \quad n \to \infty\,, \tag{40}$$

  we can easily read the CLT as establishing that the speed of convergence of $\bar{\mathsf{Y}}_n$ to $\mathbb{E}[\mathsf{Y}]$ is of order $1/\sqrt{n}$. In other words, the LLN tells that $\bar{\mathsf{Y}}_n$ converge to $\mathbb{E}[\mathsf{Y}]$, but it does not say how fast, the CLT then tells us that the error $\mathsf{E}_n = \bar{\mathsf{Y}}_n - \mathbb{E}[\mathsf{Y}]$ is on the order of $1/\sqrt{n}$.

- ***The CLT as approximation theorem***. The CLT is an approximation theorem par excellence. As often happens, the approximation is stated as a limit theorem. A limit theorem tells us that there is an approximation theorem, but it does not tell us how good the approximation is. The law of large numbers tells us that the average of sums of i.i.d. random variables tends to the expectations, but not how fast. The central limit theorem then tells us that the error E_n is on the order of $1/\sqrt{n}$. But what about the speed of convergence in the central limit theorem itself? The ***Berry-Esseen Theorem*** tells us that the distribution functions converge at a rate of about $1/\sqrt{n}$, and it gives the "about" in terms of the third moment of the underlying distribution.

- ***Why it is not possible to go beyond the convergence in distribution***. The CLT tells us that the error E_n converges to zero in distribution, but doesn't tell us whether the sequence $\mathsf{Z}_n = \sqrt{n}\mathsf{E}_n$ converges or not. In fact, it does not. The sequence not only fails to converge a.s., but its lim sup and lim inf are infinite. Neither there is convergence in probability.

- ***What, exactly, is the asymptotic behavior of the sequence*** $\mathsf{S}_n - \mathbb{E}\{\mathsf{S}_n\}$***?*** Evidently, as indicated above, the sequence is not $O(\sqrt{n})$. Hausdorff (1913) showed that it behaved as $O(\sqrt{n^{1+\epsilon}})$ and Hardy and Littlewood (1914) improved that to $O(\sqrt{n\log n})$. Then Khintchine (1922) improved that to $O(\sqrt{n\log(\log n)})$, and then finally, (1924) completed this by showing that this estimate couldn't be improved. Khinchine's theorem received the title ***Law of the Iterated Logarithm.*** So the rate of convergence of $\mathsf{E}_n = \bar{\mathsf{Y}}_n - \mathbb{E}[\mathsf{Y}]$ is not $O(1/\sqrt{n})$, but slightly—very slightly—slower. It is exactly $O(1/\sqrt{n\log(\log n)})$. To understand why this theorem is so celebrated, one must appreciate how tiny the $\log\log$ term is (e.g., $\log\log(10^{100})$ is less than five and a half).

7 Large deviations

This theory is not commonly studied in physics. Yet physicists have been using this theory for more than a hundred years. Basically, the theory started with Einstein's theory of fluctuations.

What are large deviations? A basic approximation or scaling law of the form $\rho_n \sim e^{-nI}$, where ρ_n is some probability distribution, n a parameter assumed to be large, and I some positive constant, is referred to as a ***large deviation principle***. While the central limit theorem estimates the probability of $O(1/\sqrt{n})$ deviations of $\bar{\mathsf{Y}}_n$ from its mean, it does not give any information about large deviations, i.e., those that are of the order of the mean of empirical mean itself.

7.1 Preliminaries

If a random variable Y has a moment-generating function $M(k) = \mathbb{E}\{e^{k\mathsf{Y}}\}$, then the domain of the characteristic function $\hat{\rho}(k) = \mathbb{E}\{e^{ik\mathsf{Y}}\}$ can be extended to the complex

plane, and

$$\hat{\rho}(k) = M(ik)\,. \tag{41}$$

Cumulant generating function The ***cumulants*** κ_n of a random variable Y are defined via the ***cumulant-generating function***

$$C(k) = \log \mathbb{E}\{e^{k\mathsf{Y}}\} = \log M(k)\,. \tag{42}$$

Its most important properties are

(i) $C(k)$ is convex.

(ii) $C(k) \geq k\mathbb{E}\{Y\} > -\infty$ (by Jensen's inequality).

(iii) $C(k)$ is real analytic in the interior of its essential domain and

$$C'(k) = \frac{1}{M}\frac{\partial M}{\partial k} \equiv m(k)\,, \tag{43a}$$

$$C''(k) = \frac{\partial m(k)}{\partial k} = -\frac{1}{M^2}\left(\frac{\partial M}{\partial k}\right)^2 + \frac{1}{M}\frac{\partial^2 M}{\partial k^2} \geq 0 \tag{43b}$$

The cumulants are defined by

$$\kappa_n = \left.\frac{\partial^n C}{\partial k^n}\right|_{k=0}\,. \tag{44}$$

Thus, from eq. (43a), we get the first two cumulants

$$\kappa_1 = C'(0) = \mathbb{E}\{\mathsf{Y}\} \quad \text{(since } M(0) = 1\text{).} \tag{45}$$

$$\kappa_2 = C''(0) = \mathbb{E}(\mathsf{Y}^2) - \mathbb{E}(\mathsf{Y})^2 = \sigma^2 \tag{46}$$

Second cumulant generating function Also the logarithm of a characteristic function is a cumulant generating function. It is defined by

$$\eta(k) = \log \hat{\rho}(k)\,. \tag{47}$$

and is called the second cumulant generating function.

Examples

Bernoulli(p)	$C(k) = \log(pe^k + q)$	(48)
Poisson(λ)	$C(k) = \lambda(e^k - 1)$	(49)
$N(m, \sigma)$	$C(k) = km + \frac{1}{2}\sigma^2 k^2$	(50)
Exponential(λ)	$C(k) = \log\frac{\lambda}{\lambda - k}$ for $k < \lambda$	(51)
Cauchy	$M(k) = \infty$ for $k \neq 0\,.$	(52)

N.B The normal distribution is the only absolutely continuous distribution all of whose cumulants beyond the first two (i.e., other than the mean and variance) are zero.

Rate function The rate function is defined as the Legendre transform of the cumulant generating function:

$$I(y) = \mathscr{L}(C)(y) = \sup_k [ky - C(k)] \,. \tag{53}$$

Since $C(k)$ is convex, $I(y)$ is convex too and the Legendre transform is invertible:

$$C(k) = \mathscr{L}(I)(k) = \sup_y [ky - I(y)] \,. \tag{54}$$

The most important properties of the rate function are

(i) $I(y)$ is convex, real analytic and non negative.

(ii) $I(y)$ attains its minimum value at $y = \mathbb{E}[Y] \equiv m$ and $I(m) = 0$. (By Jensen's inequality, we have $M(k) \geq e^{km}$ for all $k \in \mathbb{R}$, Taking the logarithm of both sides yields $km - C(k) \leq 0$ and it is equal to 0 for $k = 0$. We conclude that $I(m) = 0$. Consequently m is a minimum of the convex positive function $I(y)$.)

(iii) If $y > m$ and $k < 0$, then $kY - C(k) \leq km - C(k)$, whence

$$I(y) = \sup_{k \geq 0} [ky - C(k)] \,, \quad y \geq m \,. \tag{55}$$

Similarly,

$$I(y) = \sup_{k \leq 0} [ky - C(k)] \,, \quad y \leq m \,. \tag{56}$$

(iv) $I(y) \to +\infty$ as $|y| \to \infty$, and its level sets $\{y : I(y) \leq a\}$ are compact. (We have

$$\frac{I(y)}{y} \geq k - \frac{C(k)}{y} \quad \text{and} \quad \lim_{y \to +\infty} \frac{C(k)}{y} = 0 \,, \quad \text{whence} \quad \lim_{y \to +\infty} \frac{I(y)}{y} \geq k \,.$$

Consequently its level sets $\{y : I(y) \leq a\}$ are bounded, and closed by continuity of I.)

Examples

- Gaussian. The cumulant generating function C of a Gaussian $N(m, \sigma)$ is $km + \frac{1}{2}\sigma^2 k^2$. Let us compute $I(y)$. The maximum of $ky - C(k)$ is obtained at $k = k_y$ solution of $y = C'(k)$, i.e. $y = m + \sigma^2 k$, that is $k = (y - m)/\sigma^2$, whence

$$I(y) = \sup_k [ky - C(k)] = \frac{y - m}{\sigma^2} y - \frac{y - m}{\sigma^2} m - \frac{1}{2}\sigma^2 \frac{(y - m)^2}{\sigma^4} = \frac{(y - m)^2}{2\sigma^2} \,. \tag{57}$$

- Bernoulli(p). Calculation gives

$$I(y) = y \log \frac{y}{p} + (1 - y) \log \frac{1 - y}{1 - p}$$

for $0 \leq y \leq 1$. $I(y) = \infty$ outside of this range.

- Exponential(λ).

$$I(y) = \lambda y - 1 - \log(\lambda y) \quad \text{for} \quad y > 0 \quad \text{and } \infty \text{ otherwise}$$

- Cauchy.

$$I(y) = 0 \quad \text{for all} \quad y \,.$$

Cumulant generating function and the rate function for sums of i.i.d. r.v. From their very definition, the cumulant generating function and the rate function are the sums of the one-variable functions $C(k)$ and $I(Y)$,

$$C_{\mathsf{S}_n}(k) = nC(k) \tag{58}$$
$$I_{\mathsf{S}_n}(Y) = nI(Y)\,, \tag{59}$$

that is, in the language of thermodynamics, they are extensive.

7.2 The large deviation principle

Roughly, it says that the asymptotic probability distribution of $\bar{\mathsf{Y}}_n$ is

$$\rho_n(y) \asymp e^{-nI(y)} \tag{60}$$

where $I(y)$ is the rate function (76) of the one-variable distribution. The all subject of large deviations is about controlling the probabilities of atypical events analogous to those of the i.i.d. case. More specifically, the theory deals with rates at which probabilities of certain atypical events decay as a natural parameter in the problem varies.

N.B. ***The sign "$\asymp$" is used to express that, as $n \to \infty$, the dominant part of $\rho_n(Y)$ is the decaying exponential $e^{-nI(y)}$. We may also interpret the sign "$\asymp$" as expressing an equality relationship on a logarithmic scale; that is, we may interpret $a_n \asymp b_n$ as meaning that***

$$\lim_{n\to\infty} \frac{1}{n} \log a_n = \lim_{n\to\infty} \frac{1}{n} \log b_n\,.$$

Example: Gaussian empirical mean If the i.i.d. r.v. are $N(m,\sigma)$, then

$$\rho_{\bar{\mathsf{Y}}_n}(y) = \rho(y) = N(m, \sigma/\sqrt{n})(y) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} e^{-\frac{n(y-m)^2}{2\sigma^2}}\,. \tag{61}$$

Thus, asymptotically in n,

$$\rho_{\bar{\mathsf{Y}}_n}(y) \asymp \exp\left[-\frac{n(y-m)^2}{2\sigma^2}\right]\,. \tag{62}$$

in agreement with eq. (60).

7.3 The method of Laplace

For easiness of notation, we shall denote $\rho_{\bar{\mathsf{Y}}_n}$ by ρ_n

$$\hat{\rho}_n(k) = \hat{\rho}\,(k/n)^n$$

Therefore, by Fourier inversion formula and change of variables in the integral, we obtain

$$\rho_n(y) = \frac{n}{2\pi} \int_{-\infty}^{\infty} \hat{\rho}\,(k)^n\, e^{-inky} dk = \frac{n}{2\pi} \int_{-\infty}^{\infty} e^{n[\log \hat{\rho}(k) - iky]} dk\,. \tag{63}$$

We want to study the asymptotics of this integral for large n. To this end, we shall make a simplifying assumption: ***Assume that there exist positive numbers*** a ***and*** b ***such that integral***

$$\hat{\rho}(z) = \int_{-\infty}^{\infty} \rho(y) e^{izy} dy = \hat{\rho}(k+is) = \int_{-\infty}^{\infty} \rho(y) e^{-sy} e^{iky} dy \tag{64}$$

converges absolutely for $-a < s < b$, so that $\hat{\rho}(z)$ is analytic in the the strip $-ia < \mathrm{Im}(z) < ib$. Then $\hat{\rho}(z)$ is the Fourier transform of $\rho(y)e^{-sy}$ and by Fourier inversion formula we have

$$\rho(y) e^{-sy} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\rho}(k+is) e^{-iky} dk\,,$$

whence

$$\rho(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\rho}(k+is) e^{sy} e^{-iky} dk = \frac{1}{2\pi} \int_{-\infty+is}^{\infty+is} \hat{\rho}(z) e^{-izy} dz\,. \tag{65}$$

By change of variables $w = iz$, the integral above becomes

$$\rho(y) = \frac{1}{2\pi i} \int_{s-i\infty}^{s+i\infty} \hat{\rho}(-iw) e^{-wy} dw = \frac{1}{2\pi i} \int_{s-i\infty}^{s+i\infty} M(z) e^{-zy} dz\,, \tag{66}$$

where

$$M(z) = \hat{\rho}(-iz)\,. \tag{67}$$

Thus

$$\rho_n(y) = \frac{n}{2\pi i} \int_{s-i\infty}^{s+i\infty} M(z)^n e^{-nzy} dz \quad \text{with} \quad -a < s < b\,. \tag{68}$$

It is nowconvenient to introduce the function (analytic continuation of the cumulant generating function)

$$C(z) = \log M(z) \tag{69}$$

(analytic continuation of the cumulant generating function) and rewrite the integral as

$$\rho_n(y) = \frac{n}{2\pi i} \int_{s-i\infty}^{s+i\infty} e^{n[C(z)-zy]} dz \quad \text{with} \quad -a < s < b\,. \tag{70}$$

Laplace Integrals Laplace's original method is designed to approximate integrals of the form

$$I(\lambda) = \int A(x) e^{\lambda S(x)} dx \tag{71}$$

as $\lambda \to \infty$, where A and S are real-valued functions with some degree of smoothness. These integrals are similar to the one we are interested in (with $n = \lambda$ and $A = 1$), but not identical, since the integral (70) involve complex numbers.

Yet, the intuition for dealing with them is the same and it is useful to explain it first for integrals of the form (71).

Suppose that

- A and S are independent of λ.
- There is only one critical point x_0 that maximizes $S(x)$ for any $\lambda > 0$.
- $e^{\lambda S(x)}$ becomes very strongly peaked at $x = x_0$ and decreases rapidly away from x_0 as $\lambda \to \infty$; that is, $S''(x_0) < 0$.

Then approximate $S(x)$ by its leading Taylor term:

$$S(x) \approx S(x_0) + \frac{1}{2}S''(x_0)(x - x_0)^2$$

Then the main contribution to the integral comes from a small neighborhood of x_0:

$$\begin{aligned}\int A(x)e^{\lambda S(x)}dx &\sim \int_{x_0-\epsilon}^{x_0+\epsilon} A(x)e^{\lambda\left[S(x_0)+\frac{1}{2}S''(x_0)(x-x_0)^2\right]}dx \\ &\sim A(x_0)e^{\lambda S(x_0)}\int_{x_0-\epsilon}^{x_0+\epsilon} e^{\lambda\frac{S''(x_0)}{2}(x-x_0)^2}dx\end{aligned}$$

Now extend the domain of integration from $-\infty$ to $+\infty$. Though seemingly bizarre, this is legitimate because in this way are introduced corrections of order $O(1/\lambda)$, which can be neglected. Since

$$\int_{-\infty}^{\infty} e^{\lambda\frac{S''(x_0)}{2}(x-x_0)^2}dx = \frac{\sqrt{2\pi}}{\sqrt{-\lambda S''(x_0)}},$$

we finally arrive at

$$I(\lambda) \sim \sqrt{\frac{2\pi}{\lambda\left|S''(x_0)\right|}}A(x_0)e^{\lambda S(x_0)} \tag{72}$$

Derivation of Stirling's formula An elementary application of eq. (72) is the derivation of Stirling's formula: express the factorial as a Gamma function:

$$n! = \Gamma(n+1) = \int_0^\infty t^n e^{-t}dt = \int_0^\infty e^{-t+n\ln t}dt\,.$$

By substitution $t = nr$, rewrite the integral in the standard form of eq. (71)

$$n! = \int_0^\infty e^{-nr+n\ln n+n\ln r}ndr = n^{n+1}\int_0^\infty e^{n(-r+\ln r)}dr$$

So, $A(r) = 1$, $S(r) = -r + \ln r$, $S'(r) = -1 + \frac{1}{r}$, and $S''(r) = -\frac{1}{r^2}$, per $r > 0$. Thus S has a global maximum for $r = 1$ and $S''(1) = -1$. From eq. (72) we obtain the desired result

$$n! \sim n^{n+1}\frac{\sqrt{2\pi}e^{-n}}{\sqrt{n}} = \sqrt{2\pi n}\,n^n e^{-n}$$

Asymptotic approximation of the distribution of the empirical mean via the "saddle-point method" Let us return to the integral in eq. (70). Suppose now that the following condition holds:

$$\begin{aligned}&\text{the equation } C'(z) = y \text{ has a unique real solution } z = k_y \\ &\text{and } C''(k_y) > 0.\end{aligned} \tag{73}$$

Then choose the line $s = k_y$ to evaluate the integral in eq. (68). Since $C''(k_y) > 0$, $C(z) - zy$ has a minimum at $z = k_y$ along the real line. Then

- its modulus along the (imaginary) line $s = k_y$ must have a maximum (since the function is analytic);

- When we move away from the real line along $s = k_y$, the integrand become smaller than the maximum modulus, since for $z = k + is$ (k constant)

$$\left|e^{C(z)-zy}\right| = \left|M(z)e^{-zy}\right| = e^{-ky}\left|\int_{-\infty}^{\infty} e^{(k+is)y}\rho(y)dy\right| \leq e^{-ky}M(k)$$

  (the equality occurs only for $s = 0$).

- The term $e^{-n[k_y+is]}$ will oscillated wildly. Thus, by Riemann-Lebesgue lemma, only a small neighborhood of k_y need to be considered when n is large, the remaining contribution to the integral being negligible.

So, we are back to the situation encountered earlier with Laplace integrals. We obtain

$$\rho_n(y) = \sqrt{\frac{n}{2\pi C''(k_y)}}e^{n[C(k_y)-k_y y]}\left[1 + \frac{A_1}{n} + \frac{A_2}{n^2} + \ldots\right] \tag{74}$$

By a detailed computation, one finds the explicit form of the coefficients $A_1, A_2, \ldots$. Moreover, it can be shown that the (non convergent) series so obtained is an asymptotic expansion of ρ_n.

An asymptotic expansion is a series expansion of a function in a variable n which may converge or diverge but whose partial sums can be made an arbitrarily good approximation to a given function for n large enough .

7.4 The Large Deviation Approximation

The first term in the right had side of eq. (74),

$$\rho_n^{\text{LD}}(y) = \sqrt{\frac{n}{2\pi C''(k_y)}}e^{n[C(k_y)-k_y y]} \asymp e^{-nI(y)} \tag{75}$$

where

$$I(y) = \mathscr{L}(C)(y) = \sup_k \left[ky - C(k)\right], \tag{76}$$

is called the ***saddlepoint*** approximation of ρ_n or, in modern parlance, its ***large deviation*** approximation. Note that ρ_n^{LD} is always positive, but will not usually integrate to exactly one, so in practice it is renormalized. The renormalized version improves the relative order of the approximation to $O(n^{-3/2})$, which is $O(n^{-1})$ without renormalization, as shown in eq. (74).

The goodness of this approximation of eq. (74) hinges on assumption (73). It turns out that it is indeed quite general.

7.5 Cramer's theorem

Proposition 5 (Cramer's theorem (1938))

Let J be an interval $[a, b] \subset \mathbb{R}$. Assume that $M(k)$, the moment generating function of Y, is bounded in some neighborhood of $k = 0$ and that that the maximizer k_y of $ky - C(k)$ is in this neighborhood with $y \in [a, b]$. Then for any $\epsilon > 0$ and sufficiently large n

$$e^{-n(I(J)+\epsilon)} \leq \mathbb{P}\left\{\bar{Y}_n \in J\right\} \leq e^{-nI(J)}, \quad \text{where} \quad I(J) = \inf_{y\in J} I(y) \tag{77}$$

and $I(y)$ is the rate function

$$I(y) = \sup_k [ky - C(k)] = k_y y - C(k_y)\,. \tag{78}$$

Proof We first prove the upper bound and then, the lower bound.

Upper bound. If $a < m_{\mathsf{Y}} < b$, then by the LLN and the fact that $I(m_{\mathsf{Y}}) = 0$ (which implies $\inf_{y\in J} I(y) = 0$) we have that the upper bound statement is true. Now, without loss of generality, suppose that $a \geq m_{\mathsf{Y}}$ (the case $a < b < m_{\mathsf{Y}}$ is handled by just a change of sign).

Then for $k > 0$

$$\begin{aligned}
\mathbb{P}\left\{\bar{\mathsf{Y}}_n \in [a,b]\right\} &= \int_{\frac{1}{n}(y_1+\ldots+y_n)\in[a,b]} \rho(y_1)\cdots\rho(y_n)dy_1\cdots dy_n \\
&\leq e^{-ka}\int_{\frac{1}{n}(y_1+\ldots+y_n)\in[a,b]} e^{k(y_1+\ldots+y_n)/n}\rho(y_1)\cdots\rho(y_n)dy_1\cdots dy_n \\
&\leq e^{-ka}\int e^{k(y_1+\ldots+y_n)/n}\rho(y_1)\cdots\rho(y_n)dy_1\cdots dy_n \\
&\leq e^{-ka}M\left(\frac{k}{n}\right)^n .
\end{aligned}$$

This implies that

$$\mathbb{P}\left\{\bar{\mathsf{Y}}_n \in [a,b]\right\} \leq e^{-ka+nC(k/n)}\,,$$

with $C(k)$ the generating function of the the cumulants of Y. One may now replace k with nk to obtain

$$\mathbb{P}\left\{\bar{\mathsf{Y}}_n \in [a,b]\right\} \leq e^{-nka+nC(k)}\,.$$

Since this inequality is true for all $k \geq 0$, we have

$$\mathbb{P}\left\{\bar{\mathsf{Y}}_n \in [a,b]\right\} \leq e^{-n\sup_{k>0}[ka+C(k)]} = e^{-nI(a)} = e^{-n\inf_{y\in[a,b]} I(y)} = e^{-nI(J)} \tag{79}$$

(recall eq. (56)). The above result is called the *Chernoff bound*. Note that this bound actually holds for any positive integer n (it is not an asymptotic statement). The first part of the proof is completed.

Lower bound. Let $y \in (a,b)$. Let $\delta > 0$ be small enough so that $(y-\delta, y+\delta) \subset (a,b)$. It is the sufficient to show that

$$\mathbb{P}\left\{\bar{\mathsf{Y}}_n \in (y-\delta, y+\delta)\right\} \geq e^{-I(y)} \tag{80}$$

Let k_y be the point such that $I(y) = k_y y - C(k_y)$. Without any loss of generality we can suppose that y is a number greater or equal to the mean value m_{Y}. This the implies that k_y is greater or equal to 0 since $I(y)$ attains its minimum at m_{Y} and for $y > m_{\mathsf{Y}}$ is non decreasing so that its derivative at y, k_y, is is greater or equal to 0.

The key idea for proving eq. (80) is the definition of a new distribution under which the random variable has a mean at y, instead of at m_{Y}. This changes the nature of the deviation from the mean to a small deviation, as opposed to a large deviation and thus allows the use the classical limit theorems. The new distribution which is useful to introduce is

$$p(x) = \frac{e^{k_y x}\rho(x)}{M(k_y)}$$

Its moment generating function is

$$M_{\mathsf{X}}(k) = \int_{-\infty}^{+\infty} e^{kx} p(x) dx = \frac{M(k+k_y)}{M(k_y)}$$

Thus

$$m_{\mathsf{X}} = \left.\frac{\partial M_{\mathsf{X}}}{\partial k}\right|_{k=0} = \frac{M'(k_y)}{M(k_y)} = g'(k_y)\,.$$

But k_y has been defined has the point where $ky - g(k)$ is maximum, that is the solution of $y - g'(k) = 0$. Thus $g'(k_y) = y$, whence $m_{\mathsf{X}} = y$. Thus by the law of large numbers, we must have per any $\delta_1 > 0$

$$\lim_{n\to\infty} \int_{\left|\frac{1}{n}\sum_i^n x_i - y\right| < \delta_1} p(x_1)\cdots p(x_n) dx_1 \cdots dx_n = 1$$

Now note that for $\delta_1 < \delta$ we have

$$\begin{aligned}
\mathbb{P}\left\{\bar{\mathsf{Y}}_n \in (y-\delta, y+\delta)\right\} &= \int_{\left|\frac{1}{n}\sum_i x_i^n - y\right| < \delta} \rho(x_1)\cdots\rho(x_n) dx_1\cdots dx_n \\
&\geq \int_{\left|\frac{1}{n}\sum_i^n x_i - y\right| < \delta_1} \rho(x_1)\cdots\rho(x_n) dx_1\cdots dx_n \\
&\geq e^{(-ny-\delta_1 n)k_y} \int_{\left|\frac{1}{n}\sum_i^n x_i - y\right| < \delta_1} e^{k_y \sum_i^n x_i} \rho(x_1)\cdots\rho(x_n) dx_1\cdots dx_n \\
&\geq e^{(-ny-\delta_1 n)k_y} M(k_y)^n \int_{\left|\frac{1}{n}\sum_i^n x_i - y\right| < \delta_1} p(x_1)\cdots p(x_n) dx_1\cdots dx_n\,.
\end{aligned}$$

Therefore for n large enough

$$\mathbb{P}\left\{\bar{\mathsf{Y}}_n \in (y-\delta, y+\delta)\right\} \geq e^{(-ny-\delta_1 n)k_y} M(k_y)^n = e^{-n[k_y y - C(k_y) + \delta_1 k_y]} = e^{-n[I(y)+\delta_1 k_y]}\,.$$

Since δ_1 can be made arbitrarly small, we have finished the lower bound part of the proof. □

Remark It should be noted that the proof of the theorem can be easily modified to yield the following result: for any $\delta > 0$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left\{ y \leq \frac{1}{n}\sum_{i=1}^n \mathsf{Y}_i \leq y + \delta \right\} = -I(y)\,. \tag{81}$$

Noting that, for small δ, $\mathbb{P}\left\{y \leq \frac{1}{n}\sum_{i=1}^n \mathsf{Y}_i \leq y+\delta\right\}$ can be interpreted as $\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n \mathsf{Y}_i \approx y\right\}$, this result states that the probability that the sum of the random variables exceeds ny is approximately equal (up to logarithmic equivalence) to the probability that the sum is equal to y, which is indeed what eq. (60) says.

8 From large to small deviations

Expand the rate function around the theoretical mean $m = \mathbb{E}\{Y\}$

$$I(y) = I(m) + I'(m)(y-m) + \frac{1}{2} I''(m)(y-m)^2 + \ldots$$

To determine $I(m)$, $I'(m)$, and $I''(m)$ we appeal to some properties of $I(y)$.

- $I(y)$ ***is a convex function***: it is the Legendre transform of the convex function $C(k)$, i.e.,

$$I(y) = \sup_k [ky - C(k)] \, .$$

- $I(y)$ ***is non negative***: since $C(k)$ is convex, the Legendre transform is invertible,

$$C(k) = \sup_y [ky - I(y)] \, .$$

Since $C(k) = \log M(k) = \log \mathbb{E}\{e^{k\mathsf{Y}}\}$, we have that $C(0) = 0$. Thus, from the above displayed equation

$$0 = C(0) = \sup_k [-I(y)] = -\inf I(y).$$

Since the inf is 0, the function is non negative. □

- $I(y)$ ***attains its minimum value at*** $y = \mathbb{E}[\mathsf{Y}] \equiv m$ ***and*** $I(m) = 0$ ***and*** $I'(m) = 0$: assume differentiability (no linear parts). Then

$$I(y) = k(y)y - C(k(y)) \, .$$

where $k(y)$ is solution of

$$y = C'(k(y)) \tag{82}$$

But

$$C'(k) = \frac{M'(k)}{M(k)} \quad \Rightarrow \quad C'(0) = \frac{M'(0)}{M(0)} = m \, .$$

Thus $k(m) = 0$, whence

$$I(y) = 0 - C(0) = 0 - 0 = 0 \, .$$

Thus $I(m) = 0$ and and since $y = m$ is a minimum (because the inf of $I(y)$ is 0), we have $I'(m) = 0$. □

- $I''(m) = 1/\sigma^2$: consider

$$dI = kdy + ydk - dC, \quad ydk = dC \quad \Rightarrow \quad I'(y) = k(y) \, , \quad I''(y) = k'(y) \, .$$

differentiate (82) with respect to y:

$$1 = C''(k(y))k'(y) \quad \Rightarrow \quad k'(y) = \frac{1}{C''(k(y)} = I''(y)$$

$$\Rightarrow \quad I''(m) = \frac{1}{C''(0)} = \frac{1}{\sigma^2}$$

□

Conclusions

So the Taylor series of the rate function around the mean becomes

$$I(y) = \frac{(y-m)^2}{2\sigma^2} + \dots$$

Thus the small deviations of $\bar{\mathsf{Y}}_n$ ***around its mean (small =*** $O(1/\sqrt{n})$ ***) are Gaussian-distributed:***

$$\rho_{\bar{\mathsf{Y}}_n} \sim e^{-n(y-m)^2/(2\sigma^2)}$$

In this sense, large deviation theory contains the Central Limit Theorem.