CHAPTER

SUM OF INDEPENDENT RANDOM VARIABLES 7

CHAPTER CONTENTS

In this chapter, the behavior of the sum of independent random variables is first investigated. Then the limiting behavior of the mean of independent and identically distributed (i.i.d.) samples when the number of samples tends to infinity is discussed.

7.1 CONVOLUTION

Let x and y be independent discrete variables, and z be their sum:

$$z = x + y.$$

Since $x + y = z$ is satisfied when $y = z - x$, the probability of z can be computed by summing the probability of x and $z - x$ over all x. For example, let z be the sum of the outcomes of two 6-sided dice, x and y. When $z = 7$, these dice take

$$(x, y) = (1,6),(2,5),(3,4),(4,3),(5,2),(6,1),$$

and summing up the probabilities of occurring these combinations gives the probability of $z = 7$.

The probability mass function of z, denoted by $k(z)$, can be expressed as

$$k(z) = \sum_x g(x)h(z - x),$$

where $g(x)$ and $h(y)$ are the probability mass functions of x and y, respectively. This operation is called the *convolution* of x and y and denoted by $x * y$. When x and y are continuous, the probability density function of $z = x + y$, denoted by $k(z)$, is given similarly as

$$k(z) = \int g(x)h(z - x)\mathrm{d}x,$$

where $g(x)$ and $h(y)$ are the probability density functions of x and y, respectively.

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00018-2

7.2 REPRODUCTIVE PROPERTY

When the convolution of two probability distributions in the same family again yields a probability distribution in the same family, that family of probability distributions is said to be *reproductive*. For example, the normal distribution is reproductive, i.e., the convolution of normal distributions $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$ yields $N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

When x and y are independent, the moment-generating function of their sum, $x + y$, agrees with the product of their moment-generating functions:

$$M_{x+y}(t) = M_x(t)M_y(t).$$

Let x and y follow $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$, respectively. As shown in Eq. (4.1), the moment-generating function of normal distribution $N(\mu_x, \sigma_x^2)$ is given by

$$M_x(t) = \exp\left(\mu_x t + \frac{\sigma_x^2 t^2}{2}\right).$$

Thus, the moment-generating function of the sum, $M_{x+y}(t)$, is given by

$$\begin{aligned} M_{x+y}(t) &= M_x(t)M_y(t) \\ &= \exp\left(\mu_x t + \frac{\sigma_x^2 t^2}{2}\right)\exp\left(\mu_y t + \frac{\sigma_y^2 t^2}{2}\right) \\ &= \exp\left((\mu_x + \mu_y)t + \frac{(\sigma_x^2 + \sigma_y^2)t^2}{2}\right). \end{aligned}$$

Since this is the moment-generating function of $N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$, the reproductive property of normal distributions is proved.

Similarly, computation of the moment-generating function of $M_{x+y}(t)$ for independent random variables x and y proves the reproductive properties for the binomial, Poisson, negative binomial, gamma, and chi-squared distributions (see Table 7.1). The Cauchy distribution does not have the moment-generating function, but the computation of the characteristic function $\varphi_x(t) = M_{ix}(t)$ (see Section 2.4.3) shows that the convolution of $\mathrm{Ca}(a_x, b_x)$ and $\mathrm{Ca}(a_y, b_y)$ yields $\mathrm{Ca}(a_x + a_y, b_x + b_y)$.

On the other hand, the geometric distribution $\mathrm{Ge}(p)$ (which is equivalent to the binomial distribution $\mathrm{NB}(1, p)$) and the exponential distribution $\mathrm{Exp}(\lambda)$ (which is equivalent to the gamma distribution $\mathrm{Ga}(1, \lambda)$) do not have the reproductive properties for p and λ.

7.3 LAW OF LARGE NUMBERS

Let $x_1, \ldots, x_n$ be random variables and $f(x_1, \ldots, x_n)$ be their joint probability mass/density function. If $f(x_1, \ldots, x_n)$ can be represented by using a probability mass/density function $g(x)$ as

$$f(x_1, \ldots, x_n) = g(x_1) \times \cdots \times g(x_n),$$

Table 7.1 Convolution

Distribution	x	y	$x + y$
Normal	$N(\mu_x, \sigma_x^2)$	$N(\mu_y, \sigma_y^2)$	$N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$
Binomial	$\mathrm{Bi}(n_x, p)$	$\mathrm{Bi}(n_y, p)$	$\mathrm{Bi}(n_x + n_y, p)$
Poisson	$\mathrm{Po}(\lambda_x)$	$\mathrm{Po}(\lambda_y)$	$\mathrm{Po}(\lambda_x + \lambda_y)$
Negative binomial	$\mathrm{NB}(k_x, p)$	$\mathrm{NB}(k_y, p)$	$\mathrm{NB}(k_x + k_y, p)$
Gamma	$\mathrm{Ga}(\alpha_x, \lambda)$	$\mathrm{Ga}(\alpha_y, \lambda)$	$\mathrm{Ga}(\alpha_x + \alpha_y, \lambda)$
Chi-squared	$\chi^2(n_x)$	$\chi^2(n_y)$	$\chi^2(n_x + n_y)$
Cauchy	$\mathrm{Ca}(a_x, b_x)$	$\mathrm{Ca}(a_y, b_y)$	$\mathrm{Ca}(a_x + a_y, b_x + b_y)$

$x_1, \ldots, x_n$ are mutually independent and follow the same probability distribution. Such $x_1, \ldots, x_n$ are said to be i.i.d. with probability density/mass function $g(x)$ and denoted by

$$x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} g(x).$$

When $x_1, \ldots, x_n$ are i.i.d. random variables having expectation μ and variance σ^2, the sample mean (Fig. 7.1),

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

satisfies

$$E[\overline{x}] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \mu,$$
$$V[\overline{x}] = \frac{1}{n^2} \sum_{i=1}^{n} V[x_i] = \frac{\sigma^2}{n}.$$

This means that the average of n samples has the same expectation as the original single sample, while the variance is reduced by factor $1/n$. Thus, if the number of samples tends to infinity, the variance vanishes and thus the sample average $\overline{x}$ converges to the true expectation μ.

The *weak law of large numbers* asserts this fact more precisely. When the original distribution has expectation μ, the characteristic function $\varphi_{\overline{x}}(t)$ of the average of independent samples can be expressed by using the characteristic function $\varphi_x(t)$ of a single sample x as

$$\varphi_{\overline{x}}(t) = \left[\varphi_x\left(\frac{t}{n}\right)\right]^n = \left[1 + i\mu\frac{t}{n} + \cdots\right]^n.$$

The mean of samples $x_1, \ldots, x_n$ usually refers to the *arithmetic mean*, but other means such as the *geometric mean* and the *harmonic mean* are also often used:

$$\text{Arithmetic mean: } \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\text{Geometric mean: } \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}},$$

$$\text{Harmonic mean: } \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}.$$

For example, suppose that the weight increased by the factors 2%, 12%, and 4% in the last three years, respectively. Then the average increase rate is not given by the arithmetic mean $(0.02 + 0.12 + 0.04)/3 = 0.06$, but the geometric mean $(1.02 \times 1.12 \times 1.04)^{\frac{1}{3}} \approx 1.0591$. When climbing up a mountain at 2 kilometer per hour and going back at 6 kilometer per hour, the mean velocity is not given by the arithmetic mean $(2+6)/2 = 4$ but by the harmonic mean $2d/(\frac{d}{2} + \frac{d}{6}) = 3$ for distance d, according to the formula "velocity = distance/time." When $x_1, \ldots, x_n > 0$, the arithmetic, geometric, and harmonic means satisfy

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \geq \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}},$$

and the equality holds if and only if $x_1 = \cdots = x_n$. The *generalized mean* is defined for $p \neq 0$ as

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}.$$

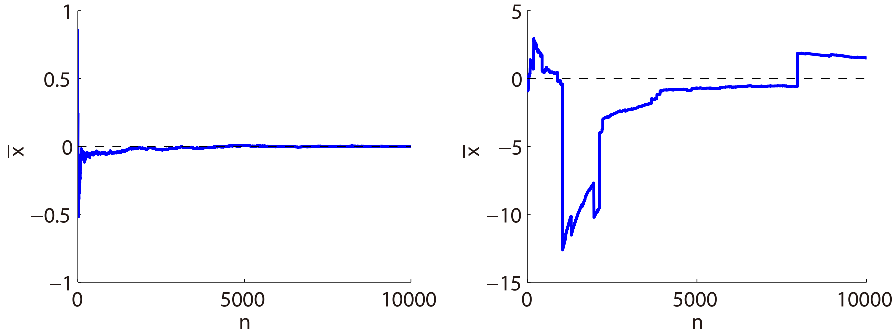
The generalized mean is reduced to the arithmetic mean when $p = 1$, the geometric mean when $p \to 0$, and the harmonic mean when $p = -1$. The maximum of $x_1, \ldots, x_n$ is given when $p \to +\infty$, and the minimum of $x_1, \ldots, x_n$ is given when $p \to -\infty$. When $p = 2$, it is called the *root mean square*.

FIGURE 7.1

Arithmetic mean, geometric mean, and harmonic mean.

Then Eq. (3.5) shows that the limit $n \to \infty$ of the above equation yields

$$\lim_{n\to\infty} \varphi_{\overline{x}}(t) = e^{it\mu}.$$



(a) Standard normal distribution $N(0, 1)$ (b) Standard Cauchy distribution Ca(0, 1)

FIGURE 7.2

Law of large numbers.

Since $e^{it\mu}$ is the characteristic function of a constant μ,

$$\lim_{n\to\infty} \Pr(|\overline{x} - \mu| < \varepsilon) = 1$$

holds for any $\varepsilon > 0$. This is the weak law of large numbers and $\overline{x}$ is said to *converge in probability* to μ. If the original distribution has the variance, its proof is straightforward by considering the limit $n \to \infty$ of Chebyshev's inequality (8.4) (see Section 8.2.2).

On the other hand, the *strong law of large numbers* asserts

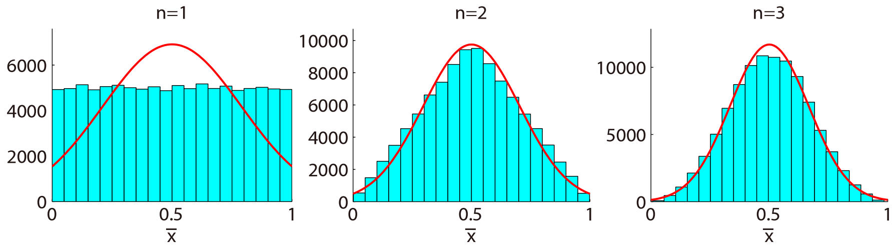
$$\Pr\left(\lim_{n\to\infty} \overline{x} = \mu\right) = 1,$$

and $\overline{x}$ is said to *almost surely converge* to μ. The almost sure convergence is a more direct and stronger concept than the convergence in probability.

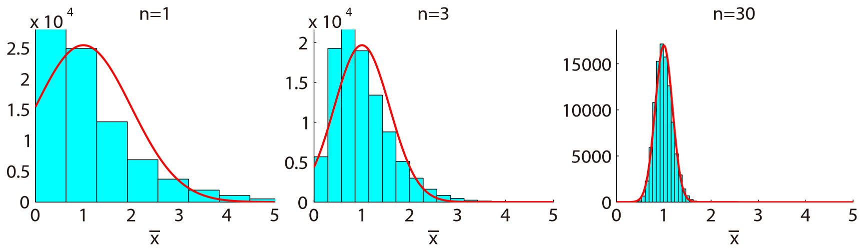
Fig. 7.2 exhibits the behavior of the sample average $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ when $x_1, \ldots, x_n$ are i.i.d. with the standard normal distribution $N(0, 1)$ or the standard Cauchy distribution Ca(0, 1). The graphs show that, for the normal distribution which possesses the expectation, the increase of n yields the convergence of the sample average $\overline{x}$ to the true expectation 0. On the other hand, for the Cauchy distribution which does not have the expectation, the sample average $\overline{x}$ does not converge even if n is increased.

7.4 CENTRAL LIMIT THEOREM

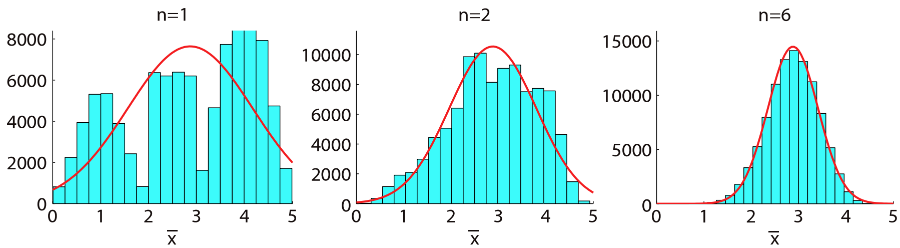
As explained in Section 7.2, the average of independent normal samples follows the normal distribution. If the samples follow other distributions, which distribution does



(a) Continuous uniform distribution $U(0, 1)$



(b) Exponential distribution Exp(1)



(c) Distribution used in Fig. 19.11

FIGURE 7.3

Central limit theorem. The solid lines denote the normal densities.

the sample average follow? Fig. 7.3 exhibits the histograms of the sample averages for the continuous uniform distribution $U(0, 1)$, the exponential distribution Exp(1), and the probability distribution used in Fig. 19.11, together with the normal densities with the same expectation and variance. This shows that the histogram of the sample average approaches the normal density as the number of samples, n, increases.

The *central limit theorem* asserts this fact more precisely: for standardized random variable

$$z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}},$$

the following property holds:

$$\lim_{n\to\infty} \Pr(a \leq z \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x.$$

Since the right-hand side is the probability density function of the standard normal distribution integrated from a to b, z is shown to follow the standard normal distribution in the limit $n \to \infty$. In this case, z is said to *converge in law* or *converge in distribution* to the standard normal distribution. More informally, z is said to *asymptotically* follow the normal distribution or z has *asymptotic normality*. Intuitively, the central limit theorem shows that, for any distribution, as long as it has the expectation μ and variance σ^2, the sample average $\overline{x}$ approximately follows the normal distribution with expectation μ and variance σ^2/n when n is large.

Let us prove the central limit theorem by showing that the moment-generating function of

$$z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

is given by the moment-generating function of the standard normal distribution, $e^{t^2/2}$. Let

$$y_i = \frac{x_i - \mu}{\sigma}$$

and express z as

$$z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i.$$

Since y_i has expectation 0 and variance 1, the moment-generating function of y_i is given by

$$M_{y_i}(t) = 1 + \frac{1}{2} t^2 + \cdots .$$

This implies that the moment-generating function of z is given by

$$M_z(t) = \left[M_{y_i/\sqrt{n}}(t)\right]^n = \left[M_{y_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 + \frac{t^2}{2n} + \cdots\right]^n.$$

If the limit $n \to \infty$ of the above equation is considered, Eq. (3.5) yields

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

which means that z follows the standard normal distribution.