CHAPTER

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

2

CHAPTER CONTENTS

In this chapter, the notions of random variables and probability distributions are introduced, which form the basis of probability and statistics. Then simple statistics that summarize probability distributions are discussed.

2.1 MATHEMATICAL PRELIMINARIES

When throwing a six-sided die, the possible outcomes are only 1, 2, 3, 4, 5, 6, and no others. Such possible outcomes are called *sample points* and the set of all sample points is called the *sample space*.

An *event* is defined as a subset of the sample space. For example, event A that any odd number appears is expressed as

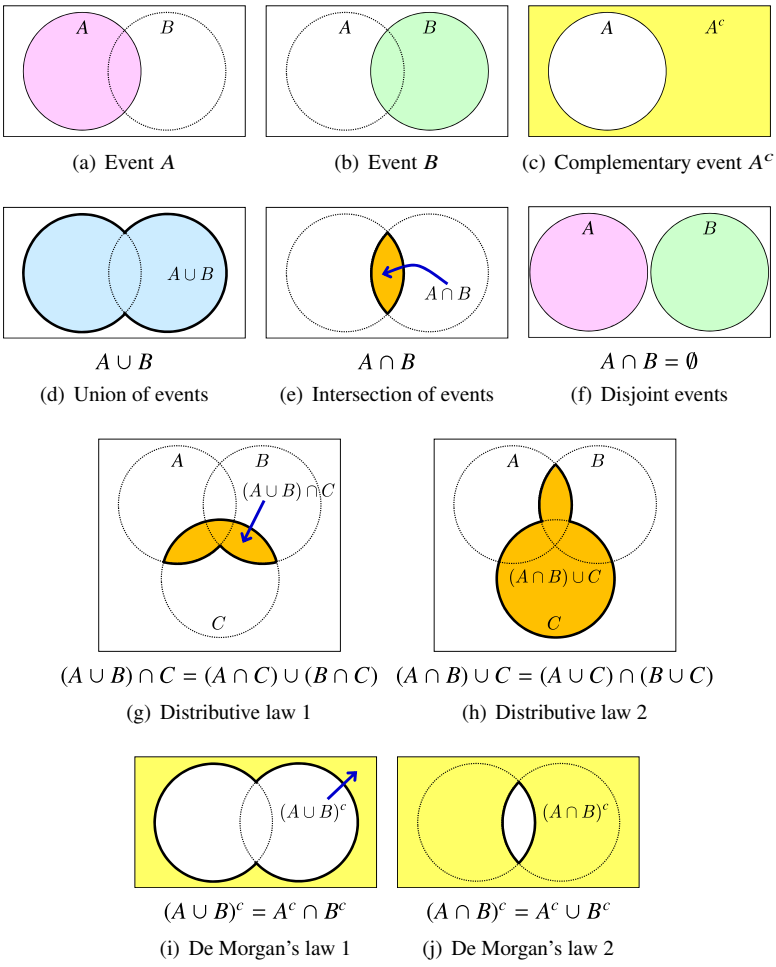
$$A = \{1,3,5\}.$$

The event with no sample point is called the *empty event* and denoted by $\emptyset$. An event consisting only of a single sample point is called an *elementary event*, while an event consisting of multiple sample points is called a *composite event*. An event that includes all possible sample points is called the *whole event*. Below, the notion of combining events is explained using Fig. 2.1.

The event that at least one of the events A and B occurs is called the *union of events* and denoted by $A \cup B$. For example, the union of event A that an odd number appears and event B that a number less than or equal to three appears is expressed as

$$A \cup B = \{1,3,5\} \cup \{1,2,3\} = \{1,2,3,5\}.$$

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00013-3



(a) Event A (b) Event B (c) Complementary event A^c

$A \cup B$

(d) Union of events

$A \cap B$

(e) Intersection of events

$A \cap B = \emptyset$

(f) Disjoint events

$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

(g) Distributive law 1

$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

(h) Distributive law 2

$(A \cup B)^c = A^c \cap B^c$

(i) De Morgan's law 1

$(A \cap B)^c = A^c \cup B^c$

(j) De Morgan's law 2

FIGURE 2.1

Combination of events.

On the other hand, the event that both events A and B occur simultaneously is called the *intersection of events* and denoted by $A \cap B$. The intersection of the above events A and B is given by

$$A \cap B = \{1,3,5\} \cap \{1,2,3\} = \{1,3\}.$$

If events A and B never occur at the same time, i.e.,

$$A \cap B = \emptyset,$$

events A and B are called *disjoint events*. The event that an odd number appears and the event that an even number appears cannot occur simultaneously and thus are disjoint. For events A, B, and C, the following *distributive laws* hold:

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C),$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C).$$

The event that event A does not occur is called the *complementary event* of A and denoted by A^c. The complementary event of the event that an odd number appears is that an odd number does not appear, i.e., an even number appears. For the union and intersection of events A and B, the following *De Morgan's laws* hold:

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$

2.2 PROBABILITY

Probability is a measure of likeliness that an event will occur and the probability that event A occurs is denoted by $\Pr(A)$. A Russian mathematician, *Kolmogorov*, defined the probability by the following three axioms as abstraction of the evident properties that the probability should satisfy.

1. **Non-negativity:** For any event A_i,

$$0 \leq \Pr(A_i) \leq 1.$$

2. **Unitarity:** For entire sample space Ω,

$$\Pr(\Omega) = 1.$$

3. **Additivity:** For any countable sequence of disjoint events $A_1, A_2, \ldots$,

$$\Pr(A_1 \cup A_2 \cup \cdots) = \Pr(A_1) + \Pr(A_2) + \cdots .$$

From the above axioms, events A and B are shown to satisfy the following *additive law*:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

This can be extended to more than two events: for events A, B, and C,

$$\begin{aligned}\Pr(A \cup B \cup C) = {} & \Pr(A) + \Pr(B) + \Pr(C) \\ & - \Pr(A \cap B) - \Pr(A \cap C) - \Pr(B \cap C) \\ & + \Pr(A \cap B \cap C).\end{aligned}$$

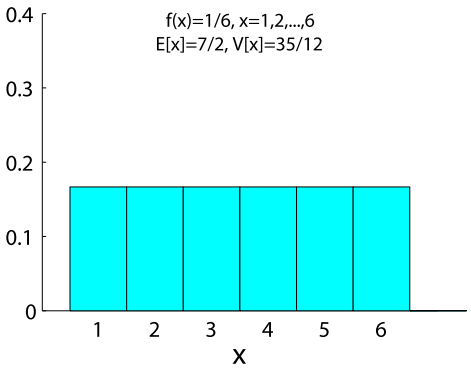


FIGURE 2.2

Examples of probability mass function. Outcome of throwing a fair six-sided dice (discrete uniform distribution $U\{1,2,\ldots,6\}$).

2.3 RANDOM VARIABLE AND PROBABILITY DISTRIBUTION

A variable is called a *random variable* if probability is assigned to each *realization* of the variable. A *probability distribution* is the function that describes the mapping from any realized value of the random variable to probability.

A *countable set* is a set whose elements can be enumerated as $1,2,3,\ldots$. A random variable that takes a value in a countable set is called a *discrete random variable*. Note that the size of a countable set does not have to be finite but can be infinite such as the set of all natural numbers. If probability for each value of discrete random variable x is given by

$$\Pr(x) = f(x),$$

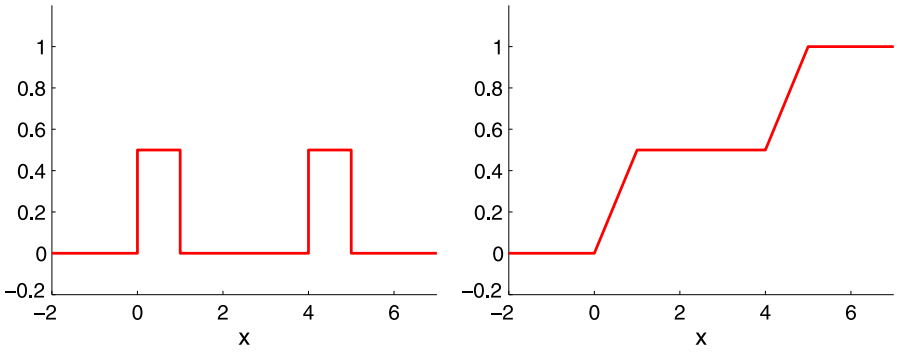
$f(x)$ is called the *probability mass function*. Note that $f(x)$ should satisfy

$$\forall x, \quad f(x) \geq 0, \quad \text{and} \quad \sum_x f(x) = 1.$$

The outcome of throwing a fair six-sided die, $x \in \{1,2,3,4,5,6\}$, is a discrete random variable, and its probability mass function is given by $f(x) = 1/6$ (Fig. 2.2).

A random variable that takes a continuous value is called a *continuous random variable*. If probability that continuous random variable x takes a value in $[a,b]$ is given by

$$\Pr(a \leq x \leq b) = \int_a^b f(x)\mathrm{d}x, \tag{2.1}$$



(a) Probability density function $f(x)$ (b) Cumulative distribution function $F(x)$

FIGURE 2.3

Example of probability density function and its cumulative distribution function.

$f(x)$ is called a *probability density function* (Fig. 2.3(a)). Note that $f(x)$ should satisfy

$$\forall x, \quad f(x) \geq 0, \quad \text{and} \quad \int f(x)\mathrm{d}x = 1.$$

For example, the outcome of spinning a roulette, $x \in [0, 2\pi)$, is a continuous random variable, and its probability density function is given by $f(x) = 1/(2\pi)$. Note that Eq. (2.1) also has an important implication, i.e., the probability that continuous random variable x exactly takes value b is actually zero:

$$\Pr(b \leq x \leq b) = \int_b^b f(x)\mathrm{d}x = 0.$$

Thus, the probability that the outcome of spinning a roulette is exactly a particular angle is zero.

The probability that continuous random variable x takes a value less than or equal to b,

$$F(b) = \Pr(x \leq b) = \int_{-\infty}^{b} f(x)\mathrm{d}x,$$

is called the *cumulative distribution function* (Fig. 2.3(b)). The cumulative distribution function F satisfies the following properties:

- **Monotone nondecreasing:** $x < x'$ implies $F(x) \leq F(x')$.
- **Left limit:** $\lim_{x \to -\infty} F(x) = 0$.
- **Right limit:** $\lim_{x \to +\infty} F(x) = 1$.

If the derivative of a cumulative distribution function exists, it agrees with the probability density function:

$$F'(x) = f(x).$$

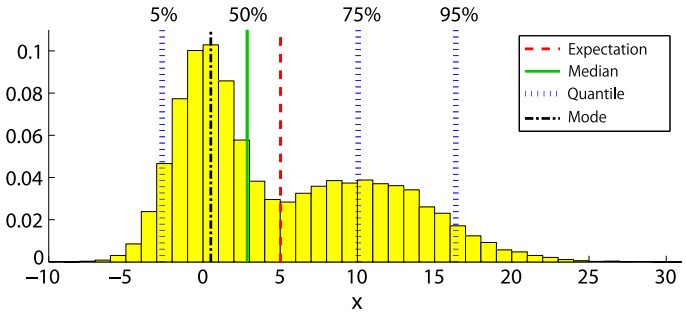


FIGURE 2.4

Expectation is the average of x weighted according to $f(x)$, and median is the 50% point both from the left-hand and right-hand sides. α-quantile for $0 \leq \alpha \leq 1$ is a generalization of the median that gives the $100\alpha\%$ point from the left-hand side. Mode is the maximizer of $f(x)$.

$\Pr(a \leq x)$ is called the *upper-tail probability* or the *right-tail probability*, while $\Pr(x \leq b)$ is called the *lower-tail probability* or the *left-tail probability*. The upper-tail and lower-tail probabilities together are called the *two-sided probability*, and either of them is called a *one-sided probability*.

2.4 PROPERTIES OF PROBABILITY DISTRIBUTIONS

When discussing properties of probability distributions, it is convenient to have simple statistics that summarize probability mass/density functions. In this section, such statistics are introduced.

2.4.1 EXPECTATION, MEDIAN, AND MODE

The *expectation* is the value that a random variable is expected to take (Fig. 2.4). The expectation of random variable x, denoted by $E[x]$, is defined as the average of x weighted according to probability mass/density function $f(x)$:

$$\text{Discrete: } E[x] = \sum_x x f(x),$$

$$\text{Continuous: } E[x] = \int x f(x) \mathrm{d}x.$$

Note that, as explained in Section 4.5, there are probability distributions such as the Cauchy distribution that the expectation does not exist (diverges to infinity).

The expectation can be defined for any function ξ of x similarly:

$$\text{Discrete: } E[\xi(x)] = \sum_x \xi(x) f(x),$$

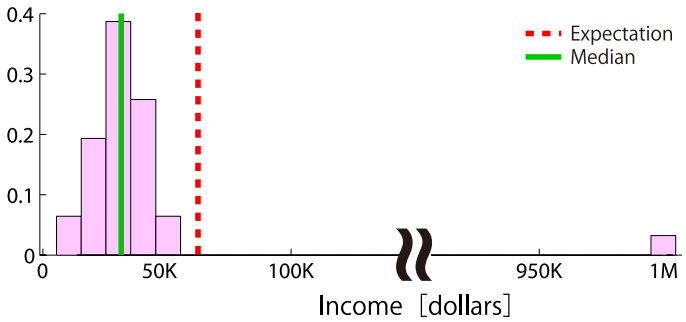


FIGURE 2.5

Income distribution. The expectation is 62.1 thousand dollars, while the median is 31.3 thousand dollars.

$$\text{Continuous: } E[\xi(x)] = \int \xi(x) f(x) \mathrm{d}x.$$

For constant c, the expectation operator E satisfies the following properties:

$$\begin{aligned} E[c] &= c, \\ E[x + c] &= E[x] + c, \\ E[cx] &= cE[x]. \end{aligned}$$

Although the expectation represents the "center" of a probability distribution, it can be quite different from what is intuitively expected in the presence of *outliers*. For example, in the income distribution illustrated in Fig. 2.5, because one person earns 1 million dollars, all other people are below the expectation, 62.1 thousand dollars. In such a situation, the *median* is more appropriate than the expectation, which is defined as b such that

$$\Pr(x \leq b) = 1/2.$$

That is, the median is the "center" of a probability distribution in the sense that it is the 50% point both from the left-hand and right-hand sides. In the example of Fig. 2.5, the median is 31.3 thousand dollars and it is indeed in the middle of everybody.

The α-*quantile* for $0 \leq \alpha \leq 1$ is a generalization of the median that gives b such that

$$\Pr(x \leq b) = \alpha.$$

That is, the α-quantile gives the $100\alpha\%$ point from the left-hand side (Fig. 2.4) and is reduced to the median when $\alpha = 0.5$.

Let us consider a probability density function f defined on a finite interval $[a, b]$. Then the minimizer of the *expected squared error*, defined by

$$E\left[(x-y)^2\right] = \int_a^b (x-y)^2 f(x)\mathrm{d}x,$$

with respect to y is shown to agree with the expectation of x. Similarly, the minimizer y of the *expected absolute error*, defined by

$$E\left[|x-y|\right] = \int_a^b |x-y| f(x)\mathrm{d}x, \tag{2.2}$$

with respect to y is shown to agree with the expectation of x. Furthermore, a weighted variant of Eq. (2.2),

$$\int_a^b |x-y|_\alpha f(x)\mathrm{d}x, \quad |x-y|_\alpha = \begin{cases} (1-\alpha)(x-y) & (x > y), \\ \alpha(y-x) & (x \leq y), \end{cases}$$

is minimized with respect to y by the α-quantile of x.

Another popular statistic is the *mode*, which is defined as the maximizer of $f(x)$ (Fig. 2.4).

2.4.2 VARIANCE AND STANDARD DEVIATION

Although the expectation is a useful statistic to characterize probability distributions, probability distributions can be different even when they share the same expectation. Here, another statistic called the *variance* is introduced to represent the spread of the probability distribution.

The variance of random variable x, denoted by $V[x]$, is defined as

$$V[x] = E\left[(x - E[x])^2\right].$$

In practice, expanding the above expression as

$$V[x] = E\left[x^2 - 2xE[x] + (E[x])^2\right] = E[x^2] - (E[x])^2$$

often makes the computation easier. For constant c, variance operator V satisfies the following properties:

$$\begin{aligned} V[c] &= 0, \\ V[x+c] &= V[x], \\ V[cx] &= c^2 V[x]. \end{aligned}$$

Note that these properties are quite different from those of the expectation.

The square root of the variance is called the *standard deviation* and is denoted by $D[x]$:

$$D[x] = \sqrt{V[x]}.$$

Conventionally, the variance and the standard deviation are denoted by σ^2 and σ, respectively.

2.4.3 SKEWNESS, KURTOSIS, AND MOMENTS

In addition to the expectation and variance, higher-order statistics such as the *skewness* and *kurtosis* are also often used. The skewness and kurtosis represent asymmetry and sharpness of probability distributions, respectively, and defined as

$$\text{Skewness: } \frac{E\left[(x - E[x])^3\right]}{(D[x])^3},$$
$$\text{Kurtosis: } \frac{E\left[(x - E[x])^4\right]}{(D[x])^4} - 3.$$

$(D[x])^3$ and $(D[x])^4$ in the denominators are for normalization purposes and -3 included in the definition of the kurtosis is to zero the kurtosis of the normal distribution (see Section 4.2). As illustrated in Fig. 2.6, the right tail is longer than the left tail if the skewness is positive, while the left tail is longer than the right tail if the skewness is negative. The distribution is perfectly symmetric if the skewness is zero. As illustrated in Fig. 2.7, the probability distribution is sharper than the normal distribution if the kurtosis is positive, while the probability distribution is duller than the normal distribution if the kurtosis is positive.

The above discussions imply that the statistic,

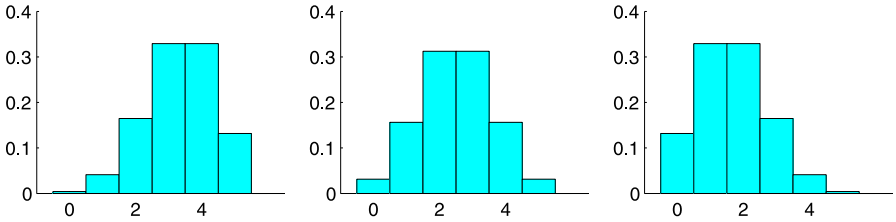
$$\nu_k = E\left[(x - E[x])^k\right],$$

plays an important role in characterizing probability distributions. ν_k is called the kth *moment* about the expectation, while

$$\mu_k = E[x^k]$$

is called the kth *moment* about the origin. The expectation, variance, skewness, and kurtosis can be expressed by using μ_k as

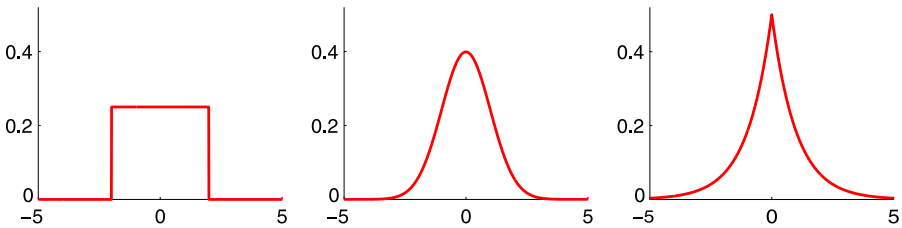
$$\text{Expectation: } \mu_1,$$
$$\text{Variance: } \mu_2 - \mu_1^2,$$
$$\text{Skewness: } \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{(\mu_2 - \mu_1^2)^{\frac{3}{2}}},$$
$$\text{Kurtosis: } \frac{\mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4}{(\mu_2 - \mu_1^2)^2} - 3.$$



(a) Skewness: −0.32 (b) Skewness: 0 (c) Skewness: 0.32

FIGURE 2.6

Skewness.



(a) Kurtosis: −1.2 (b) Kurtosis: 0 (c) Kurtosis: 3

FIGURE 2.7

Kurtosis.

Probability distributions will be more constrained if the expectation, variance, skewness, and kurtosis are specified. As the limit, if the moments of all orders are specified, the probability distribution is uniquely determined. The *moment-generating function* allows us to handle the moments of all orders in a systematic way:

$$M_x(t) = E[e^{tx}] = \begin{cases} \sum_x e^{tx} f(x) & \text{(Discrete)}, \\ \int e^{tx} f(x) \mathrm{d}x & \text{(Continuous)}. \end{cases}$$

Indeed, substituting zero to the kth derivative of the moment-generating function with respect to t, $M_x^{(k)}(t)$, gives the kth moment:

$$M_x^{(k)}(0) = \mu_k .$$

Below, this fact is proved.

The value of function g at point t can be expressed as

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

If higher-order terms in the right-hand side are ignored and the infinite sum is approximated by a finite sum, an approximation to $g(t)$ can be obtained. When only the first-order term $g(0)$ is used, $g(t)$ is simply approximated by $g(0)$, which is too rough. However, when the second-order term $tg'(0)$ is included, the approximation gets better, as illustrated below. By further including higher-order terms, the approximation gets more accurate and converges to $g(t)$ if all terms are included.

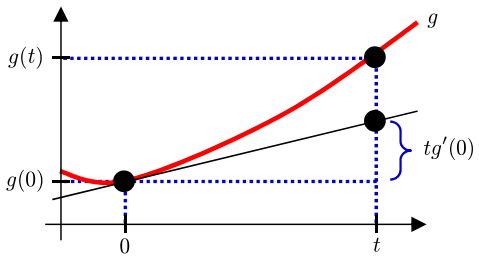


FIGURE 2.8

Taylor series expansion at the origin.

Given that the kth derivative of function e^{tx} with respect to t is $x^k e^{tx}$, the *Taylor series expansion* (Fig. 2.8) of function e^{tx} at the origin with respect to t yields

$$e^{tx} = 1 + (tx) + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \cdots .$$

Taking the expectation of both sides gives

$$E[e^{tx}] = M_x(t) = 1 + t\mu_1 + t^2\frac{\mu_2}{2!} + t^3\frac{\mu_3}{3!} + \cdots .$$

Taking the derivative of both sides yields

$$M_x'(t) = \mu_1 + \mu_2 t + \frac{\mu_3}{2!}t^2 + \frac{\mu_4}{3!}t^3 + \cdots ,$$
$$M_x''(t) = \mu_2 + \mu_3 t + \frac{\mu_4}{2!}t^2 + \frac{\mu_5}{3!}t^3 + \cdots ,$$
$$\vdots$$
$$M_x^{(k)}(t) = \mu_k + \mu_{k+1}t + \frac{\mu_{k+2}}{2!}t^2 + \frac{\mu_{k+3}}{3!}t^3 + \cdots .$$

Substituting zero into this gives $M_x^{(k)}(0) = \mu_k$.

Depending on probability distributions, the moment-generating function does not exist (diverges to infinity). On the other hand, its sibling called the *characteristic function* always exists:

$$\varphi_x(t) = M_{ix}(t) = M_x(it),$$

where i denotes the *imaginary unit* such that $i^2 = -1$. The characteristic function corresponds to the *Fourier transform* of a probability density function.

2.5 TRANSFORMATION OF RANDOM VARIABLES

If random variable x is transformed as

$$r = ax + b,$$

the expectation and variance of r are given by

$$E[r] = aE[x] + b \quad \text{and} \quad V[r] = a^2V[x].$$

Setting $a = \dfrac{1}{D[x]}$ and $b = -\dfrac{E[x]}{D[x]}$ yields

$$z = \frac{x}{D[x]} - \frac{E[x]}{D[x]} = \frac{x - E[x]}{D[x]},$$

which has expectation 0 and variance 1. This transformation from x to z is called *standardization*.

Suppose that random variable x that has probability density $f(x)$ defined on $\mathcal{X}$ is obtained by using transformation ξ as

$$x = \xi(r).$$

Then the probability density function of z is not simply given by $f\big(\xi(r)\big)$, because $f\big(\xi(r)\big)$ is not integrated to 1 in general. For example, when x is the height of a person in centimeter and r is its transformation in meter, $f\big(\xi(r)\big)$ should be divided by 100 to be integrated to 1.

More generally, as explained in Fig. 2.9, if the *Jacobian* $\frac{\mathrm{d}x}{\mathrm{d}r}$ is not zero, the scale should be adjusted by multiplying the absolute Jacobian as

$$g(r) = f\big(\xi(r)\big)\left|\frac{\mathrm{d}x}{\mathrm{d}r}\right|.$$

$g(r)$ is integrated to 1 for any transform $x = \xi(r)$ such that $\frac{\mathrm{d}x}{\mathrm{d}r} \neq 0$.

Integration of function $f(x)$ over $\mathcal{X}$ can be expressed by using function $g(r)$ on $\mathcal{R}$ such that

$$x = g(r) \quad \text{and} \quad \mathcal{X} = g(\mathcal{R})$$

as

$$\int_{\mathcal{X}} f(x)\mathrm{d}x = \int_{\mathcal{R}} f\big(g(r)\big)\left|\frac{\mathrm{d}x}{\mathrm{d}r}\right|\mathrm{d}r.$$

This allows us to change variables of integration from x to r. $\left|\frac{\mathrm{d}x}{\mathrm{d}r}\right|$ in the right-hand side corresponds to the ratio of lengths when variables of integration are changed from x to r. For example, for

$$f(x) = x \quad \text{and} \quad \mathcal{X} = [2,3],$$

integration of function $f(x)$ over $\mathcal{X}$ is computed as

$$\int_{\mathcal{X}} f(x)\mathrm{d}x = \int_2^3 x\mathrm{d}x = \left[\frac{1}{2}x^2\right]_2^3 = \frac{5}{2}.$$

On the other hand, $g(r) = r^2$ yields

$$\mathcal{R} = [\sqrt{2}, \sqrt{3}], \quad f\big(g(r)\big) = r^2, \quad \text{and} \quad \frac{\mathrm{d}x}{\mathrm{d}r} = 2r.$$

This results in

$$\int_{\mathcal{R}} f\big(g(r)\big)\left|\frac{\mathrm{d}x}{\mathrm{d}r}\right|\mathrm{d}r = \int_{\sqrt{2}}^{\sqrt{3}} r^2 \cdot 2r\mathrm{d}r = \left[\frac{1}{2}r^4\right]_{\sqrt{2}}^{\sqrt{3}} = \frac{5}{2}.$$

FIGURE 2.9

One-dimensional change of variables in integration. For multidimensional cases, see Fig. 4.2.

For linear transformation

$$r = ax + b \quad \text{and} \quad a \neq 0,$$

$x = \dfrac{r-b}{a}$ yields $\dfrac{\mathrm{d}x}{\mathrm{d}r} = \dfrac{1}{a}$, and thus

$$g(r) = \frac{1}{|a|} f\left(\frac{r-b}{a}\right)$$

is obtained.