01: Probability review

EE21: Random Signal Processing

March 2017

Zhigang SU
Professor
Sino-European Institute of Aviation Engineering
Civil Aviation University of China

Outline

Basic Probability Model

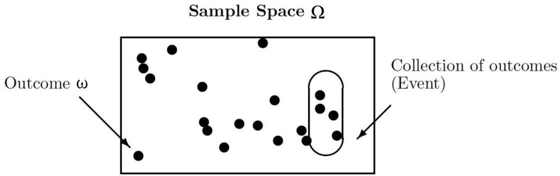
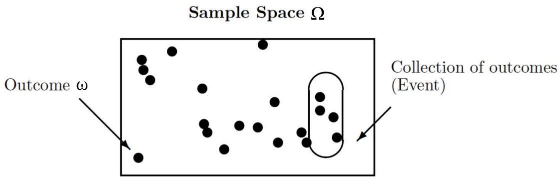


Figure: Sample space and events.

Associated with a basic probability model are the following three components

1. **Sample Space**
2. **Event Algebra**
3. **Probability Measure**

Basic Concepts

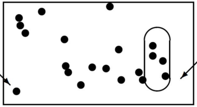


Sample Space

The sample space Ω is the set of all possible outcomes (samples) ω of the probabilistic experiment that the model represents.

We require that one and only one outcome be produced in each experiment with the model.

Basic Concepts

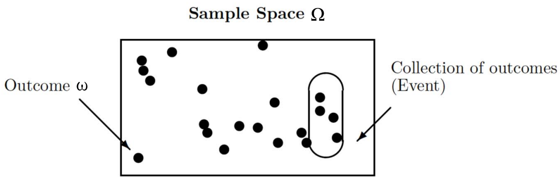


Event Algebra

An event algebra is a collection of subsets of the sample space — referred to as **events** in the sample space — chosen

- unions of events and complements of events are themselves events;
- a particular event has occurred if the outcome of the experiment lies in this event subset;
- Ω **is the "certain event"** (必然事件) because it always occurs;
- **the empty set** $\emptyset$ **is the "impossible event"** (不可能事件) because it never occurs.
- and intersections of events are also events.

Basic Concepts



Probability Measure

A probability measure associates with each event A a number $\mathrm{P}(A)$, termed the probability of A, in such a way that:

(a) $\mathrm{P}(A) \geq 0$;

(b) $\mathrm{P}(\Omega) = 1$;

(c) If $A \cap B = \emptyset$, i.e., if events A and B are mutually exclusive (不相容), then

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

Outline

Conditional Probability

The probability of event A, given that event B has occurred, is denoted by $\mathrm{P}(A|B)$.

$$\mathrm{P}(A|B) \triangleq \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \quad \text{if} \quad \mathrm{P}(B) > 0. \tag{1}$$

Knowing that B has occurred in effect reduces the sample space to the outcomes in B.

We often write $\mathrm{P}(AB)$ or $\mathrm{P}(A, B)$ for the joint probability $\mathrm{P}(A \cap B)$.

If $\mathrm{P}(B) = 0$, then the conditional probability in eqn. (1) is undefined.

By symmetry, we can also write

$$\mathrm{P}(A \cap B) = \mathrm{P}(B|A)\,\mathrm{P}(A) \tag{2}$$

Bayes' Rule (Bayes 公式)

Combining the preceding two equations

$$\mathrm{P}(A|B) \triangleq \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \quad \text{if} \quad \mathrm{P}(B) > 0.$$

and

$$\mathrm{P}(A \cap B) = \mathrm{P}(B|A)\,\mathrm{P}(A)$$

we obtain one form of **Bayes' rule (or theorem), which is at the heart of much of what we'll do with signal detection, classification, and estimation:**

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\,\mathrm{P}(B)}{\mathrm{P}(A)} \tag{3}$$

A more detailed form of Bayes'rule can be written for the conditional probability of one of a set of events $\{B_j\}$ that are mutually exclusive and collectively exhaustive, i.e. $B_l \cap B_m = \emptyset$ if $l \neq m$, and $\bigcup_j B_j = \Omega$. In this case,

$$\mathrm{P}(A) = \sum_j \mathrm{P}(A \cap B_j) = \sum_j \mathrm{P}(A|B_j)\,\mathrm{P}(B_j) \tag{4}$$

so that

$$\mathrm{P}(B_l|A) = \frac{\mathrm{P}(A|B_l)\,\mathrm{P}(B_l)}{\sum_j \mathrm{P}(A|B_j)\,\mathrm{P}(B_j)} \tag{5}$$

Independence (独立)

Events A and B are said to be independent if

$$\mathrm{P}(A|B) = \mathrm{P}(A) \tag{6}$$

or equivalently if the joint probability factors as

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B). \tag{7}$$

More generally, a collection of events is said to be mutually independent if the probability of the intersection of events from this collection, taken any number at a time, is always the product of the individual probabilities. Note that pairwise independence is not enough.

Also, two sets of events A and B are said to be independent of each other if the probability of an intersection of events taken from these two sets always factors into the product of the joint probability of those events that are in A and the joint probability of those events that are in B.

EXAMPLE: Transmission errors in a communication system

Example

A communication system transmits symbols labeled A, B, and C. Because of errors (noise) introduced by the channel, there is a nonzero probability that for each transmitted symbol, the received symbol differs from the transmitted one. The below table describes the joint probability for each possible pair of transmitted and received symbols under a certain set of system conditions.

	Symbol received		
Symbol sent	A	B	C
A	0.05	0.10	0.09
B	0.13	0.08	0.21
C	0.12	0.07	0.15

Example

	Symbol received		
Symbol sent	A	B	C
A	0.05	0.10	0.09
B	0.13	0.08	0.21
C	0.12	0.07	0.15

For notational convenience let's use A_s, B_s, C_s to denote the events that A, B or C respectively is sent, and A_r, B_r, C_r to denote A, B or C respectively being received.

So, for example, $\mathrm{P}(A_r, B_s) = 0.13$ and $\mathrm{P}(C_r, C_s) = 0.15$.

Example

	Symbol received		
Symbol sent	A	B	C
A	0.05	0.10	0.09
B	0.13	0.08	0.21
C	0.12	0.07	0.15

Questions:

1. to determine the marginal probability $\mathrm{P}(A_r)$,
2. to determine the marginal probability $\mathrm{P}(A_s)$,
3. to determine the conditional probability, that C was sent, given that B was received, i.e., $\mathrm{P}(C_s|B_r)$,
4. and to measure or calculate the probability of a transmission error P_t.

Solution

1: *To determine the marginal probability* $\mathrm{P}(A_r)$*, we sum the probabilities for all the mutually exclusive ways that* A *is received,*

$$\begin{aligned}\mathrm{P}(A_r) &= \mathrm{P}(A_r, A_s) + \mathrm{P}(A_r, B_s) + \mathrm{P}(A_r, C_s) \\ &= 0.05 + 0.13 + 0.12 = 0.3\end{aligned}$$

2: *We can determine the marginal probability* $\mathrm{P}(A_s)$ *as*

$$\mathrm{P}(A_s) = \mathrm{P}(A_r, A_s) + \mathrm{P}(B_r, A_s) + \mathrm{P}(C_r, A_s) = 0.24$$

Solution

3: *The conditional probability* $\mathrm{P}(C_s|B_r)$ *can be expressed as*

$$\mathrm{P}(C_s|B_r) = \frac{\mathrm{P}(C_s, B_r)}{\mathrm{P}(B_r)}$$

Because $\mathrm{P}(C_s, B_r) = 0.07$ *and the denominator is calculated as*

$$\mathrm{P}(B_r) = \mathrm{P}(B_r, A_s) + \mathrm{P}(B_r, B_s) + \mathrm{P}(B_r, C_s) = 0.25$$

Therefore $\mathrm{P}(C_s|B_r) = 0.28$*.*

4: *The probability of a transmission error,* P_t*, would correspond to any of the following mutually exclusive events happening:*

$$(A_s \cap B_r), (A_s \cap C_r), (B_s \cap A_r), (B_s \cap C_r), (C_s \cap A_r), (C_s \cap B_r)$$

P_t *is therefore the sum of the probabilities of these six mutually exclusive events, yielding* $\mathrm{P}_t = 0.72$*.*

Outline

Random Variables

A real-valued random variable $X(\cdot)$ is a function that maps each outcome ω of a probabilistic experiment to a real number $X(\omega)$, which is termed the **realization** of (or value taken by) the random variable in that experiment.

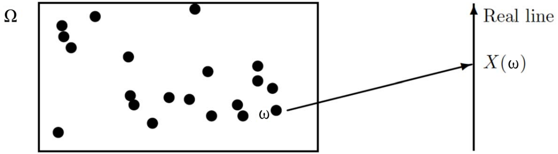


Figure: Random variable.

We shall typically just write the random variable as X instead of $X(\cdot)$ or $X(\omega)$.

It is often also convenient to consider random variables taking values that are not specified as real numbers but rather a finite or countable set of labels, say $L_0, L_1, L_2, \cdots$.

For instance,

- the random status of a machine may be tracked using the labels Idle, Busy, and Failed.
- the random presence of a target in a radar scan can be tracked using the labels Absent and Present.

We refer to such random variables as random events, mapping each outcome ω of a probabilistic experiment to the label $L(\omega)$, chosen from the possible values $L_0, L_1, L_2, \ldots$. We shall typically just write L instead of $L(\omega)$.

Outline

Cumulative Distribution Functions

For a (real-valued) random variable X, the probability of the event comprising all ω for which $X(\omega) \leq x$ is described using **the cumulative distribution function (CDF)** $F_X(x)$:

$$F_X(x) = \mathrm{P}(X \leq x) \tag{8}$$

We can therefore write

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a) \tag{9}$$

- if there is a non-zero probability that X takes a specific value x_1, i.e. if $\mathrm{P}(X = x_1) > 0$,

$$F_X(x_1) - F_X(x_1-) = \mathrm{P}(X = x_1)$$

  ,
- the CDF is nondecreasing as a function of x, it starts from $F_X(-\infty) = 0$ and rises to $F_X(\infty) = 1$.

Conditional CDF

The conditional CDF $F_{X|L}(x|L = l_i)$ is used to describe the distribution of X conditioned on some random event L taking the specific value l_i, and assuming $\mathrm{P}(L = l_i) > 0$:

$$F_{X|L}(x|L = l_i) = \mathrm{P}(X \leq x|L = l_i) = \frac{\mathrm{P}(X \leq x, L = l_i)}{\mathrm{P}(L = l_i)}. \quad (10)$$

Probability Density Functions

The probability density function (PDF) $f_X(x)$ of the random variable X is the derivative of $F_X(x)$:

$$f_X(x) = \frac{dF_X(x)}{dx}. \tag{11}$$

It is of course always non-negative because $F_X(x)$ is nondecreasing.

At points of discontinuity in $F_X(x)$, corresponding to values of x that have non-zero probability of occurring, there will be (Dirac) impulses in $f_X(x)$, of strength or area equal to the height of the discontinuity.

We can write

$$\mathrm{P}(a < X \leq b) = \int_a^b f_X(x)dx. \tag{12}$$

$$P(a < X \leq b) = \int_a^b f_X(x)dx.$$

(Any impulse of $f_X(x)$ at b would be included in the integral, while any impulse at a would be left out – i.e. the integral actually goes from $a+$ to $b+$.)

We can heuristically think of $f_X(x)dx$ as giving the probability that X lies in the interval $(x - dx, x]$:

$$P(x - dx < X \leq x) \approx f_X(x)dx. \tag{13}$$

Note that at values of x where $f_X(x)$ does not have an impulse, the probability of X having the value x is zero, i.e., $P(X = x) = 0$.

A related function is the conditional PDF $f_{X|L}(x|L_i)$, defined as the derivative of $F_{X|L}(x|L_i)$ with respect to x.

Probability Mass Function

A real-valued discrete random variable X is one that takes only a finite or countable set of real values, $\{x_1, x_2, \cdots\}$.

The CDF in this case would be a "staircase" function, while the PDF would be zero everywhere, except for impulses at the x_j , with strengths corresponding to the respective probabilities of the x_j .

These strengths/probabilities are conveniently described by the probability mass function (PMF) $p_X(x)$, which gives the probability of the event $X = x_j$:

$$\mathrm{P}(X = x_j) = p_X(x_j). \tag{14}$$

Outline

Jointly CDF and PDF

The joint CDF of two random variables X and Y is

$$F_{X,Y}(x,y) = \mathrm{P}(X \le x, Y \le y). \tag{15}$$

The corresponding joint PDF is

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \tag{16}$$

and has the heuristic interpretation that

$$\mathrm{P}(x - dx < X \le x, y - dy < Y \le y) \approx f_{X,Y}(x,y) dx dy. \tag{17}$$

Marginal PDF

The marginal PDF $f_X(x)$ is defined as the PDF of the random variable X considered on its own, and is related to the joint density $f_{X,Y}(x,y)$ by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy. \tag{18}$$

A similar expression holds for the marginal PDF $f_Y(y)$.

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

Conditional CDF

When the model involves a random variable X and a random event L, we may work with the conditional CDF

$$F_{X|L}(x|l_i) = \mathrm{P}(X \leq x | L = l_i) = \frac{\mathrm{P}(X \leq x, L = l_i)}{\mathrm{P}(L = l_i)}, \tag{19}$$

provided $\mathrm{P}(L = l_i) > 0$.

The derivative of this function with respect to x gives the conditional PDF $f_{X|L}(x|l_i)$,

$$f_{X|L}(x|l_i) = \frac{f_{X,L}(x, l_i)}{f_L(l_i)}. \tag{20}$$

Using similar reasoning, we can obtain relationships such as the following:

$$\mathrm{P}(L = l_i | X = x) = \frac{f_{X|L}(x|l_i)\,\mathrm{P}(L = l_i)}{f_X(x)} \tag{21}$$

Independent variables

Two random variables X and Y are said to be independent or statistically independent if their joint PDF (or equivalently their joint CDF) factors into the product of the individual ones:

$$\begin{aligned} &f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \text{or} \\ &F_{X,Y}(x,y) = F_X(x)F_Y(y), \end{aligned} \tag{22}$$

For a set of more than two random variables to be independent, we require that the joint PDF (or CDF) of random variables from this set factors into the product of the individual PDFs (respectively, CDFs).

EXAMPLE: Independence of events

Example

Considering two independent random variables X and Y for which the marginal PDFs are uniform between zero and one:

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

$$f_Y(y) = \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

We define the events A, B, C and D as follows:

$$A = \left\{ y > \frac{1}{2} \right\}, B = \left\{ y < \frac{1}{2} \right\}, C = \left\{ x < \frac{1}{2} \right\},$$

$$D = \left\{ x < \frac{1}{2} \text{ and } y < \frac{1}{2} \right\} \cup \left\{ x > \frac{1}{2} \text{ and } y > \frac{1}{2} \right\}$$

Example

These events are illustrated pictorially in Figure as follow

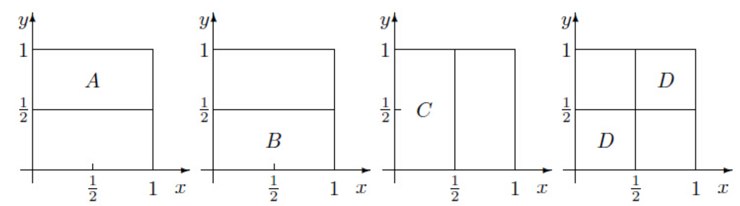


Figure: Illustration of events A, B, C, and D

Questions:

1. to calculate $\mathrm{P}(A \cap C)$
2. to calculate $\mathrm{P}(A \cap B)$
3. to judge whether A and B are independent variables
4. to prove that A, C and D are pairwise independent and aren't mutually independent.

Solution

Because X and Y are independent, the joint PDF $f_{X,Y}(x,y)$ is given by

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

1:

$$\mathrm{P}(A \cap C) = \mathrm{P}(y > \frac{1}{2}, x < \frac{1}{2}) = \mathrm{P}(A) \times \mathrm{P}(C) = \frac{1}{4}$$

2:

$$\mathrm{P}(A \cap B) = \mathrm{P}(y > \frac{1}{2}, y < \frac{1}{2}) = 0$$

3: *Since $\mathrm{P}(A) = \mathrm{P}(B) = \frac{1}{2}$, then $\mathrm{P}(A \cap B) \neq \mathrm{P}(A)\,\mathrm{P}(B)$, so events A and B are not independent.*

Solution

4: *Since there is no region where all three sets* A, C *and* D *overlap, then* $\mathrm{P}(A \cap C \cap D) = 0$
However, $\mathrm{P}(A) = \mathrm{P}(C) = \mathrm{P}(D) = \frac{1}{2}$, *so*

$$\mathrm{P}(A \cap C \cap D) \neq \mathrm{P}(A)\,\mathrm{P}(C)\,\mathrm{P}(D)$$

therefore the events A, C, and D are not mutually independent (相互独立), even though they are easily seen to be pairwise independent （两两独立）.

Outline

Expectation

Also termed the expected or mean or average value, or the first-moment of the real-valued random variable X is denoted by $E[X]$ or $\overline{X}$ or μ_X, and defined as

$$E[X] = \overline{X} = \mu_X = \int_{-\infty}^{+\infty} x f_X(x) dx \tag{23}$$

Other simple measures of where the PDF is centered or concentrated are provided

- by the **median**, which is the value of x for which $f_X(x) = 0.5$,
- and by the **mode**, which is the value of x for which$f_X(x)$ is maximum (in degenerate cases one or both of these may not be unique).

Variance

The variance or centered second-moment of the random variable X is denoted by σ_X^2 and defined as the expected squared deviation from the mean

$$\begin{aligned}\sigma_X^2 &= E[(X-\mu_X)^2] \\ &= \int_{-\infty}^{+\infty} (x-\mu_X)^2 f_X(x)dx \\ &= E[X^2] - \mu_X^2\end{aligned}$$

$$\sigma_X^2 = E[X^2] - \mu_X^2 \tag{24}$$

We refer to

- $E[X^2]$ as the second-moment of X,
- the square root of the variance, termed the standard deviation, is a widely used measure of the spread of the PDF.

EXAMPLE 1.3: Gaussian and uniform random variables

Example

Two common PDF's that we will work with are the Gaussian (or normal) density and the uniform density:

$$\text{Gaussian:} \quad f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\text{Uniform:} \quad f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

The two parameters μ and σ that define the Gaussian PDF can be shown to be its mean and standard deviation respectively.

Similarly, though the uniform density can be simply parametrized by its lower and upper limits a and b as above, an equivalent parametrization is via its mean $\mu = \frac{(a+b)}{2}$ and standard deviation $\sigma = \sqrt{\frac{(b-a)^2}{12}}$.

Chebyshev inequality

$$\mathrm{P}\left(\frac{|X - \mu_X|}{\sigma_X} \geq k\right) \leq \frac{1}{k^2} \qquad (25)$$

This inequality implies that, for any random variable, the probability it lies at or more than 3 standard deviations away from the mean (on either side of the mean) is not greater than $(\frac{1}{3^2}) = 0.11$.

Of course, for particular PDFs, much more precise statements can be made, and conclusions derived from the Chebyshev inequality can be very conservative.

- in the case of a Gaussian PDF, the probability of being more than 3 standard deviations away from the mean is only 0.0026,
- for a uniform PDF the probability of being more than even 2 standard deviations away from the mean is precisely 0.

Conditional Expectation

The conditional expectation of the random variable X, given that the random variable Y takes the value y, is the real number

$$E[X|Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx = g(y), \qquad (26)$$

i.e., this conditional expectation takes some value $g(y)$ when $Y = y$.

We may also consider the random variable $g(Y)$, namely the function of the random variable Y that, denote $g(Y)$ by $E[X|Y]$.

Note that the expectation $E[g(Y)]$ of the random variable $g(Y)$, i.e. the iterated expectation $E[E[X|Y]]$, is well defined.

How to compute $E[X]$ when we have the joint PDF $f_{X,Y}(x,y)$

One way is to evaluate the marginal density $f_X(x)$ of X, and then use the definition of expectation:

$$E[X] = \int_{-\infty}^{+\infty} x(\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy)dx. \quad (27)$$

However, it is often simpler to compute the conditional expectation of X, given $Y = y$, then average this conditional expectation over the possible values of Y , using the marginal density of Y. To derive this more precisely, recall that

$$f_{X,Y} = f_{X|Y}(x|y)f_Y(y). \quad (28)$$

and use this in (27) to deduce that

$$\begin{aligned} E[X] &= \int_{-\infty}^{+\infty} f_Y(y)(\int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx)dy \\ &= E_Y[E_{X|Y}[X|Y]] \end{aligned} \quad (29)$$

More simply, one writes

$$E[X] = E[E[X|Y]] = \int_{-\infty}^{+\infty} h(y) f_Y(y) dy \tag{30}$$

This shows that we only need $f_Y(y)$ to calculate the expectation of a function of Y; to compute the expectation of $X = h(Y)$, we do not need to determine $f_X(x)$.

Similarly, if X is a function of two random variables, $X = h(Y, Z)$, then

$$E[X] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(y, z) f_{Y,Z}(y, z) dydz. \tag{31}$$

It is easy to show from this that if Y **and** Z **are independent**, and if $h(y, z) = g(y)l(z)$, then

$$E[g(Y)l(Z)] = E[g(Y)]E[l(Z)]. \tag{32}$$

Outline

Jointly distributed random variables

Consider a pair of jointly distributed random variables X and Y . Their marginal PDFs are simply obtained by projecting the probability mass along the y-axis and x-axis directions respectively:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy \tag{33}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx. \tag{34}$$

In other words, the PDF of X is obtained by integrating the joint PDF over all possible values of the other random variable Y and similarly for the PDF of Y.

Mean value of bivariate random varialbes

The mean value of the bivariate PDF is specified by giving the mean values of each of its two component random variables:

- the mean value has an x component that is $E[X]$,
- a y component that is $E[Y]$,
- and these two numbers can be evaluated from the respective marginal densities.

The center of mass of the bivariate PDF is thus located at

$$(x, y) = (E[X], E[Y]). \tag{35}$$

Standard deviation of bivariate random varialbes

A measure of the spread of the bivariate PDF in the x direction may be obtained from the standard deviation σ_X of X, computed from $f_X(x)$;

and a measure of the spread in the y direction may be obtained from σ_Y , computed similarly from $f_Y(y)$.

However, these two numbers clearly only offer a partial view. We would really like to know what the spread is in a general direction rather than just along the two coordinate axes. We can consider, for instance, the standard deviation (or, equivalently, the variance) of the random variable Z defined as

$$Z = \alpha X + \beta Y. \tag{36}$$

for arbitrary constants α and β.

Note that by choosing α and β allows us to specify the behavior in all directions.

Z behaves

- $Z = 0$ when $\alpha X + \beta Y = 0$. This is the equation of a straight line through the origin in the (x, y) plane, a line that indicates the precise combinations of values x and y that contribute to determining $f_Z(0)$, by projection of $f_{X,Y}(x, y)$ along the line. Let us call this the reference line.
- If Z now takes a nonzero value z, the corresponding set of (x, y) values lies on a line offset from but parallel to the reference line. We project $f_{X,Y}(x, y)$ along this new offset line to determine $f_Z(z)$.

Variance of Z

The mean of Z is easily found in terms of quantities we have already computed, namely $E[X]$ and $E[Y]$:

$$E[Z] = \alpha E[X] + \beta E[Y]. \tag{37}$$

As for the variance of Z, it is easy to establish from (36) and (37) that

$$\sigma_Z^2 = E[Z^2] - (E[Z])^2 = \alpha^2 \sigma_X^2 + \beta^2 \sigma_Y^2 + 2\alpha\beta C_{X,Y} \tag{38}$$

or equivalently

$$C_{X,Y} = E[XY] - E[X]E[Y]. \tag{39}$$

The quantity $E[XY]$ that appears in (39), i.e., the expectation of the product of the random variables, is referred to as the correlation or second cross-moment of X and Y, and will be denoted by $R_{X,Y}$:

$$R_{X,Y} = E[XY]. \tag{40}$$

It is reassuring to note from (40) that the covariance $C_{X,Y}$ is the only new quantity needed when going from mean and spread computations along the coordinate axes to such computations along any axis.

Correlation Coefficient $\rho_{X,Y}$

In summary,

- we can express the location of $f_{X,Y}(x, y)$ in an aggregate or approximate way in terms of the 1st-moments, $E[X]$, $E[Y]$;
- and we can express the spread around this location in an aggregate or approximate way in terms of the (central) 2nd-moments, σ_X^2 , σ_Y^2 , $C_{X,Y}$.

It is common to work with a normalized form of the covariance, namely the correlation coefficient $\rho_{X,Y}$:

$$\rho_{X,Y} = \frac{C_{X,Y}}{\sigma_X \sigma_Y}. \tag{41}$$

This normalization ensures that the correlation coefficient is unchanged if X and/or Y is multiplied by any nonzero constant or has any constant added to it.

Example

For instance, the centered and normalized random variables

$$V = \frac{X - \mu_X}{\sigma_X}, \quad W = \frac{Y - \mu_Y}{\sigma_Y}. \tag{42}$$

each of which has mean 0 and variance 1, have the same correlation coefficient as X and Y .

The correlation coefficient might have been better called the covariance coefficient, since it is defined in terms of the covariance and not the correlation of the two random variables, but this more helpful name is not generally utilized.

Correlation Coefficient $\rho_{X,Y}$

$$|\rho_{X,Y}| \leq 1. \tag{43}$$

From the various preceding definitions, a positive correlation $R_{X,Y} > 0$ suggests that X and Y tend to take the same sign, on average, whereas a positive covariance $C_{X,Y} > 0$ or equivalently a positive correlation coefficient $\rho_{X,Y} > 0$ suggests that the deviations of X and Y from their respective means tend to take the same sign, on average.

Conversely, a negative correlation suggests that X and Y tend to take opposite signs, on average, while a negative covariance or correlation coefficient suggests that the deviations of X and Y from their means tend to take opposite signs, on average.

Uncorrelated and Independent

The random variables X and Y are said to be uncorrelated (or linearly independent, a less common and potentially misleading term) if

$$E[XY] = E[X]E[Y], \tag{44}$$

or equivalently if

$$C_{X,Y} = 0 \;\; or \;\; \rho_{X,Y} = 0. \tag{45}$$

Thus uncorrelated does not mean zero correlation (unless one of the random variables has an expected value of zero).

Rather, uncorrelated means zero covariance.

Independent random variables X and Y, i.e., those for which

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \tag{46}$$

are always uncorrelated,

but the converse is not generally true: uncorrelated random variables may not be independent.

If X and Y are independent, then $E[XY] = E[X]E[Y]$ so X and Y are uncorrelated.

The converse does not hold in general.

Example

For instance, consider the case where the combination (X, Y) takes only the values (1, 0), (-1, 0), (0, 1) and (0,-1), each with equal probability $\frac{1}{4}$. Then X and Y are easily seen to be uncorrelated but dependent, i.e., not independent.

Orthogonal

Two random variables X and Y are orthogonal if

$$E[XY] = 0$$

EXAMPLE: Perfect correlation, zero correlation

Example

Consider the degenerate case where Y is given by a deterministic linear function of a random variable X (so Y is also a random variable, of course):

$$Y = \xi X + \zeta \tag{47}$$

where ξ and ζ are constants. Then it is easy to show that $\rho_{X,Y} = 1$ if $\xi > 0$ and $\rho = -1$ if $\xi < 0$.

It is easy to establish that $\rho_{X,Y} = 0$ if

$$Y = \xi X^2 + \zeta \tag{48}$$

and X has a PDF $f_X(x)$ that is even about 0, i.e.,

$$f_X(-x) = f_X(x).$$

PDF of Bivariate Gaussian density

The random variables X and Y are said to be bivariate Gaussian or bivariate normal if their joint PDF is given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\{\frac{1}{2(1-\rho^2)}(v^2 - 2\rho vw + w^2)\}$$

where $v = \frac{x-\mu_X}{\sigma_X}$ and $w = \frac{y-\mu_Y}{\sigma_Y}$.

Properties of Bivariate Gaussian density

This density is the natural bivariate generalization of the familiar Gaussian density, and has several nice properties:

- The marginal densities of X and Y are Gaussian.
- The conditional density of Y , given $X = x$, is Gaussian with mean μ_X and variance $\sigma_Y^2(1-\rho^2)$ (which evidently does not depend on the value of x); and Y similary for the conditional density of X, given $Y = y$.
- If X and Y are uncorrelated, i.e., if $\rho = 0$, then X and Y are actually independent.
- Any two affine (i.e., linear plus constant) combinations of X and Y are themselves bivariate Gaussian (e.g., $Q = X + 3Y + 2$ and $R = 7X + Y - 3$ are bivariate Gaussian).

Joint PDF

Some of the generalizations of the preceding discussion from two random variables to many random variables are fairly evident. In particular, the mean of a joint PDF

$$f_{X_1,X_2,...,X_l}(x_1, x_2, ..., x_l) \tag{49}$$

in the l-dimensional space of possible values has coordinates that are the respective individual means, $E[X_1], \cdots, E[X_l]$.

The spreads in the coordinate directions are deduced from the individual (marginal) spreads, σ_{X_1} ,..., σ_{X_l}.

To be able to compute the spreads in arbitrary directions, we need all the additional $\frac{l(l-1)}{2}$ central 2nd moments, namely σ_{X_i,X_j} for all $1 \leq i < j \leq l$ (note that $\sigma_{X_j,X_i} = \sigma_{X_i,X_j}$) but nothing more.

Outline

Vector-space picture

To develop the vector-space picture, we represent the random variables X and Y as vectors X and Y in some abstract vector space. For the squared lengths of these vectors, we take the second-moments of the associated random variables, $E[X^2]$ and $E[Y^2]$ respectively.

Recall that in Euclidean vector space the squared length of a vector is the inner product of the vector with itself. This suggests that perhaps in our vector space interpretation the inner product $\langle X, Y \rangle$ between two general vectors X and Y should be defined as the correlation (or second cross-moment) of the associate random variables:

$$\langle X, Y \rangle = E[X, Y] = R_{X,Y}. \tag{50}$$

This indeed turns out to be the definition that's needed.

Properties of inner product

With this definition, the standard properties required of an inner product in a vector space are satisfied, namely:

Symmetry: $\langle X, Y\rangle = \langle Y, X\rangle$.

Linearity: $\langle X, \alpha_1 Y_1 + \alpha_2 Y_2\rangle = \alpha_1\langle X, Y_1\rangle + \alpha_2\langle X, Y_2\rangle$

Positivity: $\langle X, Y\rangle$ is positive for $X \neq 0$, and 0 otherwise.

This definition of inner product is also consistent with the fact that we often refer to two random variables as orthogonal when $E[XY] = 0$.

Covariance of the random variables

The centered random variables $X - \mu_X$ and $Y - \mu_Y$ can similarly be represented as vectors $\tilde{X}$ and $\tilde{Y}$ in this abstract vector space, with squared lengths that are now the variances of the random variables X and Y:

$$\sigma_X^2 = E[(X - \mu_X)^2], \qquad \sigma_Y^2 = E[(Y - \mu_Y)^2]. \tag{51}$$

respectively.

The lengths are therefore the standard deviations of the associated random variables, σ_X and σ_Y respectively. The inner product of the vectors $\tilde{X}$ and $\tilde{Y}$ becomes

$$\langle \tilde{X}, \tilde{Y} \rangle = E[(X - \mu_X)(Y - \mu_Y)] = C_{X,Y}, \tag{52}$$

namely the covariance of the random variables.

Correlation coefficient of the two random variables

In Euclidean space the inner product of two vectors is given by the product of the lengths of the individual vectors and the cosine of the angle between them:

$$\langle \tilde{X}, \tilde{Y} \rangle = C_{X,Y} = \sigma_X \sigma_Y \cos(\theta), \tag{53}$$

so the quantity

$$\theta = \cos^{-1}\left(\frac{C_{X,Y}}{\sigma_X \sigma_Y}\right) = \cos^{-1} \rho \tag{54}$$

can be thought of as the angle between the vectors. Here is the correlation coefficient of the two random variables, so evidently

$$\rho = \cos(\theta). \tag{55}$$

Thus, the correlation coefficient is the cosine of the angle between the vectors.

It is therefore not surprising at all that

$$-1 \leq \rho \leq 1. \tag{56}$$

When ρ is near 1, the vectors are nearly aligned in the same direction, whereas when ρ is near -1 they are close to being oppositely aligned.

The correlation coefficient is zero when these vectors $\tilde{X}$ and $\tilde{Y}$ (which represent the centered random variables) are orthogonal, or equivalently, the corresponding random variables have zero covariance,

$$C_{X,Y} = 0. \tag{57}$$

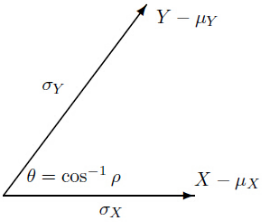


Figure: Random Variables as Vectors.