Review of Probability Theory

Tadao Hoshino （星野匡郎）

ver. 2018 Fall Semester

Random Variable

Random Variable

- A random variable is a variable whose values are determined by chance.
- Examples of random variables are:
  - the outcome of rolling a dice;
  - the number of heads in 2 flips of a coin;
  - the number of car accidents in a day;
  - the number of students in this class.
- A more formal definition of random variable is as follows:

Random Variable

A random variable is a function that maps the outcomes of a random experiment to numerical values.

Random Variable

- It is important to note that a random variable is not the same as a random number.
- A random variable is a "rule" (i.e., function) that associates a number with each outcome in the set of possible outcomes. The set of possible outcomes is called the sample space.
- In the following, we use a capital letter, say X, to denote a random variable. A realized value of X is denoted by a small letter x.
- The set of possible realizations of a random variable X, $\{x_1, x_2, ..., \}$, is called the support, denoted by $\mathcal{X}$

Random variable X	Sample space	Support $\mathcal{X}$
Outcome of rolling a dice	$\{1,2,3,4,5,6\}$	$\{1,2,3,4,5,6\}$
# of heads in 2 flips of a coin	$\{HH, TH, HT, TT\}$	$\{0,1,2\}$

Random Variable

- There are two types of random variables:

Discrete Random Variable

A discrete random variable is a random variable where its support is a discrete set.

Continuous Random Variable

A continuous random variable is a random variable where its support is an interval (or a collection of intervals).

- Intuitively, a discrete random variable is a random variable where the values in its support can be written down in a list.
- For example, letting X denote the height (in meters) of a randomly selected person, then X is a continuous random variable.
- When X is continuous, there is an infinite number of values in the support of X.

Cumulative Distribution Function: CDF

The probability that the random variable X takes a value less than or equal to x, $\Pr(X \leq x)$, is called the cumulative distribution function (CDF), or simply distribution function, of X, and we denote

$$F(x) = \Pr(X \leq x).$$

NOTE: x can be any value, even outside the support $\mathcal{X}$.

Probability Distribution

Properties of CDF

- For any X, the probability that X takes a value smaller (larger) than negative (positive) infinity is zero (one):

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to \infty} F(x) = 1$$

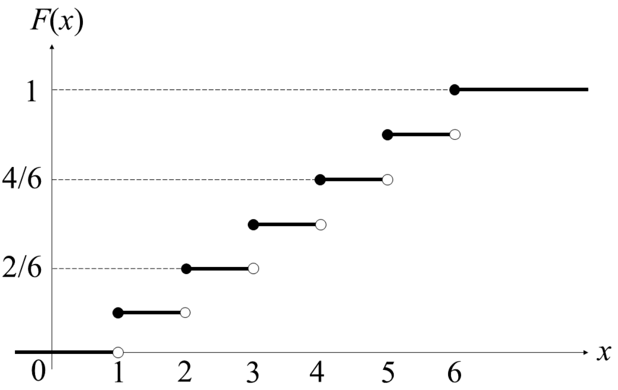
- CDF is monotonically non-decreasing :

$$x_1 \leq x_2 \Rightarrow \Pr(X \leq x_1) \leq \Pr(X \leq x_2) \iff F(x_1) \leq F(x_2)$$

- In the following, we will mainly consider only continuous random variables with continuous CDF, rather than discrete random variables.
- In the case of discrete random variables, the corresponding CDF is always a discontinuous step-function.

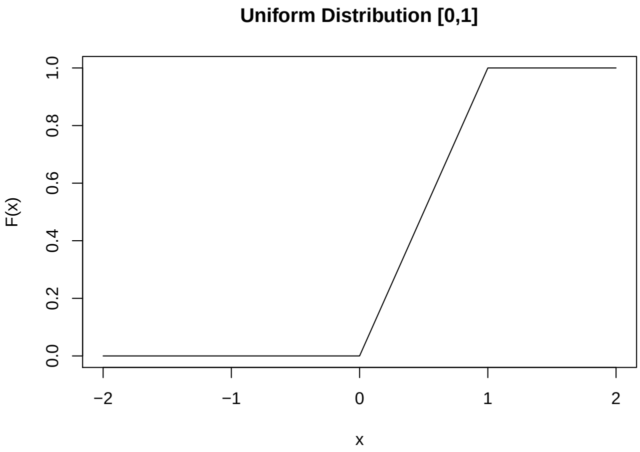
Examples of CDF's

Outcome of rolling a dice = discrete uniform distribution on $\{1, 2, ..., 6\}$



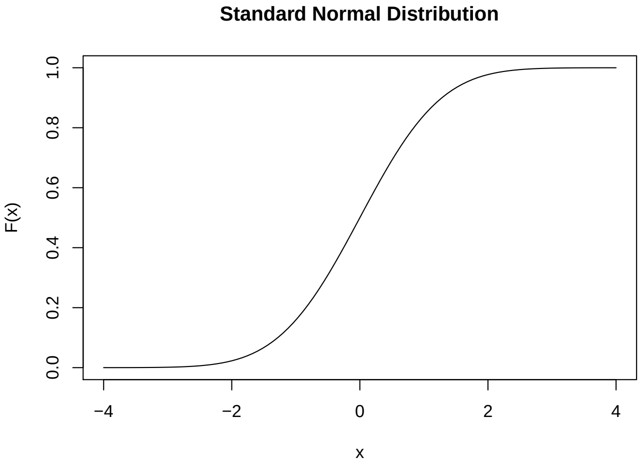
Examples of CDF's

Continuous uniform distribution on $[0, 1]$



Examples of CDF's

Standard normal distribution, with mean 0 and the standard deviation 1.



Probability Density Function

- The probability that X takes a value within an interval $[a, b]$ is given by

$$\begin{aligned}\Pr(X \in [a,b]) = \Pr(a \leq X \leq b) &= \Pr(X \leq b) - \Pr(X \leq a) \\ &= F(b) - F(a)\end{aligned}$$

- For a continuous random variable X, what is the probability of that X takes a specific value, say $\Pr(X = 1)$?
- Intuitively, it is expected that the probability $\Pr(X = 1)$ can be approximated by the probability $\Pr(X \in [1, 1 + h])$ for sufficiently small $h > 0$:

$$\begin{aligned}\Pr(X = 1) &\approx \Pr(X \in [1, 1 + h]) \\ &= F(1 + h) - F(1)\end{aligned}$$

Probability Density Function

- For example, let $F(\cdot)$ be the standard normal distribution function. Then, it holds that

$$
\begin{aligned}
\Pr(X \in [1, 1 + \overbrace{1.00}^{h}]) &= F(2.00) - F(1) = 0.1359 \\
\Pr(X \in [1, 1 + 0.50]) &= F(1.50) - F(1) = 0.0918 \\
\Pr(X \in [1, 1 + 0.10]) &= F(1.10) - F(1) = 0.0230 \\
\Pr(X \in [1, 1 + 0.01]) &= F(1.01) - F(1) = 0.0024
\end{aligned}
$$

- As h approaches to zero, we will get a smaller and smaller probability; clearly in the limit the probability will be exactly 0.
- This is due to the continuity of $F(\cdot)$. That is, the continuity of $F(\cdot)$ implies that

$$a \approx b \Rightarrow F(a) \approx F(b)$$

- In fact, $\Pr(X = x) = 0$ holds $\underline{\text{for any } x}$ if X is continuous.

Probability Density Function

- On the other hand, if we divide $F(x+h) - F(x)$ by h, the value does not degenerate to zero but converges to a specific value. In fact,

	$F(1+h) - F(1)$	$\frac{F(1+h)-F(1)}{h}$
$h = 0.1$	0.0230	0.2299
$h = 0.01$	0.0024	0.2408
$h = 0.001$	0.0002	0.2418

- By taking the limit $h \to 0$, we can define the probability density function.

Probability Density Function: PDF

We define the probability density function (PDF), or simply density function, of X as the derivative of the CDF of X:

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \frac{\partial F(x)}{\partial x}$$

Why "density" function?

- We can interpret $\frac{F(x+h)-F(x)}{h}$ as the "mass density" of the event $\{X \in [x, x+h]\}$, where the mass is measured in probability of the event:

$$\text{mass density} = \frac{\text{mass}}{\text{volume}}$$

$$\frac{F(x+h)-F(x)}{h} = \frac{\Pr(X \in [x, x+h])}{h}$$

- Thus, the probability density function $f(x)$ is the mass density of the event $\{X = x\}$.

Probability Density Function

- More intuitively, one can interpret the value of the density $f(x)$ as

  How likely X takes a value in the neighborhood of $x \in \mathcal{X}$

- However, note that $f(x)$ is not a probability, and thus its value can be larger than one (but never be negative).

When X is a discrete random variable, we can easily compute the probability $\Pr(X = x)$ for any $x \in \mathcal{X}$. The function $p(x) = \Pr(X = x)$ is called the probability function.

X	How to measure the likelihood of the event $\{X = x\}$
Discrete	probability function $p(x)$
Continuous	probability density function $f(x)$

Probability Density Function

Since PDF is the derivative of CDF, we obtain the following properties.

Properties of PDF

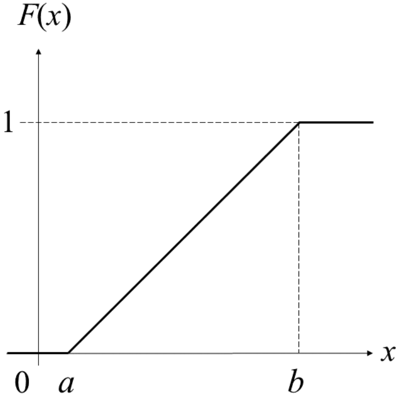
- $F(x)$ is non-decreasing $\iff f(x) \geq 0$ for all x
- $F(a) = \int_{-\infty}^{a} f(x)dx$. Thus, the probability that X takes a value in $[a, b]$ is

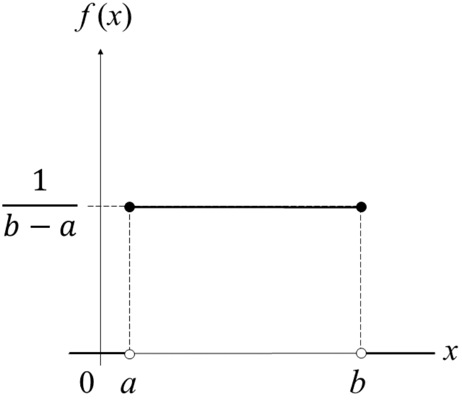
$$\begin{aligned}\Pr(X \in [a,b]) &= F(b) - F(a) \\ &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx\end{aligned}$$

- $F(\infty) = 1 \iff \int_{-\infty}^{\infty} f(x)dx = 1$

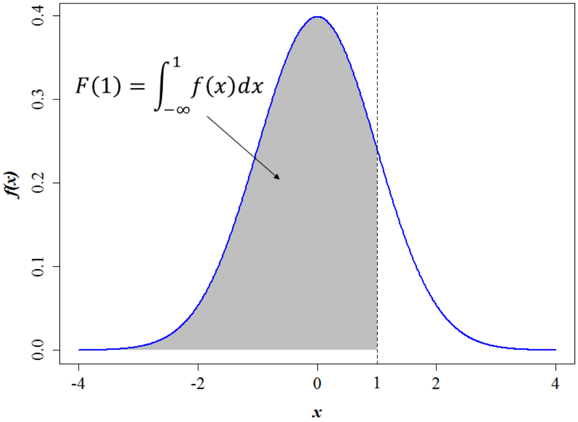
Examples of CDF and corresponding PDF

Continuous uniform distribution on $[a, b]$





Standard normal distribution, with mean 0 and the standard deviation 1.



Exercise 1

Let X be a continuous random variable whose PDF is

$$f(x) = \begin{cases} \frac{2}{3} - \frac{2}{9}x & \text{if } x \in [0,3] \\ 0 & \text{otherwise} \end{cases}$$

1. Derive the CDF $F(\cdot)$ of X, and draw a graph of $F(\cdot)$.
2. Calculate the probability $\Pr(X \leq 1)$.
3. Calculate the probability $\Pr(X > 2)$.

Expectation and Variance

Expectation

Let X be a continuous random variable with density function $f(x)$. The expectation of X, also called the mean of X, is defined by

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

In the case of discrete random variable X, the expectation of X is

$$E[X] = \sum_{i=1}^{k} x_i p(x_i),$$

where $p(x_i) = \Pr(X = x_i)$, and $\{x_1, ..., x_k\} = \mathcal{X}$.

The expectation is a weighted average of the values of X where the weights are the probabilities attached to those values.

Derivation of $E[X]$ for continuous X

- Let X be a continuous random variable with CDF $F(x)$ and PDF $f(x)$. Recall: the probability that X takes a value in $[a, b]$ is

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

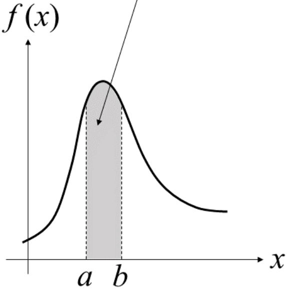
- As shown in the figure (next page), this probability can be approximated by

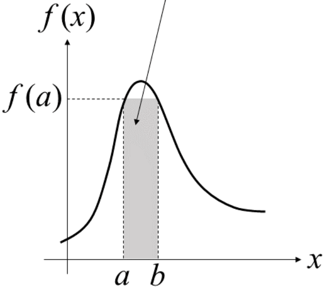
$$\Pr(a \leq X \leq b) \approx f(a)(b - a) \tag{1}$$

if $(b - a)$ is sufficiently small.

$$\Pr(a \leq X \leq b) = F(b) - F(a)$$

$$= \int_a^b f(x)dx \qquad \approx \qquad f(a)(b-a)$$





Derivation of $E[X]$ for continuous X

- Suppose that the support of X is given by a closed interval, say $\mathcal{X} = [x_{\min}, x_{\max}]$.
- Partition the interval $[x_{\min}, x_{\max}]$ into n subintervals of equal length

$$x_{\min} = x_1 < x_2 < \cdots < x_{n+1} = x_{\max}$$
$$\text{such that } x_{i+1} - x_i = h \text{ for all } i = 1, ..., n.$$

- Recall: $\Pr(X = x) \approx \Pr(X \in [x, x+h])$ for sufficiently small h.
- Thus, by analogy from the definition of $E[X]$ for discrete X, it holds that

$$E[X] \approx \sum_{i=1}^{n} x_i \Pr(x_i \leq X \leq x_{i+1})$$

(note: $x_{i+1} = x_i + h$).

Derivation of $E[X]$ for continuous X

- Further, by the approximation (1), we have

$$E[X] \approx \sum_{i=1}^{n} x_i f(x_i) h. \tag{2}$$

The above approximation becomes more accurate as h tends to zero, or, equivalently, as n increases to infinity.

- Thus by taking the limit as $n \to \infty$, we obtain

$$\begin{aligned} E[X] = \lim_{n\to\infty} \sum_{i=1}^{n} x_i f(x_i) h &= \int_{x_{\min}}^{x_{\max}} x f(x) dx \\ &= \int_{-\infty}^{\infty} x f(x) dx, \end{aligned}$$

where the last equality follows from the fact that $f(x) = 0$ for $x \notin [x_{\min}, x_{\max}]$.

Expectation

If X is distributed as uniform on $[a,b]$, its expectation is obtained by $E[X] = (b+a)/2$.

Proof : The PDF of uniform distribution on $[a,b]$ is

$$f(x) = \begin{cases} (b-a)^{-1} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

By definition of expectation, we have

$$\begin{aligned} E[X] &= \int_a^b x \cdot (b-a)^{-1} \cdot dx \\ &= [0.5 \cdot x^2]_a^b \cdot (b-a)^{-1} \\ &= 0.5(b^2 - a^2) \cdot (b-a)^{-1} = (b+a)/2 \quad \blacksquare \end{aligned}$$

NOTE : In general, when a random variable X follows a symmetric distribution, the expectation $E[X]$ coincides with the center of symmetry.

Linearity of expectation

For a random variable X and constants a and b,

$$E[aX+b] = aE[X] + b$$

This property is called linearity.

Proof : By the definition of expectation

$$\begin{aligned} E[aX+b] &= \int_{-\infty}^{\infty} (ax+b)f(x)dx \\ &= \underbrace{a\int_{-\infty}^{\infty} xf(x)dx}_{E[X]} + \underbrace{b\int_{-\infty}^{\infty} f(x)dx}_{1} \\ &= aE[X] + b \quad \blacksquare \end{aligned}$$

Expectation of a function of a random variable

For a random variable X, the expectation of $g(X)$, $E[g(X)]$, is given by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Proof* : Let $Y = g(X)$, and suppose that $g(\cdot)$ is monotonic such that $Y = y \iff X = g^{-1}(y)$. Then, the CDF of Y is given by

$$F_Y(y) = \Pr(Y \leq y) = \Pr(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Noting that $\partial g^{-1}(y)/\partial y = 1/g'(g^{-1}(y))$, the PDF of Y is

$$f_Y(y) = F'_Y(y) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} = \frac{f_X(x)}{g'(x)}.$$

By the change of variable $y = g(x)$,

$$E[Y] = \int_{-\infty}^{\infty} y f_y(y)dy = \int_{-\infty}^{\infty} g(x)f_y(g(x))g'(x)dx = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

When $g(\cdot)$ is non-monotonic, one can split the function into monotonic segments, and applying the above argument to each segment repeatedly leads to the desired result. ■

- An important special case is when $g(X)$ is an indicator function:

$$g(X) = \mathbf{1}\{X \leq a\} = \begin{cases} 1 & \text{if } X \leq a \\ 0 & \text{if } X > a \end{cases}$$

- In this case, we have $E(\mathbf{1}\{X \leq a\}) = F(a)$:

$$\begin{aligned} E(\mathbf{1}\{X \leq a\}) &= \int_{-\infty}^{\infty} \mathbf{1}\{x \leq a\} f(x) dx \\ &= \int_{-\infty}^{a} 1 \cdot f(x) dx + \int_{a}^{\infty} 0 \cdot f(x) dx = F(a). \end{aligned}$$

Variance

- For a random variable X, the variance of X is defined by

$$\begin{aligned} V(X) &= E[\{X - E(X)\}^2] \\ &= E(X^2) - (E[X])^2 \end{aligned}$$

- In particular, if $E(X) = 0$, $V(X) = E(X^2)$.

NOTE : Suppose there is a sample $\{X_1, ..., X_n\}$ of n observations. Then, the sample variance of the data is defined by

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

The definition of the variance of a random variable is obtained by replacing the sample average $n^{-1}\sum_{i=1}^{n}$ with the expectation E.

Standard deviation

The square root of the variance of X, $\sqrt{V(X)}$, is called the standard deviation of X.

It is often convenient to standardize X by subtracting its expectation $E(X)$ and dividing by its standard deviation $\sqrt{V(X)}$. The standardized random variable has mean zero and variance (and also standard deviation) one:

$$
\begin{aligned}
Z &= \frac{X - E(X)}{\sqrt{V(X)}} \\
E(Z) &= E\left[\frac{X - E(X)}{\sqrt{V(X)}}\right] = \frac{E(X) - E(X)}{\sqrt{V(X)}} = 0 \\
V(Z) &= \frac{E[\{X - E(X)\}^2]}{V(X)} = \frac{V(X)}{V(X)} = 1
\end{aligned}
$$

Exercise

Exercise 2

1. Let X be a continuous random variable whose PDF is

$$f(x) = \begin{cases} \frac{2}{3} - \frac{2}{9}x & \text{if } x \in [0,3] \\ 0 & \text{otherwise} \end{cases}$$

   Compute $E(X)$ and $V(X)$.
2. Let a and b constants. Derive the variance of $aX + b$.
3. Let a be a constant. Derive the variance of $(X + a)/\sqrt{V(X)}$.

Multiple Random Variables

Conditional probability and joint probability

- Consider two events, A and B. For example,
  - A: a randomly selected person has a part-time job.
  - B: a randomly selected person is an undergraduate student.
- **Conditional probability** $\Pr(A|B)$: the probability of selecting a person who has a part-time job given that she is an undergraduate student.
  $=$ the proportion of part-time workers among undergraduate students.
- **Joint probability** $\Pr(A, B)$: the probability of selecting a person who has a part-time job and is an undergraduate student.
  $=$ the proportion of part-time workers who are undergraduate students among the whole population.

Conditional Distribution Function

For a pair of random variables (X, Y) and constants (x, y), the conditional probability of $Y \leq y$ given $X = x$ is called the conditional distribution function, and it is denoted by $F_{Y|X}(y|X = x)$:

$$F_{Y|X}(y|X = x) = \Pr(Y \leq y|X = x)$$

For example, let $Y =$ annual income in million JPY, $X =$ education in years. $Y = 2 \Rightarrow$ anninc 2 mill JPY, $X = 9 \Rightarrow$ junior high graduate, etc

- $F_{Y|X}(2|X = 9) =$ the probability that anninc is less than or equal to 2 mill JPY for junior high graduates.
- $F_{Y|X}(5|X = 16) =$ the probability that anninc is less than or equal to 5 mill JPY for college graduates.

Conditional Density Function

The derivative of the conditional distribution function $F_{Y|X}(y|X=x)$ with respect to y is called the conditional density function, and it is denoted by $f_{Y|X}(y|X=x)$:

$$
\begin{aligned}
f_{Y|X}(y|X=x) &= \lim_{h \to 0} \frac{F_{Y|X}(y+h|X=x) - F_{Y|X}(y|X=x)}{h} \\
&= \frac{\partial F_{Y|X}(y|X=x)}{\partial y}
\end{aligned}
$$

- For example, you can interpret
  - $f_{Y|X}(2|X=9) =$ how the event $\{\text{anninc} = 2 \text{ mill JPY}\}$ is likely to occur for junior high graduates.
  - $f_{Y|X}(5|X=16) =$ how the event $\{\text{anninc} = 5 \text{ mill JPY}\}$ is likely to occur for college graduates.

Conditional Density Function

Since the conditional density function is the derivative of the conditional distribution function, it has the following properties:

- $F_{Y|X}(y|X=x)$ is non-decreasing in $y \iff f_{Y|X}(y|X=x) \geq 0$ for all y
- $F_{Y|X}(a|X=x) = \int_{-\infty}^{a} f_{Y|X}(y|X=x)dy$. Thus, the probability that Y takes a value in $[a,b]$ conditional on $X=x$ is

$$\begin{aligned}
\Pr(Y \in [a,b]|X=x) &= F_{Y|X}(b|X=x) - F_{Y|X}(a|X=x) \\
&= \int_{-\infty}^{b} f_{Y|X}(y|X=x)dy - \int_{-\infty}^{a} f_{Y|X}(y|X=x)dy \\
&= \int_{a}^{b} f_{Y|X}(y|X=x)dy
\end{aligned}$$

- $F_{Y|X}(\infty|X=x) = 1 \iff \int_{-\infty}^{\infty} f_{Y|X}(y|X=x)dy = 1$

Conditional Expectation

Let $f_{Y|X}(y|X=x)$ be the conditional density function of Y given $X=x$. Then,

$$E(Y|X=x)=\int_{-\infty}^{\infty} y f_{Y|X}(y|X=x)dy$$

is called the conditional expectation of Y given $X=x$.

- $E(Y|X=x)$ is the expectation of Y obtained when restricting the population to those who satisfy $X=x$.
- For example
  - $E(Y|X=9)=$ the expectation of anninc for junior high graduates.
  - $E(Y|X=16)=$ the expectation of anninc for college graduates.

  In this example, it is expected that $E(Y|X=9)\leq E(Y|X=16)$.

NOTE : The value of $E(Y|X=x)$ can vary with x. That is, $E(Y|X=x)$ is the value obtained by plugging x into the so-called conditional expectation function $E(Y|X=\cdot)$.

Joint Distribution and Joint Density

Joint Distribution Function, Joint Density Function

For random variables (Y, X) and constants (y, x), the joint probability of $\{Y \leq y, X \leq x\}$ is called the joint distribution function, and it is denoted by $F_{YX}(y, x)$:

$$F_{YX}(y, x) = \Pr(Y \leq y, X \leq x)$$

In addition, by taking the cross-partial derivative of $F_{YX}(y, x)$, we obtain the joint density function

$$f_{YX}(y, x) = \frac{\partial^2 F_{YX}(y, x)}{\partial y \partial x}$$

- You can interpret the joint density as the likelihood of the joint event. For example,
  - $f_{YX}(2, 9)$ = the likelihood that a randomly selected person is a junior high graduate and her anninc is 2 mill JPY.
  - $f_{YX}(5, 16)$ = the likelihood that a randomly selected person is a college graduate and her anninc is 5 mill JPY.

Joint Distribution and Joint Density

- It is important not to confuse conditional distribution with joint distribution.
- When we say "conditional ...", we are focusing on a sub-population satisfying the condition. When we say "joint ...", we are focusing on the whole population.

The joint density function has the following properties:

- $F_{YX}(y,x)$ is non-decreasing in both $(y,x) \iff f_{YX}(y,x) \geq 0$ for all (y,x)
- $F_{YX}(a,b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f_{YX}(y,x) dy dx$.
- The probability that Y and X take values in $[a_1, b_1]$ and $[a_2, b_2]$, respectively, is

$$\Pr(Y \in [a_1, b_1], X \in [a_2, b_2]) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f_{YX}(y,x) dy dx$$

Marginal Distribution and Marginal Density

- Given a joint distribution of random variables, the distribution and density function of each random variable are called the margianl distribution function and marginal density function, respectively.
- The marginal distribution function of Y, $F_Y(y)$, can be derived from the joint distribution function $F_{YX}(y, x)$ in the following manner:

$$\begin{aligned} F_Y(y) = \Pr(Y \leq y) &= \Pr(Y \leq y, X \leq \infty) \\ &= F_{YX}(y, \infty) \end{aligned}$$

- Namely,

$$\int_{-\infty}^{a} f_Y(y)dy = \int_{-\infty}^{a} \int_{-\infty}^{\infty} f_{YX}(y, x)dxdy$$

which further implies that

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

Conditional Density and Joint Density

- Recall: Relationship between the joint probability $\Pr(A, B)$ and the conditional probability $\Pr(A|B)$

$$\Pr(A|B) = \frac{\Pr(A \bigcap B)}{\Pr(B)}$$

- The same representation holds for the conditional density function:

$$f_{Y|X}(y|X = x) = \frac{f_{YX}(y, x)}{f_X(x)} = \frac{f_{YX}(y, x)}{\int f_{YX}(y, x)dy}$$

$\Rightarrow$ if the marginal density $f_X(x)$ is equal to zero, the conditional density $f_{Y|X}(y|X = x)$ is not well-defined.

Independence

- **Independence of events**: For two events $\{Y \leq y\}$ and $\{X \leq x\}$, if

$$\Pr(Y \leq y, X \leq x) = \Pr(Y \leq y) \cdot \Pr(X \leq x)$$

in other words, $F_{YX}(y, x) = F_Y(y) \cdot F_X(x)$

then, we say that these events are independent.

Independence of random variables

Let $F_{YX}(y, x)$ be the joint distribution function of (Y, X), and $F_Y(y)$ and $F_X(x)$ be the marginal distribution function of Y and X, respectively. We say that X and Y are independent if

$$F_{YX}(y, x) = F_Y(y) \cdot F_X(x)$$

for any (y, x).

Independence of random variables

- When two random variables Y and X are independent, we also have

$$f_{YX}(y,x) = f_Y(y) \cdot f_X(x)$$

because

$$\begin{aligned} f_{YX}(y,x) &= \frac{\partial^2 F_{YX}(y,x)}{\partial y \partial x} \\ &= \frac{\partial F_Y(y)}{\partial y} \frac{\partial F_X(x)}{\partial x} = f_Y(y) \cdot f_X(x) \end{aligned}$$

- Intuitively, the independence between Y and X means that they are "irrelevant" to each other.
- The concept of independence is "stronger" than uncorrelatedness:

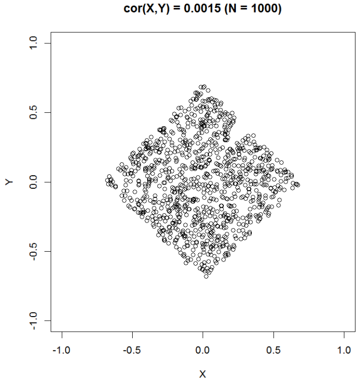
Y and X are independent $\Rightarrow$ they are uncorrelated, but the converse is not necessarily true.

Example

- Suppose that the joint density of Y and X is given by

$$f_{YX}(y,x) = \begin{cases} 1 & \text{if } |y| + |x| \le \frac{1}{\sqrt{2}} \\ 0 & \text{otherwise} \end{cases}$$

- Scatter plot of X and Y values ($N = 1,000$)



Example (cont')

- Suppose that the joint density of Y and X is given by

$$f_{YX}(y,x) = \begin{cases} 1 & \text{if } |y| + |x| \leq \frac{1}{\sqrt{2}} \\ 0 & \text{otherwise} \end{cases}$$

- As clearly seen from this figure, there is no correlation between X and Y. Indeed, the sample correlation coefficient is almost zero (0.0015).
- However, they are not independent (Exercise: Prove this fact.); for example, when $X = \frac{1}{\sqrt{2}}$, the only value that Y can take is zero, i.e., the value of Y depends on X, and vice versa.

Independence of random variables

- An important property of independent variables is that if Y and X are independent, the conditional expectation of Y given $X = x$ coincides with its unconditional expectation.
- Note that if Y and X are independent,

$$f_{Y|X}(y|X = x) = \frac{f_{YX}(y, x)}{f_X(x)} = \frac{f_Y(y) \cdot f_X(x)}{f_X(x)} = f_Y(y)$$

  Thus, for any x, it holds that

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|X = x) dy = \int_{-\infty}^{\infty} y f_Y(y) dy = E(Y)$$

- In the example of $Y =$ anninc and $X =$ education, the independence of them means that the level of education has no effect on the mean annual income.

Expectation of product of random variables

The expectation of the product of random variables Y and X is given by

$$E(YX) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yx f_{YX}(y,x) dx dy$$

- If Y and X are independent, since $f_{YX}(y,x) = f_Y(y) f_X(x)$, we have

$$\begin{aligned} E(YX) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yx f_Y(y) f_X(x) dx dy \\ &= \int_{-\infty}^{\infty} y f_Y(y) dy \int_{-\infty}^{\infty} x f_X(x) dx = E(Y)E(X) \end{aligned}$$

Covariance

Covariance

Covariance

- For two random variables Y and X, the covariance is defined by

$$\begin{aligned} Cov(Y,X) &= E[\{Y - E(Y)\}\{X - E(X)\}] \\ &= E(YX) - E(Y)E(X) \end{aligned}$$

- In particular, if either X or Y has mean zero, $Cov(Y,X) = E(YX)$.

NOTE : $Cov(Y,X)$ provides the strength of the correlation between Y and X. Hence, if they are independent, by $E(YX) = E(Y)E(X)$, we obtain

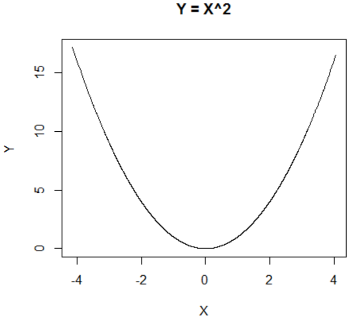
$$Cov(Y,X) = 0$$

However, $Cov(Y,X) = 0$ does not imply the independence.

Example

- Let X be a random variable distributed as the standard normal with $E(X) = 0$ and $V(X) = 1$. Also, define $Y = X^2$.
- Noting that for a standard normal X we have $E[X^3] = 0$,

$$\begin{aligned} Cov(Y, X) &= E(YX) - E(Y)E(X) \\ &= E(X^3) - E(X)E(X^2) = 0 - 0 \cdot 1 = 0 \end{aligned}$$



Variance Covariance Matrix

Consider a vector of random variables $\mathbf{X} = (X_1, ..., X_k)^\top$, where $\top$ denotes transpose. Then, the following $k \times k$ matrix is caleld the variance covariance matrix, or simply covariance matrix, of $\mathbf{X}$:

$$
\begin{aligned}
VCM(\mathbf{X}) &= E[\{\mathbf{X} - E(\mathbf{X})\}\{\mathbf{X} - E(\mathbf{X})\}^\top] \\
&= \begin{bmatrix} V(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & V(X_2) & \cdots & Cov(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & V(X_k) \end{bmatrix}
\end{aligned}
$$

- $VCM(\mathbf{X})$ is a symmetric matrix.
- When X_i and X_j are independent for all $i \neq j$, all the off-diagonals of $VCM(\mathbf{X})$ are zero, i.e., $VCM(\mathbf{X})$ becomes a diagonal matrix.

Exercise 3

1. Suppose that X and Y are independent. Prove the following result:

$$V(X+Y) = V(X) + V(Y)$$

Glossary II