

# Random Variables

## 1.1 Density function

By definition, a random variable  $X$  is a function with domain the sample space and range a subset of the real numbers. For example, in rolling two dice  $X$  might represent the sum of the points on the two dice. Similarly, in taking samples of college students  $X$  might represent the number of hours per week a student studies, a student's GPA, or a student's height. The notation  $X(s) = x$  means that  $x$  is the value associated with the outcome  $s$  by the random variable  $X$ .

There are three types of random variables: discrete random variables, continuous random variables, and mixed random variables.

**Example 1.1.1.** A committee of 4 is selected from a group consisting of 5 men and 5 women. Let  $X$  be the random variable that represents the number of women in the committee. Find the probability mass distribution of  $X$ .

**Solution** For  $x = 0, 1, 2, 3, 4$  we have

$$p_X(x) = \frac{\binom{5}{x} \binom{5}{4-x}}{\binom{10}{4}} \quad x = 0, 1, 2, 3, 4.$$

The probability mass function can be described by the table

$x$	0	1	2	3	4
$p(x)$	$\frac{5}{210}$	$\frac{50}{210}$	$\frac{100}{210}$	$\frac{50}{210}$	$\frac{5}{210}$

□

## 1.2 Cumulative Distribution

First, we prove that the probability is a continuous set function. In order to do that, we need the following definitions:

### Definition 1.1 Increasing and Decreasing sequence of events

A sequence of sets  $\{E_n\}_{n=1}^{\infty}$  is said to be increasing if

$$E_1 \subset E_2 \subset \dots \subset E_n \subset E_{n+1} \subset \dots$$

whereas it is said to be a decreasing sequence if

$$E_1 \supset E_2 \supset \dots \supset E_n \supset E_{n+1} \supset \dots$$

**Lemma 1.1**

If  $\{E_n\}_{n \geq 1}$  is either an increasing or decreasing sequence of events then

$$\lim_{n \rightarrow \infty} \mathbb{P}[E_n] = \mathbb{P}[\lim_{n \rightarrow \infty} E_n]. \quad (1.1)$$

that is

$$\mathbb{P} \left[ \bigcup_{n=1}^{\infty} E_n \right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] \quad \text{for increasing sequence,} \quad (1.2)$$

and

$$\mathbb{P} \left[ \bigcap_{n=1}^{\infty} E_n \right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] \quad \text{for decreasing sequence,} \quad (1.3)$$

*Proof.* Firstly, suppose that  $E_n \subset E_{n+1}$  for all  $n \geq 1$ . Define the events

$$\begin{aligned} F_1 &= E_1 \\ F_n &= E_n \cap E_{n-1}^c, \quad n > 1 \end{aligned}$$

Note that for  $n > 1$ ,  $F_n$  consists of those outcomes in  $E_n$  that are not in any of the earlier  $E_n$   $\forall i < n$ . Clearly, for  $i \neq j$  we have  $F_i \cap F_j = \emptyset$ . Also,  $\bigcup_{n=1}^{\infty} F_n = \bigcup_{n=1}^{\infty} E_n$  and for  $n \geq 1$  we have  $\bigcup_{i=1}^n F_i = \bigcup_{i=1}^n E_i$ . From these properties we have

$$\begin{aligned} \mathbb{P} \left[ \lim_{n \rightarrow \infty} E_n \right] &= \mathbb{P} \left[ \bigcup_{n=1}^{\infty} E_n \right] = \mathbb{P} \left[ \bigcup_{n=1}^{\infty} F_n \right] \\ &= \sum_{n=1}^{\infty} \mathbb{P}[F_n] \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{P}[F_i] \\ &= \lim_{n \rightarrow \infty} \mathbb{P} \left[ \bigcup_{i=1}^n F_i \right] \\ &= \lim_{n \rightarrow \infty} \mathbb{P} \left[ \bigcup_{i=1}^n E_i \right] \\ &= \lim_{n \rightarrow \infty} \mathbb{P}[E_n]. \end{aligned}$$

On the other hand, now suppose that the sequence  $\{E_n\}_{n \geq 1}$  is a decreasing sequence of events. Then  $\{E_n^c\}_{n \geq 1}$  is an increasing sequence of events. Hence, from the previous part we have

$$\mathbb{P} \left[ \bigcup_{n=1}^{\infty} E_n^c \right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n^c].$$

By De Morgan's law we have  $\bigcup_{n=1}^{\infty} E_n^c = (\bigcap_{n=1}^{\infty} E_n)^c$ . And

$$\mathbb{P} \left[ \left( \bigcap_{n=1}^{\infty} E_n \right)^c \right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n^c].$$

Equivalently,

$$1 - \mathbb{P} \left[ \bigcap_{n=1}^{\infty} E_n \right] = \lim_{n \rightarrow \infty} (1 - \mathbb{P}[E_n]) = 1 - \lim_{n \rightarrow \infty} \mathbb{P}[E_n]$$

or

$$\mathbb{P} \left[ \bigcap_{n=1}^{\infty} E_n \right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n].$$

□

### Theorem 1.1 Properties of Cumulative Distribution Function

If  $F_X(x)$  is a cumulative distribution function, then

1.  $F_X(-\infty) = \lim_{x \downarrow -\infty} F_X(x) = 0$ .
2.  $F_X(+\infty) = \lim_{x \rightarrow +\infty} F_X(x) = 1$ .
3.  $F_X(x)$  is always *monotonically increasing*. That said, if  $x_1 < x_2$ , then  $F_X(x_1) < F_X(x_2)$ .

*Proof.* 1. Note that  $\lim_{x \downarrow -\infty} F(x) = \lim_{n \rightarrow \infty} F(x_n)$  where  $\{x_n\}$  is a decreasing sequence such that  $x_n \downarrow -\infty$ . Define

$$E_n = \{s \in \Omega : X(s) \leq x_n\}.$$

Then we have the nested chain  $E_1 \supseteq E_2 \supseteq \dots$ . Moreover,

$$\emptyset = \bigcap_{n=1}^{\infty} E_n.$$

By previous proposition, we find

$$\lim_{x \rightarrow -\infty} F(x) = \lim_{n \downarrow -\infty} F(x_n) = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] = \mathbb{P} \left[ \bigcap_{n=1}^{\infty} E_n \right] = \mathbb{P}[\emptyset] = 0.$$

2. In the other hand, suppose that  $\lim_{x \rightarrow \infty} F(x) = \lim_{n \rightarrow \infty} F(x_n)$  where  $\{x_n\}$  is a increasing sequence such that  $x_n \rightarrow \infty$ . We reuse back the definition of  $E_n$  that is

$$E_n = \{s \in \Omega : X(s) \leq x_n\}.$$

Then we have the nested chain in the opposite direction  $E_1 \subseteq E_2 \subseteq \dots$ . Moreover,

$$\Omega = \bigcup_{n=1}^{\infty} E_n$$

By previous proposition, we find


$$\lim_{x \rightarrow \infty} F(x) = \lim_{n \rightarrow \infty} F(x_n) = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] = \mathbb{P} \left[ \bigcup_{n=1}^{\infty} E_n \right] = \mathbb{P}[\Omega] = 1.$$

3. Consider two real numbers  $a, b$  such that  $a < b$ . Then

$$\{s \in \Omega : X(s) \leq a\} \subset \{s \in \Omega : X(s) \leq b\}.$$


This implies that  $\mathbb{P}[X \leq a] < \mathbb{P}[X \leq b]$ . Hence,  $F(a) < F(b)$ .

□

 **Example 1.2.1.** Let  $X$  be a random variable with probability density function

$$f_X(x) = \begin{cases} 2 - 4|x| & \text{if } \frac{1}{2} < x < -\frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

1. Find the variance of  $X$ .
2. Find the cumulative function  $F(x)$  of  $X$ .

 **Solution** 1. Since the density function  $f(x)$  is odd in  $\left(-\frac{1}{2}, \frac{1}{2}\right)$ , we have  $\mathbb{E}[X] = 0$ . Therefore

$$\begin{aligned} \text{Var}[X] &= \mathbb{E}[X^2] - 0 = \int_{-1/2}^0 x^2(2 + 4x) \, dx + \int_0^{1/2} x^2(2 - 4x) \, dx \\ &= \frac{1}{24}. \end{aligned}$$

2. The cumulative function is


$$\begin{aligned} F(x) &= \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ \int_{-1/2}^x (2 + 4t) \, dt & \text{if } -\frac{1}{2} \leq x \leq 0 \\ \int_{-1/2}^0 (2 + 4t) \, dt + \int_0^x (2 - 4t) \, dt & \text{if } 0 \leq x \leq \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases} \\ &= \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ 2x^2 + 2x + \frac{1}{2} & \text{if } -\frac{1}{2} \leq x \leq 0 \\ -2x^2 + 2x + \frac{1}{2} & \text{if } 0 \leq x \leq \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases} \end{aligned}$$

□


## 1.3 Percentiles and Quantiles

### 1.3.1 Mode

In the discrete case, the mode is the value that is most likely to be sampled. In the continuous case, the mode is where  $f(x)$  is at its peak.

 **Example 1.3.1.** The lifetime of a light bulb has density function,  $f_X$ , where  $f_X(x)$  is proportional to  $\frac{x^2}{1+x^3}$ ,  $0 < x < 5$ , and 0 otherwise.

Calculate the mode of this distribution.

 **Solution** Given the lifetime of a light bulb  $X$  has density function

$$f_X(x) = \frac{cx^2}{1+x^3}.$$

Compute the first and second order derivative of  $f$ .

$$\frac{df}{dx} = \frac{(1+x^3) \frac{d}{dx}(cx^2) - cx^2 \frac{d}{dx}(1+x^3)}{(1+x^3)^2} = \frac{2cx - cx^4}{(1+x^3)^2}.$$

$$\begin{aligned} \frac{d^2f}{dx^2} &= \frac{(1+x^3)^2 \frac{d}{dx}(2cx - cx^4) - (2cx - cx^4) \frac{d}{dx}(1+x^3)^2}{(1+x^3)^4} \\ &= \frac{(1+x^3)^2(2c - 4cx^3) - 6x^2(2cx - cx^4)(1+x^3)}{(1+x^3)^4} \\ &= \frac{(1+x^3)(2c - 4cx^3) - 6x^2(2cx - cx^4)}{(1+x^3)^3} \\ &= \end{aligned}$$

□

By inspection,  $\frac{d^2f}{dx^2} < 0$ . And so  $\frac{df}{dx} = 0$  is maximum point in  $(0, 5)$ . Solve for  $x$  of the following equation:

$$\frac{2cx - cx^4}{(1+x^3)^2} = 0$$

Since  $(1+x^3)^2 > 0$ , we can remove it safely from the equation.

## 1.4 Expected Value and Moments

For a random variable  $X$ , the expected value is denoted  $\mathbb{E}[X]$ , or  $\mu_X$  or simply  $\mu$ . The expected value is called the expectation of  $X$ , which is the "average" over the range of values that distribution  $X$  can be. You may said the expectation is the "center" of the distribution.

### Definition 1.2 Expectation value

Let  $(\Omega, \mathbb{P})$  be a probability space, let  $E \subseteq \mathbb{R}$  be countable, and let  $X$  be a  $E$ -valued random variable on  $(\Omega, \mathbb{P})$ . The expectation of  $X$ , if it exists, is defined by

$$\mathbb{E}[X] := \sum_{e \in E} e f_X(e). \quad (1.4)$$

### Lemma 1.2

Let  $(\Omega, \mathbb{P})$  be a probability space, let  $E \subseteq \mathbb{R}$  be countable set and let  $X$  be an  $E$ -valued random variable on  $(\Omega, \mathbb{P})$ . Then

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}[\{\omega\}].$$

*Proof.* Recall that

$$\Omega = \bigcup_{e \in E} \{X = e\}$$

and the events  $\{X = e\}$  are mutually exclusive. Hence

$$\begin{aligned}\sum_{\omega \in \Omega} X(\omega) \mathbb{P}[\{\omega\}] &= \sum_{e \in E} \sum_{\omega \in \{X=e\}} X(\omega) \mathbb{P}[\{\omega\}] \\ &= \sum_{e \in E} e \mathbb{P}\{X = e\} \\ &= \sum_{e \in E} e f_X(e)\end{aligned}$$

as what we expected. □

**Example 1.4.1.** Let  $X$  be a random variable representing the value shown a fair six-sided die is rolled. Then  $X \sim \text{discreteU}(\{1, 2, 3, 4, 5, 6\})$ , and  $f_X(k) = \frac{1}{6}$  for each number.

$$\mathbb{E}[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

So the expected value of a die rolled is 3.5.

**Example 1.4.2.** If  $f(x) = (k+1)x^2$  for  $0 < x < 1$ , find the moment generating function.

**Solution** Since  $f$  is a density function, thus  $\int_0^1 f(x) dx = 1$ , it follows that

$$(k+1) \times \frac{1}{3} = 1$$

so that  $k = 2$  and now  $f(x) = 3x^2$  for  $0 < x < 1$ . Then the moment generating function is

$$\begin{aligned}M_X(t) &= \int_0^1 e^{tx} (3x^2) dx = \int_0^1 3x^2 d\left(\frac{e^{tx}}{t}\right) \\ &= \frac{3x^2 e^{tx}}{t} \Big|_{x=0}^{x=1} - \int_0^1 \frac{6x e^{tx}}{t} dx \\ &= \frac{3e^t}{t} - \left[ \frac{6x e^{tx}}{t^2} \Big|_{x=0}^{x=1} - \int_0^1 \frac{6x e^{tx}}{t^2} dx \right] \\ &= \frac{3e^t}{t} - \frac{6e^t}{t^2} + \frac{6(e^t - 1)}{t^3} \\ &= \frac{e^t(6 - 6t + 3t^2)}{t^3} - \frac{6}{t^3}.\end{aligned}$$

□

### 1.4.1 Variance

Variance is a measure of the "dispersion" of  $X$  about the mean.

#### Definition 1.3 Variance

The variance of distribution  $X$  is sum of squared loss

$$\text{Var}[X] := \mathbb{E}_X(X_i - \mu_X)^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad (1.5)$$

A large variance indicates significant levels of probability or density from points far away from the mean. The variance must be always  $\geq 0$  (Since everything is squared). The variance of  $X$  is equal to zero only if  $X$  is a fixed single point and with probability 1 at that point; In other words, the function of  $X$  is a constant function (For example,  $x \sim f_X(x) = 6$ , then  $Var[X] = 0$ ).

The standard deviation of the random variable  $X$  is the square root

**Theorem 1.2 Chebyshev's Theorem**

Let  $X$  be a random variable with mean  $\mu_X$  and finite variance  $\sigma^2$ . Then,

$$\mathbb{P}[|X - \mu_X| < k\sigma] \geq 1 - \frac{1}{k^2} \quad (1.6)$$

or

$$\mathbb{P}[|X - \mu_X| \geq k\sigma] \leq \frac{1}{k^2} \quad (1.7)$$

for some constant  $k > 0$ .

## 1.4.2 The Coefficient of Variation

**Definition 1.4 Coefficient of Variation**

The coefficient of variation is

$$CV = \frac{\sigma_X}{\mu_X} = \frac{\sqrt{Var[x]}}{\mathbb{E}[X]}. \quad (1.8)$$

A higher CV implies greater variability, while a lower CV suggests more consistency or reliability of the data. Imagine if we have two datasets:

- ❖ Dataset  $A$  has a mean of 10 and standard deviation of 2, and  $CV_A = 2/10 = 1/5$ .
- ❖ Dataset  $B$  has a mean of 100 and standard deviation of 10, and  $CV_B = 10/100 = 1/10$ .

While dataset  $B$  has higher standard deviation, but it has a lower CV compared to dataset  $A$ . This indicating  $B$  less reliable variation to the mean.

## 1.5 Discrete Random Variables

### 1.5.1 Binomial distribution

A Bernoulli trial is an experiment with only two outcomes: **Success** and **failure**. The probability of a success is denoted by  $p$  and that of a failure by  $q = 1 - p$ . Moreover,  $p$  and  $q$  are related by the formula

$$p + q = 1.$$

A Bernoulli experiment is a sequence of independent Bernoulli trials. Let  $X$  represent the number of successes that occur in  $n$  independent Bernoulli trials. Then  $X$  is said to be a Binomial random variable  $(n, p)$ . If  $n = 1$ , then  $X$  is said to be a Bernoulli random variable.

**Theorem 1.3**

Let  $(\Omega, \mathbb{P})$  be a probability space, let  $p \in [0, 1]$  and let  $X_1, X_2, \dots, X_n : \Omega \rightarrow \{0, 1\}$  be inde-

pendent random variables such that each  $X_i \sim \text{Bernoulli}(p)$ . Then

$$X_1 + X_2 + \cdots + X_n \sim \text{Bin}(n, p).$$

## 1.5.2 Geometric distribution

A geometric random variable with parameter  $p$ ,  $0 < p < 1$  has a probability mass function

$$p_X(n) = \mathbb{P}(X = n) = p(1 - p)^{n-1}, \quad n = 1, 2, \dots \quad (1.9)$$


Note that  $p_X(n) \geq 0$  and

$$\sum_{n=1}^{\infty} p(1 - p)^{n-1} = \frac{p}{1 - (1 - p)} = 1. \quad (1.10)$$

A geometric random variable models the number of successive independent Bernoulli trials that must be performed to obtain the  $r$ -st success. For example, the number of flips of a fair coin until the  $r$ -st head appears follows a geometric distribution.

 **Example 1.5.1.** Consider the experiment of rolling a pair of fair dice.

1. What is the probability of getting a sum of 11?
2. If you roll two dice repeatedly, what is the probability that the first sum of 11 occurs on the 8-th roll?

 **Solution** 1. A sum of 11 occurs when the pair of dice show either (5, 6) or (6, 5) so that the required probability is  $\frac{2}{36} = \frac{1}{18}$ .

2. Let  $X$  be the number of rolls on which the first sum of 11 happened. Then  $X$  is a geometric random variable with probability  $p = \frac{1}{18}$ . Thus

$$\mathbb{P}[X = 8] = \frac{1}{18} \left(1 - \frac{1}{18}\right)^7 = 0.0372.$$

□

## 1.6 Continuous Probability Distributions

### 1.6.1 Uniform Probability Distribution

#### Definition 1.5 Uniform Distribution

If  $a < b$ , a random variable  $X$  is said to have a continuous uniform distribution if

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases} \quad (1.11)$$



**Theorem 1.4 Mean and variance of Uniform Distribution**

If  $X$  is a continuous uniform distribution on the interval  $[a, b]$ , then the mean is

$$\mu_X = \mathbb{E}[X] = \frac{a + b}{2} \quad (1.12)$$

and

$$\sigma_X^2 = \text{Var}[X] = \frac{(b - a)^2}{12} \quad (1.13)$$

*Proof.* Given  $X$  is a continuous uniform distribution on the interval  $[a, b]$ , with  $a < b$ . Then the expectation of  $X$  is

$$\begin{aligned} \mu_X = \mathbb{E}[X] &= \int_a^b x \frac{1}{b - a} dx = \frac{x^2}{2(b - a)} \Big|_{x=a}^{x=b} \\ &= \frac{b^2 - a^2}{2(b - a)} \\ &= \frac{(b - a)(b + a)}{2(b - a)} \\ &= \frac{a + b}{2}. \end{aligned}$$

Now we continue to work on the variance for  $X$ . But before that, we need to find the expectation of  $X^2$ .

$$\begin{aligned} \mathbb{E}[X^2] &= \int_a^b x^2 \frac{1}{b - a} dx = \frac{x^3}{3(b - a)} \Big|_{x=a}^{x=b} \\ &= \frac{b^3 - a^3}{3(b - a)} \\ &= \frac{(b - a)(b^2 + ab + a^2)}{3(b - a)} \\ &= \frac{a^2 + ab + b^2}{3} \end{aligned}$$

The variance of  $X$  is

$$\begin{aligned}
 \sigma_X^2 &= \text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\
 &= \frac{4(b^2 + ab + a^2) - 3(a+b)^2}{12} \\
 &= \frac{a^2 - 2ab + b^2}{12} \\
 &= \frac{(b-a)^2}{12}
 \end{aligned}$$

which is what we expected. □

## 1.7 Normal Distribution

## 1.8 Gamma Distribution

Some random variables can yield distributions of data are skewed right and is non-symmetric.

### Definition 1.6 Gamma Distribution

Let  $X$  be a random variable followed *gamma distribution* with parameters  $\alpha > 0$  and  $\beta$ . The density function of  $X$  is

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases} \quad (1.14)$$

### Theorem 1.5 Mean and variance of Gamma Distribution

If  $X$  is a gamma distribution with parameters  $\alpha > 0$  and  $\beta > 0$ , then the mean and variance are

$$\mu_X = \alpha\beta \quad (1.15)$$

$$\sigma^2 = \alpha\beta^2. \quad (1.16)$$

*Proof.* Using the moment generating function approach to find mean and variance,

$$\begin{aligned}
M_X(t) &= \int_0^{\infty} e^{tx} \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^{\alpha} \Gamma(\alpha)} dx \\
&= \frac{1}{\beta^{\alpha}} \int_0^{\infty} \frac{x^{\alpha-1} e^{-x\left(\frac{1}{\beta}-t\right)}}{\Gamma(\alpha)} dx \\
&= \frac{1}{\beta^{\alpha}} \int_0^{\infty} \frac{x^{\alpha-1} e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha)} dx \\
&= \frac{\left(\frac{\beta}{1-\beta t}\right)^{\alpha}}{\beta^{\alpha}} \int_0^{\infty} \frac{x^{\alpha-1} e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha) \left(\frac{\beta}{1-\beta t}\right)^{\alpha}} dx
\end{aligned}$$

Now observe that

$$\int_0^{\infty} \frac{x^{\alpha-1} e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha) \left(\frac{\beta}{1-\beta t}\right)^{\alpha}} dx = 1$$

as integrating  $x$  along the entire curve will give us 1. Thus,

$$M_X(t) = \frac{\left(\frac{\beta}{1-\beta t}\right)^{\alpha}}{\beta^{\alpha}} \times 1 = \frac{1}{(1-\beta t)^{\alpha}}. \quad (1.17)$$

That is to say, recall that the  $k$ -th derivative of  $M_X(t)$  with respect to  $t$  as at  $t = 0$ , is the  $\mathbb{E}[X^k]$ . From here we compute

$$\mathbb{E}[X] = \dot{M}_X(0) = \alpha\beta(1-\beta t)^{-\alpha-1}\big|_{t=0} = \alpha\beta.$$

$$\mathbb{E}[X^2] = \ddot{M}_X(0) = -\alpha(1+\alpha)\beta^2(1-\beta t)^{-\alpha-2}\big|_{t=0} = \alpha(1+\alpha)\beta^2.$$

And so the variance is

$$\begin{aligned}
Var[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 \\
&= \alpha\beta^2.
\end{aligned}$$

and we are done. □

### 1.8.1 Chi-square distribution

**Definition 1.7**

If  $X$  is a gamma distribution with parameters  $\alpha = \nu/2$  and  $\beta = 2$ , then  $X$  is a  $\chi^2$ -distribution with  $\nu$  degree of freedom. (with  $\nu > 0$ )

# Limiting Distribution

## 2.1 Coverage in distribution


Convergence in distribution state that the sequence of random variables  $X_1, X_2, \dots, X_n$  converges to some distribution  $X$  when  $n$  goes to infinity. It does not require any dependence between the  $X_n$  and  $X$ . Convergence in distribution is consider as the weakest type of convergence.

### Definition 2.1 Coverage in distribution

Let  $\{X_n\}_{n \geq 1}$  be a sequence of random variable and let  $X$  be a random variable. Let  $F_{X_n}$  and  $F_X$  be the cdf of  $X_n$  and  $X$  respectively. And let  $C(F)$  be the set of all continuous points of  $F$ . We say that  $X_n$  converges in distribution to  $X$  if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \quad (2.1)$$

for all  $x \in C(F)$ . We denote  $X_n \xrightarrow{d} X$ .

 **Example 2.1.1.** Let  $X_1, X_2, X_3, \dots$  be a sequence of random variable such that

$$X_n \sim \text{Geom}(\lambda/n), \quad \forall n = 1, 2, 3, \dots$$

where  $\lambda > 0$  is a constant. Define a new sequence  $Y_n$  as

$$Y_n = \frac{1}{n}X_n, \quad \forall n = 1, 2, 3, \dots$$

Show that  $Y_n$  converges in distribution to  $\text{Exp}(\lambda)$ .

 **Solution** The cdf of  $Y_n$  is

$$\begin{aligned} F_{Y_n}(y) &= \mathbb{P}\left[\frac{1}{n}X_n \leq y\right] \\ &= \mathbb{P}[X_n \leq ny] \\ &= 1 - \left(1 - \frac{\lambda}{n}\right)^{\lfloor ny \rfloor}. \end{aligned}$$

and taking limit for  $n \rightarrow +\infty$  we have

$$\begin{aligned}\lim_{n \rightarrow \infty} \left( 1 - \left( 1 - \frac{\lambda}{n} \right)^{\lfloor ny \rfloor} \right) &= 1 - \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{\left( -\frac{n}{\lambda} \right)} \right)^{-n(-y)} \\ &= 1 - \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{\left( -\frac{n}{\lambda} \right)} \right)^{-\frac{n}{\lambda}(-\lambda y)} \\ &= 1 - e^{-\lambda y}\end{aligned}$$

which is the cdf of exponential distribution with parameter  $\lambda$ .  $\square$

**Example 2.1.2.** Let  $\{X_n\}_{n \geq 1}$  be a sequence of random variables with probability mass function

$$f_{X_n}(x) = \mathbb{P}[X_n = x] = \begin{cases} 1 & \text{if } x = 2 + \frac{1}{n} \\ 0 & \text{otherwise.} \end{cases}$$

Find the limiting distribution of  $X_n$ .

### 2.1.1 Almost sure converges

We said that  $X_n$  converges to  $X$  **almost surely** if the probability that the sequence  $X_n(s)$  converges to  $X(s)$  is equal to 1.

**Example 2.1.3.** Consider the sample space  $S = [0, 1]$  with uniform probability distribution, for instance,

$$\mathbb{P}([a, b]) = b - a \quad \forall 0 \leq a \leq b \leq 1.$$

Define the sequence  $\{X_n, n = 1, 2, 3, \dots\}$  as

$$X_n(s) = \frac{n}{n+1}s + (1-s)^n.$$

Also, define the random variable  $X$  on the sample space as  $X(s) = s$ . Show that  $X_n$  *almost sure* converges to  $X$ .

**Solution** For any  $s \in [0, 1]$ , taking the limit when  $n \gg \infty$  we have

$$\begin{aligned}\lim_{n \rightarrow \infty} X_n(s) &= \lim_{n \rightarrow \infty} \left[ \frac{n}{n+1}s + (1-s)^n \right] \\ &= \lim_{n \rightarrow \infty} \frac{n}{n+1}s + \lim_{n \rightarrow \infty} (1-s)^n \\ &= 1 \cdot s + 0 \\ &= s = X(s).\end{aligned}$$

However, if  $s = 0$  then

$$\lim_{n \rightarrow \infty} X_n(0) = \lim_{n \rightarrow \infty} \left[ \frac{n}{n+1}(0) + (1-0)^n \right] = 1.$$

Thus, we conclude that  $\lim_{n \rightarrow \infty} X_n(s) = X(s)$  for all  $s$  lies between 0 and 1. And because  $\mathbb{P}([0, 1]) = 1$ , we conclude that

$$X_n \xrightarrow{a.s.} X$$

and we are done. □

## 2.2 Law of Large Numbers

In this section we will discuss the Weak and Strong Law of Large Numbers. The Law of Large Numbers are considered as a form of convergence in probability.

We will first state the Weak Law of Large Numbers (WLLN),

### Theorem 2.1 Weak Law of Large Numbers (WLLN)

Let  $\{X_n\}_{n \geq 1}$  be a sequence of i.i.d random variables, each with mean  $\mathbb{E}[X_i] = \mu$  and standard deviation  $\sigma$ , we define

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

The Weak Law of Large Numbers (WLLN) states that for all  $\epsilon > 0$ , we have

$$\lim_{n \rightarrow \infty} \mathbb{P}[|\bar{X}_n - \mu| > \epsilon] = 0. \quad (2.2)$$

*Proof.* Suppose that  $\text{Var}[X_i] = \sigma^2 > 0$  for finite  $i$ . Since  $X_1, X_2, \dots, X_n$  are identically independent, there is no correlation between them, thus

$$\begin{aligned} \text{Var}[\bar{X}_n] &= \text{Var}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n^2} \text{Var}[X_1 + X_2 + \cdots + X_n] \\ &= \frac{1}{n^2} [\text{Var}X_1 + \text{Var}X_2 + \cdots + \text{Var}X_n] \\ &= \frac{1}{n^2} (\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ times}}) \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Notice that the mean of each  $X_i$  in the sequence is also equal to the mean of the sample average, said  $\mathbb{E}[X_i] = \mu$ . We can now apply Chebyshev's inequality on  $\bar{X}_n$  to get, for all  $\epsilon > 0$ ,

$$\mathbb{P}[|\bar{X}_n - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

So that □

### Theorem 2.2 Strong Law of Large Numbers

Let  $\{X_n\}_{n \geq 1}$  be a sequence of i.i.d random variables, each with mean  $\mathbb{E}[X_i] = \mu$  and standard deviation  $\sigma$ , then

$$\mathbb{P}[\lim_{n \rightarrow \infty} \bar{X}_n = \mu] = 1. \quad (2.3)$$

*Proof.* By Markov's inequality that

$$\mathbb{P}\left[\frac{1}{n}|\bar{X}_n - n\mu| \geq n^{-\gamma}\right] \leq \frac{\mathbb{E}[(\frac{\bar{X}_n}{n} - \mu)^4]}{n^{-4\gamma}} = Kn^{-2+4\gamma}.$$

Define for all  $\gamma \in (0, \frac{1}{4})$ , and let

$$A_n = \left\{ \frac{1}{n} |\bar{X}_n - n\mu| \geq n^{-\gamma} \right\} \Rightarrow \sum_{n \geq 1} \mathbb{P}[A_n] < \infty \Rightarrow \mathbb{P}[A] = 0$$

by the first Borel-Cantelli lemma, where  $A = \bigcap_{n \geq 1} \bigcup_{m \geq n} A_m$ . But now, the event  $A^c$  happened if and only if

$$\exists N \forall n \geq N \left| \frac{\bar{X}_n}{n} - \mu \right| < n^{-\gamma} \Rightarrow \frac{\bar{X}_n}{n} \xrightarrow{p} \mu.$$

□


### Theorem 2.3 Central Limit Theorem


Let  $\{X_n\}_{n \geq 1}$  be a sequence of i.i.d random variable whose memoment generating function exist in a neighborhood of 0. Let  $\mathbb{E}[X_i] = \mu$  and  $\text{Var}[X_i] = \sigma^2 > 0$ . Define  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Then

$$Z = \frac{\sqrt{n}(\bar{X}_i - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad (2.4)$$

or

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad (2.5)$$

 **Example 2.2.1.** Let  $\bar{X}_n$  be the sample mean from a random sampling of size  $n = 100$  from  $\chi_{50}^2$ . Compute approximate value of  $\mathbb{P}(49 < \bar{X} < 51)$ .

 **Solution** Because  $\bar{X}$  followed Chi-squared distribution with degree of freedom 50, then the mean and variance are  $\mathbb{E}[X_i] = 50$  and  $\text{Var}[X_i] = 2(50) = 100$ . By Central Limit Theorem,


$$\begin{aligned} \mathbb{P}(49 < \bar{X} < 51) &\simeq \mathbb{P} \left[ \frac{\sqrt{100}(49 - \mathbb{E}[X_i])}{\sqrt{\text{Var}[X_i]}} < Z < \frac{\sqrt{100}(51 - \mathbb{E}[X_i])}{\sqrt{\text{Var}[X_i]}} \right] \\ &\simeq \mathbb{P} \left[ \frac{\sqrt{100}(49 - 50)}{\sqrt{100}} < Z < \frac{\sqrt{100}(51 - 50)}{\sqrt{100}} \right] \\ &\simeq \mathbb{P}[-1 < Z < 1] \\ &\simeq \Phi(1) - \Phi(-1) \\ &\simeq 0.84134 - 0.15866 = 0.68268. \end{aligned}$$

□

### Theorem 2.4 Slutsky's Theorem

If  $X_n$  converges in distribution to a random variable  $X$ , and  $Y_n$  converges in probability to a constant  $c$ , then

- $Y_n X_n \xrightarrow{d} cX$
- $X_n + Y_n \xrightarrow{d} X + c.$

 **Example 2.2.2.** If the random variable  $X \sim \text{Gamma}(\mu, 1)$ , show that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\bar{X}_n}} \xrightarrow{d} \mathcal{N}(0, 1)$$



⇒ **Solution** Slutsky's theorem stated that If  $X_n$  converges in distribution to a random variable  $X$  and if  $Y_n$  converges in probability to a constant  $c$ . Then  $X_n/Y_n$  converges in distribution to  $X/c$ . By the central limit theorem we have

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \text{Var}[X_i]).$$

and in this case  $\mathbb{E}X_i = \text{Var}[X_i] = \mu$ , thus we obtained

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \mu).$$

Replacing the theorem denominator  $Y_n$  with  $\bar{X}_n$ , which  $\bar{X}_n$  converges to constant  $\mu$  in probability. Hence

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\bar{X}_n}} \xrightarrow{d} \mathcal{N}\left(\frac{0}{\mu}, \frac{\mu}{\mu}\right) \Rightarrow \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\bar{X}_n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

and we are done with the proof. □

#### Theorem 2.5

If the random variable  $X_n$  converges to constant  $c$  in probability, then

$$\sqrt{X_n} \xrightarrow{p} \sqrt{c}, \quad c > 0. \quad (2.6)$$

#### Theorem 2.6

If the random variable  $X_n$  converges to constant  $c$  in probability, and  $Y_n$  converges to constant  $d$  in probability, then

- $aX_n + bY_n \xrightarrow{p} ac + bd$ .
- $X_n Y_n \xrightarrow{p} cd$ .
- $\frac{1}{X_n} \xrightarrow{p} \frac{1}{c}$  for all  $c \neq 0$ .

## 2.3 Order Statistics

We can ordering the observed random variables based on their magnitudes or ranking. These ordered variables are known as **order statistics**.

Consider  $X_1, X_2, \dots, X_n$  are independent continuous random variables with cdf  $F_X(y)$  and mass function  $f_X(y)$ . We ordered them into order statistics  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  such that

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

In this notion, the maximum random variable is

$$X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$$

and the minimum random variable is

$$X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$$

$X_{(n)}$  is the largest among  $X_1, X_2, \dots, X_n$ , the event  $X_{(n)} \leq y$  will happened only if each  $X_i \leq y$ .

Then the joint probability is

$$G_{(n)}(y) = \mathbb{P}[X_{(n)} \leq y] = \mathbb{P}[X_1 \leq y, X_2 \leq y, \dots, X_n \leq y] = \prod_{i=1}^n \mathbb{P}[X_i \leq y]. \quad (\heartsuit)$$

Because  $\mathbb{P}[X_i \leq y] = F_X(y)$  for all  $i = 1, 2, \dots, n$ . It follows that

$$(\heartsuit) \Rightarrow \mathbb{P}[X_1 \leq y] \mathbb{P}[X_2 \leq y] \cdots \mathbb{P}[X_n \leq y] = [F_X(y)]^n.$$

Now letting  $g_{(n)}$  denote the density function of  $Y_{(n)}$ , we see that, on taking derivative on  $G_{(n)}$  with respect to  $y$ .

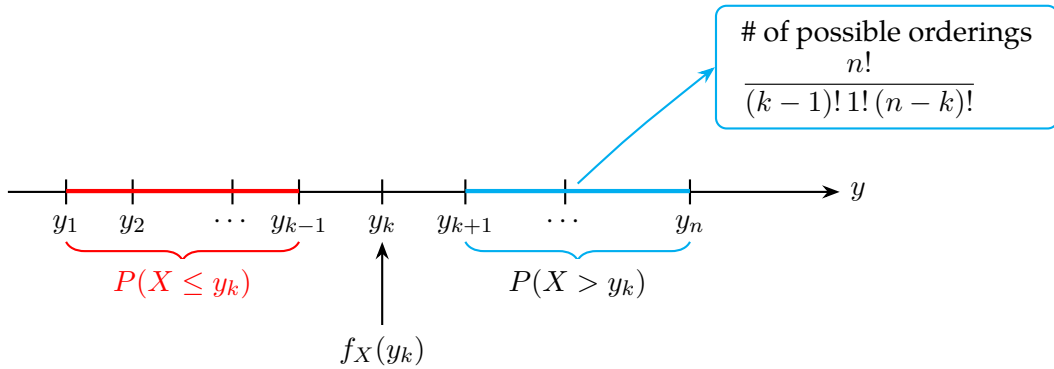
$$\begin{aligned} g_{(n)}(y) &= \frac{d}{dy} [F_X(y)]^n \\ &= n[F_X(y)]^{n-1} \frac{d}{dy} F_X(y) && \text{By Chain rule of derivative} \\ &= n[F_X(y)]^{n-1} f_X(y) \end{aligned}$$

Now we get the maximum variable. For the minimum variable  $X_{(1)}$  can be found using the similar way. The cdf of  $X_{(1)}$  is

$$F_{(1)}(y) = \mathbb{P}[X_{(1)} \leq y] = 1 - \mathbb{P}[X_{(1)} > y].$$

Since  $X_{(1)}$  is the minimum of  $X_1, X_2, \dots, X_n$ , and the event  $Y_i > y$  can be occurs for  $i = 1, 2, 3, \dots, n$ . In other words, any  $X_i$  in  $X_1, X_2, \dots, X_n$  can be the minimum variable. Hence


$$\begin{aligned} F_{(1)}(y) &= \mathbb{P}[X_{(1)} \leq y] = 1 - 1 - \mathbb{P}[X_{(1)} > y] \\ &= 1 - \mathbb{P}[X_1 > y, X_2 > y, \dots, X_n > y] \\ &= 1 - \mathbb{P}[X_1 > y] \mathbb{P}[X_2 > y] \cdots \mathbb{P}[X_n > y] \\ &= 1 - [1 - F_X(y)]^n. \end{aligned}$$




### Theorem 2.7 k-th order statistics

Let  $X_1, X_2, \dots, X_n$  be i.i.d continuous random variable with common cdf  $F_X(y)$  and common density function  $f_X(y)$ . Let  $X_{(k)}$  denote the  $k$ -th order Statistics, then the density function of  $X_{(k)}$  is

$$g_{(n)}(y) = \frac{n!}{(k-1)!(n-k)!} [F_X(y)]^{k-1} [1 - F_X(y)]^{n-k} f_X(y), \quad -\infty < y < \infty. \quad (2.7)$$

 **Example 2.3.1.** Let  $Y \sim \text{Uniform}(0, \theta)$  be the waiting time of bus arrival. A random samples of size  $n = 5$  is taken. Then,

1. Find the distribution of minimum variable.
2. Find the probability that  $Y_{(3)}$  is less than  $\frac{2}{3}\theta$ .
3. Suppose that the waiting time for bus arrival is uniformly distributed on 0 to 15 minutes, find  $\mathbb{P}[Y_{(5)} < 10]$ .

 **Solution** 1. The density of  $X_{(1)}$  is

$$\begin{aligned} Y_{(1)} \sim g_{(1)}(y) &= \frac{5!}{(1-1)!(5-1)!} [F_Y(y)]^{1-1} [1 - F_Y(y)]^{5-1} f_Y(y) \\ &= \frac{5!}{0!4!} [1 - F_Y(y)]^4 f_Y(y) \\ &= 5 \left(1 - \frac{y}{\theta}\right)^4 \left(\frac{1}{\theta}\right) \\ &= \frac{5(\theta - y)^4}{\theta^5}. \end{aligned}$$

Hence compute the mean of  $X_{(1)}$ ,

$$\mathbb{E}[Y_{(1)}] = \int_0^\theta y \left[ \frac{5(\theta - y)^4}{\theta^5} \right] dy = \int_0^\theta \frac{5y(\theta - y)^4}{\theta^5} dy \quad (\clubsuit)$$

using the substitution method and letting  $u = \theta - y$ , and for that

$$y = \theta - u \implies -du = dy$$

substitute back into  $(\clubsuit)$  and we have

$$\begin{aligned} (\clubsuit) &= \int_0^\theta \frac{5(\theta - u)u^4}{\theta^5} (-du) = -\frac{1}{\theta^5} \int_0^\theta (5\theta u^4 - u^5) du \\ &= -\frac{1}{\theta^5} \left[ \theta u^5 - \frac{1}{6} \theta^6 \right]_{u=0}^{u=\theta} \\ &= -\frac{1}{\theta^5} \left[ 0 - \frac{1}{6} \theta^6 \right] \\ &= \frac{\theta}{6} = \mathbb{E}[Y_{(1)}]. \end{aligned}$$

2. First we need to find the probability density function of  $Y_{(3)}$ , that is,

$$\begin{aligned} Y_{(3)} \sim g_{(3)}(y) &= \frac{5!}{(3-1)!(5-3)!} [F_Y(y)]^{3-1} [1 - F_Y(y)]^{5-3} f_Y(y) \\ &= 30 \left(\frac{y}{\theta}\right)^2 \left(1 - \frac{y}{\theta}\right)^2 \frac{1}{\theta}, \quad 0 < y < \theta. \end{aligned}$$

Compute the probability on which that  $Y_{(3)}$  is smaller than  $\frac{2\theta}{3}$ .

$$\begin{aligned}
 \mathbb{P}[Y_{(3)} < \frac{2}{3}\theta] &= \int_0^{\frac{2}{3}\theta} 30 \left(\frac{y}{\theta}\right)^2 \left(1 - \frac{y}{\theta}\right)^2 \frac{1}{\theta} dy \\
 &= \frac{30}{\theta^5} \int_0^{\frac{2}{3}\theta} y^2 (\theta^2 - 2\theta y + y^2) dy \\
 &= \frac{30}{\theta^5} \left[ \frac{1}{3} \theta^2 y^3 - \frac{1}{2} \theta y^4 + \frac{1}{5} y^5 \right]_{y=0}^{y=\frac{2}{3}\theta} \\
 &= 30 \left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^3 - 15 \left(\frac{2}{3}\right)^4 + 6 \left(\frac{2}{3}\right)^5 \\
 &= \frac{64}{81}.
 \end{aligned}$$

3. The probability that  $Y_{(5)}$  less than 10 minutes is equivalent to taking the bus five times. That is

$$\begin{aligned}
 \mathbb{P}[Y_{(5)} < 10] &= \mathbb{P}[Y_{(1)} < 10, Y_{(2)} < 10, \dots, Y_{(5)} < 10] \\
 &= \mathbb{P}[Y_{(1)} < 10] \times \mathbb{P}[Y_{(2)} < 10] \times \dots \times \mathbb{P}[Y_{(5)} < 10] \\
 &= \left(\frac{10}{15}\right)^5 = \frac{32}{243}.
 \end{aligned}$$

□

## Decision theory

For the given observation  $\mathcal{X}$ , we decide to take an action  $a \in \mathcal{A}$ . An action is a map  $a : \mathcal{X} \rightarrow \mathcal{A}$  with  $a(X)$  being the decision taken.


$L(\theta, a)$  denoted as the "loss function", it is the loss incurred when state is  $\theta$  and an action  $a$  is taken.

$$L : \Theta \times \mathcal{A} \rightarrow \mathbb{R}. \quad (3.1)$$

### 3.1 Conditional Distributions


Recall the definition of conditional probabilities: For two sets  $A$  and  $B$ , with  $P(A) \neq 0$ , the conditional probability of  $B$  given that  $A$  is true is defined as

$$\mathbb{P}(B | A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \quad (3.2)$$

 **Example 3.1.1.** Let  $X$  and  $Y$  be two jointly continuous random variable with joint density function

$$f_{XY}(x, y) = \begin{cases} x^2 + \frac{1}{3}y, & -1 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

For  $0 \leq y \leq 1$ , find the conditional pdf of  $X$  given  $Y = y$ .

 **Solution** First we find the marginal distribution of  $Y$ , which we can obtain by integrating along with  $x$ .

$$\begin{aligned} f_Y(y) &= \int_{-1}^1 f_{XY}(x, y) \, dx = \int_{-1}^1 \left( x^2 + \frac{1}{3}y \right) \, dx \\ &= \frac{1}{3}x^3 + \frac{1}{3}xy \Big|_{-1}^1 \\ &= \frac{2}{3}(1 + y). \end{aligned}$$


The conditional distribution of  $X$  given  $Y = y$  is

$$f_{X|Y=y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{x^2 + \frac{1}{3}y}{\frac{2}{3}(1 + y)} = \frac{3x^2 + y}{2(1 + y)}, \quad -1 \leq x \leq 1, 0 \leq y \leq 1$$

□

**Definition 3.1 Unbiased estimator**

The estimator  $\hat{\mu}$  is unbiased if  $\text{Bias}(\hat{\mu} | \theta) = 0$

 **Example 3.1.2 (Two-sample mean problems).** Consider the observations  $X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu, \sigma^2)$  response under control treatment. And  $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu + \Delta, \sigma^2)$  are explanatory data response under test treatment where  $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$ .  $\sigma^2$  is unknown variance and  $\Delta \in \mathbb{R}$  is unknown treatment effect.

We define two testing hypotheses:

$$H_0 : P \in \{P : \Delta = 0\} = \{P_\theta : \theta \in \Theta_0\}$$

$$H_1 : P \in \{P : \Delta \neq 0\} = \{P_\theta : \theta \notin \Theta_0\}$$

By construct decision rule accepting null hypothesis  $H_0$  if estimate of  $\Delta$  is significantly far away from zero. For instance,  $\hat{\Delta} = \bar{Y} - \bar{X}$  to be the estimate difference in sample means. Since  $\sigma$  is unknown, we use  $\hat{\sigma}$  to estimate true  $\sigma$ . The decision procedure is


$$\delta(X, Y) = \begin{cases} 1 & \text{if } |\frac{\hat{\Delta}}{\hat{\sigma}}| < c \\ 0 & \text{if } |\frac{\hat{\Delta}}{\hat{\sigma}}| \geq c \end{cases}$$

We again define a zero-one loss function to make decision

$$L(\theta, a) = \begin{cases} 0 & \text{if } \theta \in \Theta_a \quad (\text{correct action}) \\ 1 & \text{if } \theta \notin \Theta_a \quad (\text{wrong action}) \end{cases}.$$

The risk function is linear combination of the loss of correct and wrong actions,

$$\begin{aligned} R(\theta, \delta) &= L(\theta, 0)P_\theta(\delta(X, Y) = 0) + L(\theta, 1)P_\theta(\delta(X, Y) = 1) \\ &= \begin{cases} P_\theta(\delta(X, Y) = 1) & \text{if } \theta \in \Theta_0 \\ P_\theta(\delta(X, Y) = 0) & \text{if } \theta \notin \Theta_0 \end{cases} \end{aligned}$$

 **Example 3.1.3 (Statistical testing).** We are going to use the random variable  $X \sim P_\theta$  with sample space  $\mathcal{X}$  and parameter space  $\Theta$ , we want to test the testing hypothesis

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \notin \Theta_0.$$

We construct the critical region of a test  $\delta$  as

$$C = \{x : \delta(x) = 1\}.$$

with zero-one loss. Note that


❖ Type I error: the test  $\delta(X)$  rejects  $H_0$  when  $H_0$  is true.

❖ Type II error: the test  $\delta(X)$  accepts  $H_0$  when  $H_0$  is false.

The risk under zero-one loss as

$$R(\theta, \delta) = P_\theta(\delta(X) = 1 \mid \theta) \quad \text{if } \theta \in \Theta_0 \\ = \text{Probability of Type I error.}$$

$$R(\theta, \delta) = P_\theta(\delta(X) = 0 \mid \theta) \quad \text{if } \theta \notin \Theta_0 \\ = \text{Probability of Type II error.}$$

 **Example 3.1.4** (Statistical testing with two different hypothesis subspace). We are going to use the random variable  $X \sim P_\theta$  with sample space  $\mathcal{X}$  and parameter space  $\Theta$ , we want to test the testing hypothesis

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \notin \Theta_0.$$

We construct the critical region of a test  $\delta$  as

$$C = \{x : \delta(x) = 1\}.$$

with zero-one loss. Note that


- ❖ Type I error: the test  $\delta(X)$  rejects  $H_0$  when  $H_0$  is true.
- ❖ Type II error: the test  $\delta(X)$  accepts  $H_0$  when  $H_0$  is false.

The risk under zero-one loss as

$$R(\theta, \delta) = P_\theta(\delta(X) = 1 \mid \theta) \quad \text{if } \theta \in \Theta_0 \\ = \text{Probability of Type I error.}$$

$$R(\theta, \delta) = P_\theta(\delta(X) = 0 \mid \theta) \quad \text{if } \theta \notin \Theta_0 \\ = \text{Probability of Type II error.}$$

## 3.2 Value-at-risk

 **Example 3.2.1** (Confidence Interval). We altering the previous decision framework setup:

- ❖  $X$  is a random variable with probability  $P_\theta$ .
- ❖ The parameter of interest is  $\mu(\theta)$ .
- ❖ Define  $\mathcal{U} = \{\mu = \mu(\theta) : \theta \in \Theta\}$ .
- ❖ Objective: we want to construct an interval estimation of  $\mu(\theta)$ .
- ❖ Action space:  $\mathcal{A} = \{\mathbf{a} = [\underline{a}, \bar{a}] : \underline{a} < \bar{a} \in \mathcal{U}\}$ .
- ❖ Interval Estimator: define a map  $\hat{\mu}(X) : \mathcal{X} \rightarrow \mathcal{A}$ , that is  $\hat{\mu}(X) = [\hat{\mu}_{\text{Lower}}(X), \hat{\mu}_{\text{Upper}}(X)]$

Note that  $\theta$  is not random, the interval is random given a fixed  $\theta$ . We have to use Bayesian models to compute

$$\mathbb{P}[\mu(\theta) \in [\hat{\mu}_{\text{Lower}}(X), \hat{\mu}_{\text{Upper}}(X)] \mid X = x].$$

We define the zero-one loss function

$$L(\theta, (\underline{a}, \bar{a})) = \begin{cases} 1 & \text{if } \underline{a} > \mu(\theta) \text{ or } \bar{a} < \mu(\theta) \\ 0 & \text{otherwise.} \end{cases}$$

The risk function under zero-one loss is

$$\begin{aligned} R(\theta, \hat{\mu}(X)) &= \mathbb{E}_X[L(\theta, \hat{\mu}(X)) \mid \theta] \\ &= P_\theta(\hat{\mu}_{\text{Lower}}(X) > \mu(\theta) \text{ or } \hat{\mu}_{\text{Upper}}(X) < \mu(\theta)) \\ &= 1 - P_\theta(\hat{\mu}_{\text{Lower}}(X) \leq \mu(\theta) \leq \hat{\mu}_{\text{Upper}}(X) \mid \theta). \end{aligned}$$

It is said that the interval estimator  $\hat{\mu}(\theta)$  has confidence level  $1 - \alpha$  if

$$P_\theta(\hat{\mu}_{\text{Lower}}(X) \leq \mu(\theta) \leq \hat{\mu}_{\text{Upper}}(X) \mid \theta) \geq (1 - \alpha) \quad \forall \theta \in \Theta.$$

Equivalently, we can said  $R(\theta, \hat{\mu}(X)) \leq \alpha$  for all  $\theta \in \Theta$ .

### 3.3 Admissible

On basis of performance measure by the risk function  $R(\theta, \delta)$ , some rules are obviously bad. We said that a decision procedure  $\delta(\cdot)$  is inadmissible if  $\exists \delta'$  such that

$$R(\theta, \delta') \leq R(\theta, \delta) \quad \forall \theta \in \Theta \quad (3.3)$$

with strict inequality for some  $\theta$ .

**Example 3.3.1.** Suppose, for  $n \geq 2$ , the observations  $X_1, X_2, \dots, X_n$  be i.i.d with mean  $g(\theta) := \mathbb{E}_\theta[X_i] = \mu$ , and  $\text{Var}[X_i] = 1$  for all  $i$ . We take quadratic loss

$$L(\theta, a) := |\mu_X - a|^2.$$

Consider the decision

$$\delta'(X_1, X_2, \dots, X_n) := \bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n},$$

and  $\delta(X_1, X_2, \dots, X_n) := X_1$ . Then for all  $\theta$ , we have

$$R(\theta, \delta') = \frac{1}{n}, \quad R(\theta, \delta) = 1.$$

Therefore  $\delta$  is inadmissible.



# Estimation

## Definition 4.1 Estimator

An **estimator** is a formula, that tells how to calculate the value of an estimate based on the observations contained in a sample.

For example, the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is a rule that tells us how to calculate the estimate of the population mean  $\mu$  based on the observations in a sample.

## Theorem 4.1 Mean Squared Error

For an estimator  $\hat{\mu}(X)$  of  $\mu = \mu(\theta)$ , the mean-squared error is

$$MSE(\hat{\mu}) = Var[\hat{\mu}(X) | \theta] + Bias(\hat{\mu} | \theta)^2 \quad (4.1)$$

where  $Bias(\hat{\mu} | \theta) = \mathbb{E}_\theta[\hat{\mu}(X) | \theta] - \mu$ .

*Proof.* Consider the following decision framework:

- ❖  $X \sim P_\theta, \theta \in \Theta$ .
- ❖ The parameter of interest,  $\mu(\theta)$  is a certain function.
- ❖ Action space,  $\mathcal{A} = \{\mu = \mu(\theta), \theta \in \Theta\}$ .
- ❖ Decision procedure (or estimator),  $\hat{\mu}(X) : \mathcal{X} \rightarrow \mathcal{A}$ .
- ❖ Squared error loss as loss function:  $L(\theta, a) = [a - \mu(\theta)]^2$ .

with the setup above, the MSE is equal to the risk of decision,

$$\begin{aligned} R(\theta, \hat{\mu}(X)) &= \mathbb{E}[L((\theta, \hat{\mu}(X)) | \theta)] \\ &= \mathbb{E}[(\hat{\mu}(X) - \mu(\theta))^2 | \theta] \\ &= \mathbb{E}[(\hat{\mu}(X) - \mu)^2 | \theta] \\ &= Var[\hat{\mu}(X) | \theta] + \underbrace{(\mathbb{E}[\hat{\mu}(X) | \theta] - \mu)^2}_{Bias(\hat{\mu}|\theta)} \end{aligned}$$

□

## 4.1 Evaluating the Estimators

**Example 4.1.1.** Let  $X_1, X_2, X_3$  be a random sample of size 3 from a population with pmf

$$f(x|\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

where  $\lambda > 0$  is a parameter. Are the following estimators of  $\lambda$  unbiased?

$$\hat{\lambda}_1 = \frac{1}{4}(X_1 + 2X_2 + X_3), \quad \hat{\lambda}_2 = \frac{1}{9}(4X_1 + 3X_2 + 2X_3)$$

Given,  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  which one is more efficient?

Hence, find an unbiased estimator of  $\lambda$  that is more efficient than both  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$ .

**Solution** Given the observations  $X_1, X_2, X_3$  are i.i.d with  $X_i \sim \text{Poisson}(\lambda)$ , we have

$$\mathbb{E}[X_i] = \text{Var}[X_i] = \lambda \quad \forall i = 1, 2, 3.$$

It is easy to see that

$$\begin{aligned} \mathbb{E}[\hat{\lambda}_1] &= \frac{1}{4}(\mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_3]) = \frac{1}{4}(\lambda + 2\lambda + \lambda) = \lambda, \\ \mathbb{E}[\hat{\lambda}_2] &= \frac{1}{9}(4\mathbb{E}[X_1] + 3\mathbb{E}[X_2] + 2\mathbb{E}[X_3]) = \frac{1}{9}(4\lambda + 3\lambda + 2\lambda) = \lambda. \end{aligned}$$

Thus, both  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  are unbiased estimators of  $\lambda$ . Next, we compute the variances of both estimators,

$$\begin{aligned} \text{Var}[\hat{\lambda}_1] &= \frac{1}{16}(\text{Var}[X_1] + 4\text{Var}[X_2] + \text{Var}[X_3]) = \frac{1}{16}(\lambda + 4\lambda + \lambda) = \frac{3\lambda}{8}, \\ \text{Var}[\hat{\lambda}_2] &= \frac{1}{81}(16\text{Var}[X_1] + 9\text{Var}[X_2] + 4\text{Var}[X_3]) = \frac{1}{81}(16\lambda + 9\lambda + 4\lambda) = \frac{29\lambda}{81}. \end{aligned}$$

By inspection, since  $\frac{3}{8} = 0.375 > \frac{29}{81} \approx 0.358$ , the estimator  $\hat{\lambda}_2$  is more efficient than  $\hat{\lambda}_1$ . We have seen in previous section that the sample mean is always an unbiased estimator of the population mean irrespective of the population distribution. The variance of the sample mean is always equal to  $\frac{\sigma^2}{n}$ , where  $\sigma^2$  is the population variance and  $n$  is the sample size. Thus

$$\text{Var}[\bar{X}] = \frac{\text{Var}[X_i]}{3} = \frac{1}{3}\lambda.$$

The sample mean has even smaller variance than both  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$ . Thus,  $\bar{X} = \frac{1}{3}\lambda$  is an unbiased estimator of  $\lambda$  that is more efficient than both  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$ .  $\square$

**Rule of thumb choosing a good estimator:**

❖ Unbiasedness:  $\mathbb{E}[\hat{\theta}] = \theta$ .


❖ Minimum variance: A good estimator should have smaller  $\text{Var}[\hat{\theta}]$ , the smaller the better.

## 4.2 Point Estimators


A **point estimator** is a function of the sample data that provides a single value as an estimate of an unknown population parameter. Since the estimator is calculated from a random sample, it is itself a random variable and has a probability distribution, called the **sampling distribution**.

The sampling distribution of a point estimator describes how the estimator varies from sample to sample. Key properties of the sampling distribution include its mean (which relates to bias) and its variance (which relates to the precision of the estimator). Understanding the sampling distribution is fundamental for assessing the reliability of an estimator, constructing confidence intervals, and performing hypothesis tests.

	Target Parameter	Sample size	Point Estimator	$\mathbb{E}[\theta]$	Standard Error
Population Mean	$\mu$	$n$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	$\mu$	$\frac{\sigma}{\sqrt{n}}$
Proportion	$p$	$n$	$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$	$p$	$\sqrt{\frac{p(1-p)}{n}}$
Difference in Means	$\mu_1 - \mu_2$	$m, n$	$\bar{X} - \bar{Y}$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
Difference in Proportions	$p_1 - p_2$	$m, n$	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$

 **Example 4.2.1.** In a random sample of 80 components of a certain type, 12 are found to be defective.

1. Find a point estimate of the proportion of non-defective components.
2. Find the standard error of the point estimate.

 **Solution** 1. With  $p$  as the proportion of non-defective components, the point estimate for proportion is

$$\hat{p} = \frac{80 - 12}{80} = 0.85.$$

2. The standard error of the point estimate of non-defective proportion is

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.85 \times 0.15}{80}} \approx 0.0399.$$

□