

# Lecture Note:

# Foundations of Probability Theory

PROBABILITY AND STATISTICS B

---

Teruo Nakatsuma

Fall Semester 2019

Faculty of Economics, Keio University

# Aims Of This Course

1. Learn foundations of probability theory.
2. Study about basic concepts in probability theory such as conditional probabilities, random variables, probability distributions.
3. Learn the law of large numbers and the central limit theorem.
4. Hands-on practice of Python.

# Reading List i

1. “Helicopter tour” of mathematical statistics
  - Casella, G., and R.L. Berger, (2001) *Statistical Inference* 2nd ed., Cengage Learning.
  - Mittelhammer, R.C. (2013) *Mathematical Statistics for Economics and Business*, 2nd. ed., Springer.
  - Wasserman, L., (2010) *All of Statistics: A Concise Course in Statistical Inference*, Springer.
2. Measure-theoretic treatment of probability
  - Chung, K.L., (2001) *A Course in Probability Theory*, 3rd ed., Academic Press.

# Reading List ii

- Resnick , S.I., (2014) *A Probability Path*, 3rd ed., Birkhäuser.
- Rosenthal, J.S., (2006) *A First Look At Rigorous Probability Theory*, 2nd ed., WSPC.

## 3. Theoretical foundations of statistical inference

- Lehmann, E.L., and J.P. Romano, (2005) *Testing Statistical Hypothesis*, 3rd ed., Springer.
- Lehmann, E.L., and G. Casella, (1998) *Theory of Point Estimation*, 2nd ed., Springer.
- van der Vaart, A.W., (1998) *Asymptotic Statistics*, Cambridge University Press.

# Definitions of Probability and Related Concepts

- **Probability:** the extent to which something is probable; the likelihood of something happening or being the case:
- **Probable:** likely to be the case or to happen
- **Possibility:** a thing that may happen or be the case
- **Possible:** able to be done; within the power or capacity of someone or something
- **Chance:** a possibility of something happening

*New Oxford American Dictionary*, Oxford University Press, 2017.

# Experiments

The term **experiment** is used in probability theory to describe any procedure whose outcome is not known in advance with certainty. Examples of experiments are as follows:

1. In an experiment in which a coin is to be tossed 10 times, the researcher wants to know the probability that at most 4 heads will be obtained.
2. A bank lends 10 billion yens to a company. The bank wants to know the probability that the company will go bankrupt in a year.
3. The manager of the convenient store wants to know how many ham-and-egg sandwiches should be ordered for the next day.

# Sample Space and Events

- The collection of all possible outcomes of an experiment is called the **sample space** of the experiment. An **event** is a collection of some outcomes of the experiment, which characterizes the result of the experiment.
- Let  $\Omega$  denote the sample space of some experiment and let  $\omega$  denote a possible outcome of the experiment. Since any outcome is a member of the sample space, the statement that  $\omega$  is an outcome of the experiment is denoted by  $\omega \in \Omega$ .
- An event is a collection of some outcomes. Thus the event  $A$  is a subset of the sample space  $\Omega$ , i.e.,  $A \subseteq \Omega$ .

## Example: Bankruptcy

Suppose that the bank wants to know whether a company will go bankrupt or not. In this case, the sample space  $\Omega$  consists of two outcomes:

$$\begin{cases} \omega_1 : & \text{The company will go bankrupt;} \\ \omega_2 : & \text{The company will not go bankrupt.} \end{cases}$$

Let  $A$  denote the event that the company owing 10 billion yen to the bank will go bankrupt. Then

$$A = \{\omega_1\}.$$



## Example: Tossing a Coin i

In an experiment in which a coin is to be tossed twice, all possible outcomes are ***HH***, ***HT***, ***TH***, and ***TT*** where ***H*** indicates a head and ***T*** indicates a tail. Thus the sample space  $\Omega$  consists of the following four outcomes:

$$\left\{ \begin{array}{ll} \omega_1 : & \mathbf{HH}; \\ \omega_2 : & \mathbf{HT}; \\ \omega_3 : & \mathbf{TH}; \\ \omega_4 : & \mathbf{TT}. \end{array} \right.$$

That is,  $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\} = \{\mathbf{HH}, \mathbf{HT}, \mathbf{TH}, \mathbf{TT}\}$ .

## Example: Tossing a Coin ii

Let us define the following three events:

$$\mathbf{A} = \{\omega_1, \omega_2, \omega_3\} = \{HH, HT, TH\},$$

$$\mathbf{B} = \{\omega_2, \omega_4\} = \{HT, TT\},$$

$$\mathbf{C} = \{\omega_4\} = \{TT\}.$$

**A** is the event that at least one head is obtained; **B** is the event that a tail is obtained on the second toss; **C** is the event that no heads are obtained.

# Set Operations i

Since events are mathematically equivalent to subsets of the sample space, we can apply ordinary set operations to them. For example, an intersection of two events  $\mathbf{A} \cap \mathbf{B}$  is the event that both  $\mathbf{A}$  and  $\mathbf{B}$  occur. A union of two events  $\mathbf{A} \cup \mathbf{B}$ , on the other hand, is the one that either  $\mathbf{A}$  or  $\mathbf{B}$  occurs. The complement of an event  $\mathbf{A}$ , denote by  $\mathbf{A}^c$ , is the event that  $\mathbf{A}$  does not occur. If  $\mathbf{A} \cap \mathbf{B} = \emptyset$ , we say that  $\mathbf{A}$  and  $\mathbf{B}$  are **mutually exclusive** or **pairwise disjoint**. Useful set operations are listed in the following table.

## Set Operations ii

$A \cap B$  intersection,  $\{\omega : \omega \in A \text{ and } \omega \in B\}$

event that both  $A$  and  $B$  occurs

$A \cup B$  union,  $\{\omega : \omega \in A \text{ or } \omega \in B\}$

event that  $A$  and/or  $B$  occurs

$A^c$  complement of  $A$ ,  $\{\omega : \omega \notin A\}$

event that  $A$  does not occurs

$A \setminus B$  difference,  $\{\omega : \omega \in A \text{ and } \omega \notin B\} = A \cap B^c$

$A$  occurs but  $B$  does not

$A \subseteq B$   $A$  is a subset of  $B$ ,  $\forall \omega \in A, \omega \in B$

$A$  occurs, then  $B$  occurs

$A = B$   $A$  and  $B$  are equivalent

$A \subset B$   $\forall \omega \in A, \omega \in B$  but  $\exists \omega \in B$  such that  $\omega \notin A$

## Set Operations iii

The intersection and the union of a sequence of events  $\{A_i\}_{i=1}^n$  are defined as follows:

$$\bigcap_{i=1}^n A_i = \{\omega \in \Omega : \forall i \in \{1, \dots, n\}, \omega \in A_i\},$$
$$\bigcup_{i=1}^n A_i = \{\omega \in \Omega : \exists i \in \{1, \dots, n\}, \omega \in A_i\}.$$

The famous **de Morgan's law**

$$\left(\bigcup_{i=1}^n A_i\right)^c = \bigcap_{i=1}^n A_i^c, \quad \left(\bigcap_{i=1}^n A_i\right)^c = \bigcup_{i=1}^n A_i^c,$$

is also applicable to events.

## Example: Tossing a Coin

$$A = \{\omega_1, \omega_2, \omega_3\} = \{HH, HT, TH\},$$

$$B = \{\omega_2, \omega_4\} = \{HT, TT\},$$

$$C = \{\omega_4\} = \{TT\}.$$

Various relations among these events can be derived. For example,

$$C = A^c,$$

$$C \subset B,$$

$$A \cup B = \Omega,$$

$$A \cap C = \emptyset.$$

# $\sigma$ -Field

## Field

A **field** (also **algebra**)  $\mathcal{F}$  on  $\Omega$  is a collection of sets such that

1.  $\Omega \in \mathcal{F}$ .
2. if  $A \in \mathcal{F}$ , then  $A^c \in \mathcal{F}$ .
3. if  $A \in \mathcal{F}$  and  $B \in \mathcal{F}$ , then  $A \cup B \in \mathcal{F}$ .

The simplest field is  $\mathcal{F} = \{\Omega, \emptyset\}$ . #3 implies

$$A_1 \cup A_2 \cup \dots \cup A_n \in \mathcal{F} \quad \text{if} \quad A_1, A_2, \dots, A_n \in \mathcal{F}.$$

When  $n$  can be infinite, such  $\mathcal{F}$  is called the  **$\sigma$ -field**. A ( $\sigma$ -)field is regarded as an “exhaustive” collection of events which are of our interest.

## Example: Bankruptcy

$$\begin{cases} \omega_1 : & \text{The company will go bankrupt;} \\ \omega_2 : & \text{The company will not go bankrupt.} \end{cases}$$

The sample space is  $\Omega = \{\omega_1, \omega_2\}$ .

The simplest field on  $\Omega$  is

$$\mathcal{F}_0 = \{\{\omega_1, \omega_2\}, \emptyset\}.$$

More exhaustive one is

$$\mathcal{F}_1 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_1, \omega_2\}, \emptyset\}.$$



## Example: Asset Price $i$

Suppose there are four possible paths the future asset prices will take.

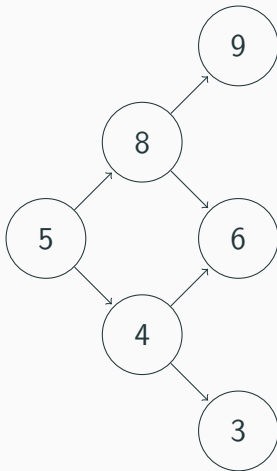
**Table 1:** Asset Price Fluctuations

Time	Path			
	$\omega_1$	$\omega_2$	$\omega_3$	$\omega_4$
0	5	5	5	5
1	8	8	4	4
2	9	6	6	3

The initial price is 5 for all paths.

## Example: Asset Price ii

$t = 0$     $t = 1$     $t = 2$



**Figure 1:** Tree of Price Fluctuations

# Filtration

Let  $\mathcal{F}_t$  denote the  $\sigma$ -field at time  $t$ . The  $\sigma$ -field evolves as

$$\mathcal{F}_0 = \{\{\omega_1, \omega_2, \omega_3, \omega_4\}, \emptyset\},$$

$$\mathcal{F}_1 = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \{\omega_1, \omega_2, \omega_3, \omega_4\}, \emptyset\},$$

$$\mathcal{F}_2 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\},$$

$$\{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \{\omega_1, \omega_4\}, \{\omega_2, \omega_3\}, \{\omega_2, \omega_4\}, \{\omega_3, \omega_4\},$$

$$\{\omega_1, \omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_4\}, \{\omega_1, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_4\},$$

$$\{\omega_1, \omega_2, \omega_3, \omega_4\}, \emptyset\}.$$

Note that  $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2$ . A sequence of  $\sigma$ -fields such that

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_t \subseteq \cdots$$

is called a **filtration**.

# Definition of Probability

The formal (axiomatic) definition of probability is as follows:

## Axiomatic Definition of Probability

Suppose  $\Omega$  is a sample space and  $\mathcal{F}$  is a ( $\sigma$ -)field on  $\Omega$ .

**Axiom 1.** For any event  $A \in \mathcal{F}$ ,  $P(A) \geq 0$ .

**Axiom 2.**  $P(\Omega) = 1$ .

**Axiom 3.** For any pairwise disjoint events  $A_1, \dots, A_n \in \mathcal{F}$ ,

$$P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n).$$

When  $\mathcal{F}$  is a  $\sigma$ -field,  $n$  must be infinite.

The triplet  $(\Omega, \mathcal{F}, P)$  is called the **probability space**.

# Properties

$A$  and  $B$  are events, and  $\{A_n\}_{n=1}^{\infty}$  is a sequence of events.

1.  $P(A) \leq P(B)$  if  $A \subseteq B$ .
2.  $P(A) \leq 1$ .
3.  $P(A^c) = 1 - P(A)$ .
4.  $P(\emptyset) = 0$ .
5.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .
6.  $P(B \setminus A) = P(B) - P(A)$  if  $A \subseteq B$ .
7.  $P(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n)$ .

## Example: Bankruptcy

Let  $\mathbf{A}$  denote the event that a company owing 10 billion yen to the bank will go bankrupt. The complement  $\mathbf{A}^c$  is the event that the company will not go bankrupt. We suppose the probability of  $\mathbf{A}$  is 1 percent, that is,

$$P(\mathbf{A}) = \frac{1}{100}.$$

Then the probability that the company will not go bankrupt is given by

$$P(\mathbf{A}^c) = 1 - P(\mathbf{A}) = 1 - \frac{1}{100} = \frac{99}{100}.$$

## Example: Tossing a Coin

We suppose that all outcomes are equally likely in the experiment in which a coin is to be tossed twice. This means

$$P(HH) = P(HT) = P(TH) = P(TT).$$

Since

$$P(HH) + P(HT) + P(TH) + P(TT) = 1,$$

we have

$$P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}.$$

# Permutations

A permutation is a sequence of elements taken from a finite set in which no element should be duplicated.

## Example: Anagram

All permutations of  $\{A,B,C\}$  are

“ABC,” “ACB,” “BAC,” “BCA,” “CAB,” “CBA.”

In general, we may consider a permutation of  $k$  out of  $n$  distinct elements. In this case, the number of all permutations is given by

$${}_nP_k = \frac{n!}{(n - k)!}.$$



## Application: Birthday Paradox i

The birthday paradox is about determining the probability that at least two persons have the same birthday (they were born on the same day of the same month, but not necessarily in the same year) in a group of people.

To make the problem simple, let us ignore the leap year. Then we have 365 possible birthdays for each person. If we have  $n$  persons, the number of all outcomes of this experiment is  **$365^n$** . Furthermore, the number of outcomes in which everyone was born on different days is

$$365 \times 364 \times \cdots \times (365 - n + 1) = {}_{365}P_n.$$

## Application: Birthday Paradox ii

Thus, assuming that all outcomes are equally possible, the probability that all  $n$  persons have distinct birthdays is given by

$$\frac{\text{\# of outcomes in which all birthdays are distinct}}{\text{\# of all outcomes}} = \frac{{}^{365}P_n}{365^n}.$$

Since the probability that at least two persons have the same birthday is the complement of the above event, we have

$$p = 1 - \frac{{}^{365}P_n}{365^n} = 1 - \frac{365!}{(365 - n)!365^n}.$$

$n$	5	10	20	23	30	40	50	60
$p$	0.027	0.117	0.411	0.507	0.706	0.891	0.970	0.994

# Combinations

A combination is a subset of a finite set in which the order of elements is ignored. For example, combinations of two letters among  $\{A,B,C\}$  are “AB,” “AC,” “BC.”

The number of combinations of  $k$  out of  $n$  elements is given by

$${}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

1.  $\binom{n}{k} = \binom{n}{n-k}.$
2.  $\binom{n}{0} = \binom{n}{n} = 1.$
3.  $(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$

## Example: Baseball Tournament

Suppose  $n$  baseball teams are entered in a tournament. In each round of the tournament, teams are randomly paired one against another. The winner will proceed to the next round but the loser will be eliminated. What is the probability that two specific teams will play against each other?

- The number of games:  $n - 1$ .
- The number of all possible pairs of teams:  $\binom{n}{2}$ .
- The probability that two specific teams will play in a specific round during a tournament:  $1/\binom{n}{2}$ .

The probability is

$$p = \frac{n - 1}{\binom{n}{2}} = \frac{(n - 1)2!(n - 2)!}{n!} = \frac{2}{n}.$$

# Independence

Two events are called **independent** if and only if

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A})P(\mathbf{B}).$$

In general, the  $n$  events  $\mathbf{A}_1, \dots, \mathbf{A}_n$  are independent if, for any subset  $\mathbf{A}_{i_1}, \dots, \mathbf{A}_{i_m}$ ,

$$P(\mathbf{A}_{i_1} \cap \dots \cap \mathbf{A}_{i_m}) = P(\mathbf{A}_{i_1}) \times \dots \times P(\mathbf{A}_{i_m}).$$

## Example: Defective Products i

Suppose that a factory produces an electronic device. It produces a defective device with probability  $p$  ( $0 < p < 1$ ) and produces a non-defective device with probability  $q = 1 - p$ .

Further suppose that 6 devices are randomly chosen for inspection and 2 of them turn out to be defective. The event that each of 6 devices is defective is independent each other.

We want to know the probability that 2 out of 6 devices are defective.

## Example: Defective Products ii

1	2	3	4	5	6
D	N	N	D	N	N
N	D	N	N	D	N
N	N	N	D	N	D
⋮					

where D stands for **defective** and N stands for **non-defective**. The number of all outcomes are  $\binom{6}{2}$ . The probability of each outcome is  $p^2q^4$ . Thus the probability is  $\binom{6}{2}p^2q^4$ .

# Conditional Probability

The **conditional probability** is defined as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

If **A** and **B** are independent,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

In other words, the probability of **A** will not be affected whether **B** occurs or not.



# Properties

1.  $P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B})$ .
2.  $P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) = P(\mathbf{B}|\mathbf{A})P(\mathbf{A})$ .
3.  $P(\mathbf{A} \cap \mathbf{B} \cap \mathbf{C}) = P(\mathbf{A}|\mathbf{B} \cap \mathbf{C})P(\mathbf{B}|\mathbf{C})P(\mathbf{C})$ .
4.  $P\left(\bigcap_{i=1}^n \mathbf{A}_i\right) = P(\mathbf{A}_1)P(\mathbf{A}_2|\mathbf{A}_1)P(\mathbf{A}_3|\mathbf{A}_1 \cap \mathbf{A}_2) \cdots P(\mathbf{A}_n|\bigcap_{i=1}^{n-1} \mathbf{A}_i)$ .
5.  $P(\mathbf{A}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) + P(\mathbf{A}|\mathbf{B}^c)P(\mathbf{B}^c)$ .
6. Suppose  $\mathbf{B}_1, \dots, \mathbf{B}_n$  are pairwise disjoint and  $\bigcup_{i=1}^n \mathbf{B}_i = \Omega$ . Then  $P(\mathbf{A}) = \sum_{i=1}^n P(\mathbf{A}|\mathbf{B}_i)P(\mathbf{B}_i)$ .

## Example: Urn i

Suppose we have an urn containing 6 red balls and 3 white balls and conduct the following experiment:

1. Pick one ball out of the urn.
2. If the ball is red, put it back and add one red ball to the urn.
3. If the ball is white, put it back and add one white ball to the urn.

## Example: Urn ii

**Table 2:** Balls in the Urn

		Picked Color	
		Red	White
Red	6	7	6
White	3	3	4

Let  $R_i$  denote the event that a red ball is picked in the  $i$ -th experiment.

The probability to pick a red ball twice in row

$$P(R_1 \cap R_2) = P(R_2|R_1)P(R_1) = \frac{7}{10} \times \frac{6}{9} = \frac{7}{15}.$$

## Example: Urn iii

The probability to pick a red ball in the second experiment

$$\begin{aligned}P(R_2) &= P(R_2|R_1)P(R_1) + P(R_2|R_1^c)P(R_1^c) \\&= \frac{7}{10} \times \frac{6}{9} + \frac{6}{10} \times \frac{3}{9} = \frac{2}{3}.\end{aligned}$$

Note that  $P(R_1 \cap R_2) \neq P(R_1)P(R_2)$ . Thus  $R_1$  and  $R_2$  are not independent.

# Bayes' Theorem

Bayes' theorem is defined as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

Suppose  $A_1, \dots, A_n$  are pairwise disjoint and  $\bigcup_{i=1}^n A_i = \Omega$ . Such a collection of events is called a **partition** of  $\Omega$ . Then a general form of Bayes' theorem is given by

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}.$$

## Example: Diagnosis Test i

Suppose a patient takes a diagnosis test for detecting a certain type of disease. The probability that the patient will have a positive reaction is 0.99 if he has this type of disease; otherwise, the probability that the patient will have a negative reaction is 0.99.

**Table 3:** Diagnosis Test

Reaction	Does the patient have the disease?	
	Yes ( $A$ )	No ( $A^c$ )
Positive ( $B$ )	0.99	0.01
Negative ( $B^c$ )	0.01	0.99

## Example: Diagnosis Test ii

In the general population, one out of 100,000 people has this type of disease. What is the probability that the patient has this type of disease when the reaction to the diagnosis test is positive?

$$\begin{aligned}P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\&= \frac{\frac{99}{100} \times \frac{1}{100,000}}{\frac{99}{100} \times \frac{1}{100,000} + \frac{1}{100} \times \frac{99,999}{100,000}} \\&= \frac{99}{100,099} \approx 9.89 \times 10^{-4}.\end{aligned}$$

## Example: Diagnosis Test iii

Instead, suppose that one out of 100 people has this type of disease. Then the conditional probability that the patient has this type of disease is

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{\frac{99}{100} \times \frac{1}{100}}{\frac{99}{100} \times \frac{1}{100} + \frac{1}{100} \times \frac{99}{100}} \\ &= \frac{99}{198} = \frac{1}{2}. \end{aligned}$$



Since

$$\begin{aligned}P(A) &= P(A|B)P(B) + P(A|B^c)P(B^c) \\ P(A^c) &= P(A^c|B)P(B) + P(A^c|B^c)P(B^c)\end{aligned}$$

we have

$$\begin{bmatrix} P(A) \\ P(A^c) \end{bmatrix} = \begin{bmatrix} P(A|B) & P(A|B^c) \\ P(A^c|B) & P(A^c|B^c) \end{bmatrix} \begin{bmatrix} P(B) \\ P(B^c) \end{bmatrix}$$

## Markov Chain ii

Let  $\mathbf{A}_t$  denote the event that  $\mathbf{A}$  occur at time  $t$  and we replace  $\mathbf{A}$  and  $\mathbf{B}$  with  $\mathbf{A}_{t+1}$  and  $\mathbf{A}_t$  respectively. Then

$$\begin{bmatrix} P(\mathbf{A}_{t+1}) \\ P(\mathbf{A}_{t+1}^c) \end{bmatrix} = \begin{bmatrix} P(\mathbf{A}_{t+1}|\mathbf{A}_t) & P(\mathbf{A}_{t+1}|\mathbf{A}_t^c) \\ P(\mathbf{A}_{t+1}^c|\mathbf{A}_t) & P(\mathbf{A}_{t+1}^c|\mathbf{A}_t^c) \end{bmatrix} \begin{bmatrix} P(\mathbf{A}_t) \\ P(\mathbf{A}_t^c) \end{bmatrix}$$

This is called a **Markov chain**. In the context of the Markov chain, each event is often referred to as a **state** and the conditional probability  $P(\mathbf{A}_{t+1}|\mathbf{A}_t^c)$  is interpreted as the probability that the chain moves from state  $\mathbf{A}^c$  at  $t$  to state  $\mathbf{A}$  at  $t + 1$ .

The matrix

$$\begin{bmatrix} P(\mathbf{A}_{t+1}|\mathbf{A}_t) & P(\mathbf{A}_{t+1}|\mathbf{A}_t^c) \\ P(\mathbf{A}_{t+1}^c|\mathbf{A}_t) & P(\mathbf{A}_{t+1}^c|\mathbf{A}_t^c) \end{bmatrix}$$

is called the **transition matrix**. When the transition matrix does not change over time, the Markov chain is **time-homogenous**. The vector of probabilities at time  $t = 1$  is called the **initial probability vector**.

In economics, finance and other fields, the Markov chain is widely used for modeling a time-varying probability.

# Markov Chain iv

Let  $\mathbf{p}_{1t} = P(\mathbf{A}_t)$ ,  $\mathbf{p}_{2t} = P(\mathbf{A}_t^c)$ , and let  $\pi_{ij}$  denote the  $(i, j)$  element of the time-homogenous transition matrix  $(i, j = 1, 2)$ . Then the Markov chain is rewritten as

$$\begin{bmatrix} \mathbf{p}_{1,t+1} \\ \mathbf{p}_{2,t+1} \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} \mathbf{p}_{1,t} \\ \mathbf{p}_{2,t} \end{bmatrix}$$

In general, the Markov chain with  $k$  states is given by

$$\underbrace{\begin{bmatrix} \mathbf{p}_{1,t+1} \\ \vdots \\ \mathbf{p}_{k,t+1} \end{bmatrix}}_{\mathbf{p}_{t+1}} = \underbrace{\begin{bmatrix} \pi_{11} & \cdots & \pi_{1k} \\ \vdots & \ddots & \vdots \\ \pi_{k1} & \cdots & \pi_{kk} \end{bmatrix}}_{\Pi} \underbrace{\begin{bmatrix} \mathbf{p}_{1,t} \\ \vdots \\ \mathbf{p}_{k,t} \end{bmatrix}}_{\mathbf{p}_t}$$

# Properties

1.  $\mathbf{p}_{t+k} = \mathbf{\Pi}^k \mathbf{p}_t$  ( $k = 1, 2, \dots$ ).
2. In particular,  $\mathbf{p}_{t+1} = \mathbf{\Pi}^t \mathbf{p}_1$ . Thus all  $\mathbf{p}_t$  in a Markov chain is determined by the initial probability vector  $\mathbf{p}_1$  and the transition matrix  $\mathbf{\Pi}$ .
3. The probability vector  $\mathbf{p}^*$  satisfies

$$\mathbf{p}^* = \mathbf{\Pi} \mathbf{p}^*$$

is called the **stationary probability vector** or **stationary distribution** of a Markov chain.

4. When the number of states is two, the stationary distribution is given by

$$p_1^* = \frac{1 - \pi_{22}}{2 - \pi_{11} - \pi_{22}}, \quad p_2^* = \frac{1 - \pi_{11}}{2 - \pi_{11} - \pi_{22}}$$

## Example: Business Cycle

Suppose that the business cycle of a country is determined by the following transition matrix:

**Table 4:** Transition Matrix  $\Pi$

Current Quarter	Previous Quarter	
	Expansion ( $E$ )	Recession ( $E^c$ )
Expansion ( $E$ )	0.9	0.25
Recession ( $E^c$ )	0.1	0.75

Then the stationary distribution is given by

$$P(E) = \frac{1 - 0.75}{2 - 0.9 - 0.75} = \frac{5}{7}, \quad P(E^c) = 1 - P(E) = \frac{2}{7}$$

# Random Variable i

A **random variable (r.v.)** is an association between events in a  $\sigma$ -field and real numbers in  $\mathbb{R}$ .

## Definition of a Random Variable

A random variable  $X$  is a function from  $\Omega$  onto  $\mathbb{R}$  such that

$$\forall B \in \mathcal{B}, \quad X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F},$$

where  $\mathcal{B}$  is the Borel  $\sigma$ -field on  $\mathbb{R}$ . A **Borel set** in  $\mathbb{R}$  is any set that can be formed as a union or an intersection of open intervals. The **Borel  $\sigma$ -field** on  $\mathbb{R}$  is the  $\sigma$ -field that contains all Borel sets in  $\mathbb{R}$ .

## Random Variable ii

Given the probability space  $(\Omega, \mathcal{F}, P)$ , we can define the probability that the value of  $X$  will belong to  $B \in \mathcal{B}$  as

$$Q(B) = P(X^{-1}(B)) = P(\{\omega : X(\omega) \in B\}).$$

It is straightforward to show that the above  $Q$  satisfies all three axioms of probability in terms of the Borel  $\sigma$ -field  $\mathcal{B}$ . Therefore  $Q$  is regarded as the probability on  $\mathcal{B}$  and the triplet  $(\mathbb{R}, \mathcal{B}, Q)$  is the probability space.



## Example: Tossing a Coin

Define

$$X = \begin{cases} 1, & \text{if a head is obtained;} \\ 0, & \text{if a tail is obtained.} \end{cases}$$

$X$  is a random variable since

$$\{\omega : X(\omega) = 1\} = \{\omega_1\},$$

$$\{\omega : X(\omega) = 0\} = \{\omega_2\},$$

$$\{\omega : X(\omega) \in \{0, 1\}\} = \{\omega_1, \omega_2\} = \Omega,$$

$$\{\omega : X(\omega) \in B\} = \emptyset \text{ for any } B \subseteq \mathbb{R} \setminus \{0, 1\}.$$

If  $P(\{\omega_1\}) = P(\{\omega_2\}) = \frac{1}{2}$ ,  $Q(X = 1) = Q(X = 0) = \frac{1}{2}$ .

## Example: Tossing a Coin

Consider an experiment in which a coin is to be tossed 10 times and let  $\mathbf{X}$  be the number of heads we observe. The possible values of  $\mathbf{X}$  are  $\mathbf{0}, \mathbf{1}, \dots, \mathbf{9}, \mathbf{10}$ . Let  $\mathbf{p}$  be the probability that we will obtain a head and suppose all outcomes are independent. Then the probability that we will obtain  $\mathbf{x}$  heads is given by

$$P(\mathbf{X} = \mathbf{x}) = \binom{\mathbf{10}}{\mathbf{x}} \mathbf{p}^{\mathbf{x}} (\mathbf{1} - \mathbf{p})^{\mathbf{10} - \mathbf{x}}$$

Alternatively, if we define a binary random variable as

$$\mathbf{Y}_i = \begin{cases} \mathbf{1} & \text{if a head is obtained;} \\ \mathbf{0} & \text{if a tail is obtained,} \end{cases} \quad (\mathbf{i} = \mathbf{1}, \dots, \mathbf{10}).$$

$\mathbf{X}$  is obtained by  $\mathbf{X} = \mathbf{Y}_1 + \dots + \mathbf{Y}_{10}$ .

## Example: Random Number Generation

We can construct a 10-digit binary random variable  $Z$  by using  $Y_i$  as the  $i$ -th digit of the number. For example, if we obtain

$$\begin{aligned} Y_1 = 0, & \quad Y_2 = 0, & \quad Y_3 = 1, & \quad Y_4 = 0, & \quad Y_5 = 1, \\ Y_6 = 0, & \quad Y_7 = 0, & \quad Y_8 = 0, & \quad Y_9 = 1, & \quad Y_{10} = 0. \end{aligned}$$

$Z$  is given by

$$\begin{aligned} Z &= Y_{10} \cdots Y_1 \\ &= 0100010100 = 2^8 + 2^4 + 2^2 = 256 + 16 + 4 = 276 \end{aligned}$$

and

$$P(Z = 276) = p^3(1 - p)^7$$

# Discrete Distribution

It is said that a random variable  $X$  has a **discrete distribution** if  $X$  can take only a countable number of different values. The term “countable” means that the number of values is either finite or as many as natural numbers  $(1, 2, 3, \dots)$ . Such  $X$  is often called a discrete random variable.

If  $X$  has a discrete distribution, the **probability function** or p.f.  $f(x)$  is defined as

$$f(x) = P(X = x).$$

# Properties

Suppose that  $\mathbf{X}$  is a discrete random variable and  $f(\mathbf{x})$  is its probability function.

1. For any  $\mathbf{x}$ ,  $0 \leq f(\mathbf{x}) \leq 1$ .
2. If  $\mathbf{x}$  is not one of the possible values of  $\mathbf{X}$ ,  $f(\mathbf{x}) = 0$ .
3. Let  $\{\mathbf{x}_i\}_{i=1}^{\infty}$  be a sequence of all possible values of  $\mathbf{X}$ .  
Then  $\sum_{i=1}^{\infty} f(\mathbf{x}_i) = 1$ .
4.  $P(\mathbf{X} \in \mathbf{A}) = \sum_{\mathbf{x}_i \in \mathbf{A}} f(\mathbf{x}_i)$ .

## Example: Discrete Uniform Distribution

The probability function of the **discrete uniform distribution** is

$$f(x) = \begin{cases} \frac{1}{k}, & \text{for } x = x_1, \dots, x_k; \\ 0, & \text{otherwise.} \end{cases}$$

For example, the number on a dice follows the discrete uniform distribution with  $x = 1, 2, \dots, 6$ .

## Example: Bernoulli Distribution

A random variable  $X$  has a **Bernoulli distribution** if

$$X = \begin{cases} 1, & \text{with probability } p; \\ 0, & \text{with probability } 1 - p. \end{cases}$$

The p.f. of the Bernoulli distribution is

$$f(x|p) = p^x(1 - p)^{1-x}, \quad x = 0, 1$$

We already studied this type of random variable in the example of a coin toss.

$$X = \begin{cases} 1, & \text{if a head is obtained;} \\ 0, & \text{if a tail is obtained.} \end{cases}$$

## Example: Binomial Distribution

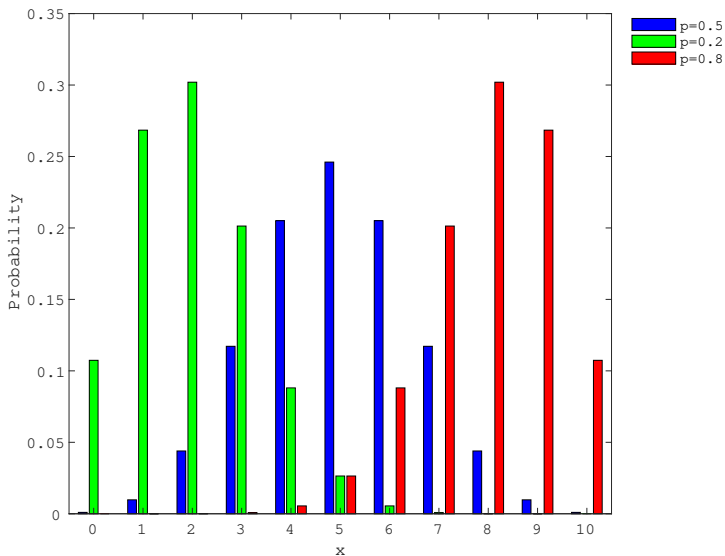
The p.f. of the **binomial distribution** is

$$f(x|n, p) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n,$$

where  $0 < p < 1$ . When a coin is to be tossed  $n$  times independently and the probability that a head is obtained is  $p$ , the number of heads follows this distribution.

Note that the binomial random variable  $X$  is defined as the sum of independent  $n$  Bernoulli random variables.





**Figure 2:** The p.f. of the binomial distribution

## Example: Geometric Distribution i

Suppose that we keep tossing a coin until we obtain a head. We consider a Bernoulli random variable  $Y_i$  ( $i = 1, 2, \dots$ ) with  $P(Y_i = 1) = p$  where  $Y_i = 1$  if we obtain a head; otherwise,  $Y_i = 0$ . We also assume that  $Y_1, Y_2, \dots$  are independent. Then the probability of the event that we obtain a head for the first time after we obtain  $x$  consecutive tails is given by

$$P(\text{a head after } x \text{ consecutive tails}) = p(1 - p)^x$$

In this case, the number of tails is a discrete random variable that has a **geometric distribution**.

## Example: Geometric Distribution ii

# of Tails	Sequence of $Y_i$
0	1
1	01
2	001
3	0001
$\vdots$	$\vdots$

In general, a geometric distribution is defined as the number of failures until the first success is obtained. The p.f. is

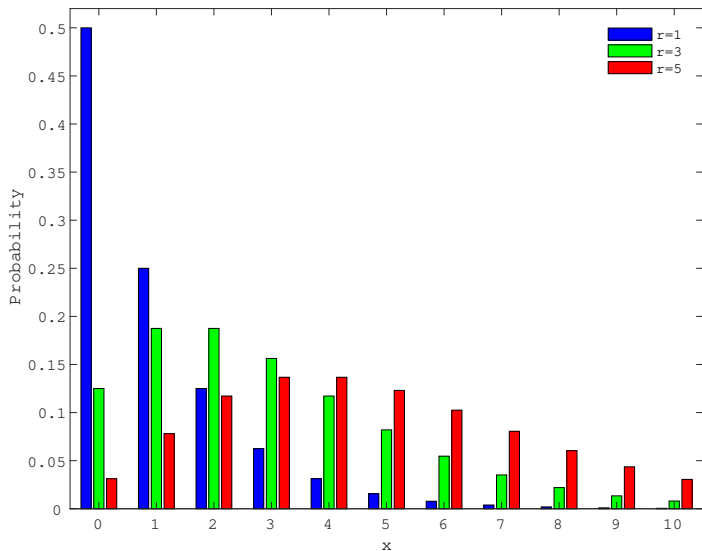
$$f(x|p) = p(1 - p)^x, \quad x = 0, 1, 2, \dots$$

## Example: Negative Binomial Distribution

Suppose that we need to obtain  $r$  heads, instead of just one, in the previous experiment. Note that for any sequence of heads and tails the last outcome should be a head, that is, the  $r$ -th head should be obtained in the very last toss. Since the number of all combinations with  $(r - 1)$  heads and  $x$  tails are  $\binom{x+r-1}{x}$ , the probability that we obtain  $x$  tails until we obtain  $r$  heads is given by

$$f(x|p) = \binom{x+r-1}{r-1} p^r (1-p)^x, \quad x = 0, 1, 2, \dots$$

which is the p.f. of a **negative binomial distribution**. The geometric distribution is a special case of the negative binomial distribution with  $r = 1$ .



**Figure 3:** The p.f. of the negative binomial distribution

## Example: Poisson Distribution i

The p.f. of a binomial distribution is

$$\begin{aligned}f(x|n, p) &= \binom{n}{x} p^x (1 - p)^{n-x} \\&= \frac{n(n-1) \cdots (n-x+1)}{x!} p^x (1 - p)^{n-x}\end{aligned}$$

If we let  $\lambda = np$ , we have

$$\begin{aligned}f(x|\lambda, n) &= \frac{n(n-1) \cdots (n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\&= \frac{\lambda^x}{x!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-x+1}{n} \left(1 - \frac{\lambda}{n}\right)^{-x} \left(1 - \frac{\lambda}{n}\right)^n\end{aligned}$$

## Example: Poisson Distribution ii

Let  $n \rightarrow \infty$  while we keep  $\lambda$  constant (so  $p$  is negligibly small). Then

$$\lim_{n \rightarrow \infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-x+1}{n} \left(1 - \frac{\lambda}{n}\right)^{-x} = 1$$
$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

Thus we have

$$f(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

which is the p.f. of a **Poisson distribution**.

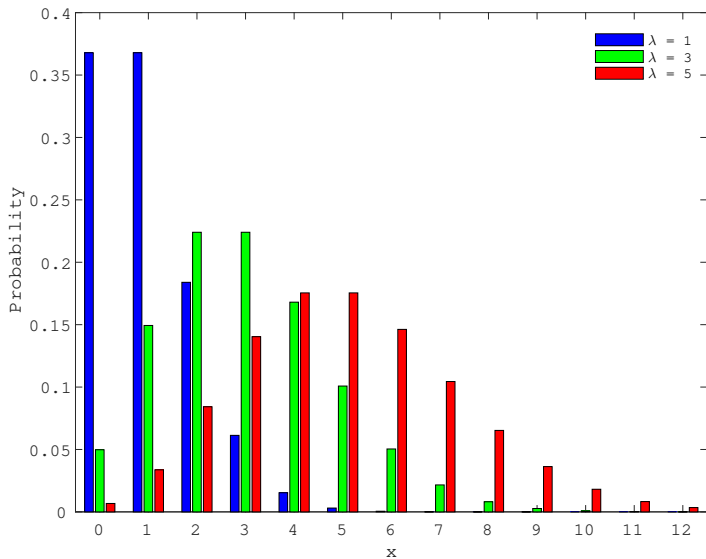
## Example: Poisson Distribution iii

Note that  $x$  in the Poisson distribution is still interpreted as the number of occurrences, but the probability  $p$  is now extremely small.

A Poisson distribution is often used for modeling occurrence of a rare phenomenon such as

- car accidents at a crossroad
- crimes committed in a district
- arrival of customers in a short interval





**Figure 4:** The p.f. of the Poisson distribution

# Continuous Distribution

It is said that a random variable  $\mathbf{X}$  has a **continuous distribution** if there exists a non-negative function  $\mathbf{f(x)}$  such that

$$\mathbf{P(X \in A)} = \int_{\mathbf{A}} \mathbf{f(x)} d\mathbf{x},$$

where  $\mathbf{A}$  is any Borel set on  $\mathbb{R}$ . The function  $\mathbf{f(x)}$  is called the **probability density function** or p.d.f. Every p.d.f. must satisfy

1.  $\mathbf{f(x)} \geq 0$ .
2.  $\int_{-\infty}^{\infty} \mathbf{f(x)} d\mathbf{x} = 1$ .

## Example: Continuous Uniform Distribution

The p.d.f. of a **continuous uniform distribution** on the interval  $[a, b]$  is

$$f(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b; \\ 0 & \text{otherwise.} \end{cases}$$

In particular, when  $a = 0$  and  $b = 1$ ,

$$f(x|0, 1) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1; \\ 0 & \text{otherwise.} \end{cases}$$

The probability that the value of  $\mathbf{X}$  is within an interval  $[c, d]$  is given by

$$P(c \leq \mathbf{X} \leq d) = \frac{\min\{b, d\} - \max\{a, c\}}{b - a}.$$

## Example: Exponential Distribution

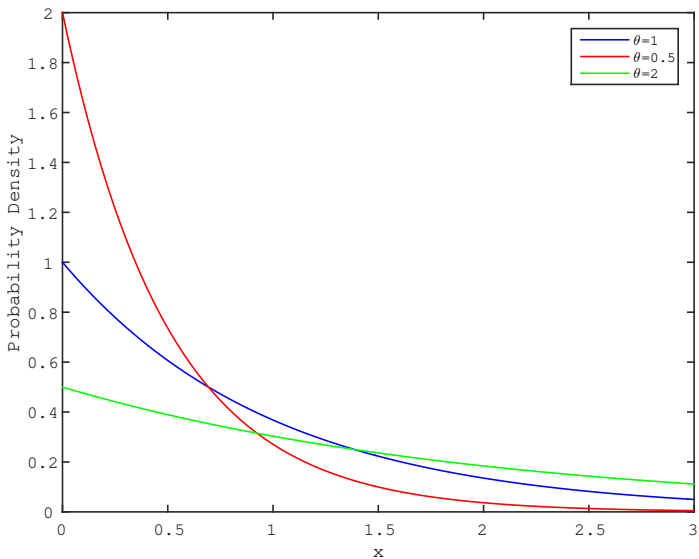
The p.d.f. of an **exponential distribution** with parameter  $\theta$  ( $\theta > 0$ ) is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{for } x > 0; \\ 0 & \text{for } x \leq 0. \end{cases}$$

The probability that the value of  $\mathbf{X}$  is within an interval  $[s, t]$  is given by

$$\begin{aligned} P(s \leq \mathbf{X} \leq t) &= \int_s^t \frac{1}{\theta} e^{-\frac{x}{\theta}} dx = -e^{-\frac{x}{\theta}} \Big|_s^t \\ &= e^{-\frac{s}{\theta}} - e^{-\frac{t}{\theta}}. \end{aligned}$$

An exponential distribution is used for longevity or duration.



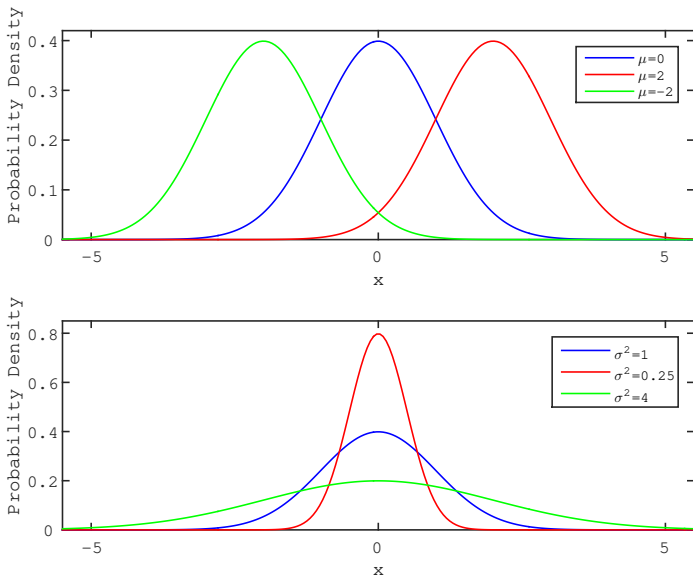
**Figure 5:** The p.d.f. of the exponential distribution

## Example: Normal Distribution

The p.d.f. of a normal (Gaussian) distribution is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right].$$

The p.d.f. of the normal distribution is bell-shaped and symmetric around the center.

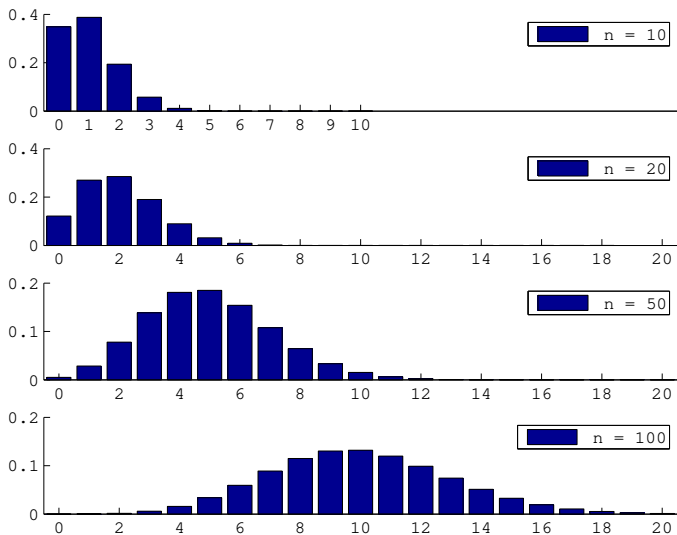


**Figure 6:** The p.d.f. of the normal distribution

# Usage of the Normal Distribution

1. The normal distribution is the mainstay of statistics and econometrics.
2. Many economic data are supposed to have a normal distribution, though it is not the case for some data (e.g., financial data).
3. Many sophisticated statistical/econometric models are built upon the normal distribution.
4. The normal distribution is often a limit of the other distribution.
5. The central limit theorem





**Figure 7:** The binomial distribution converges to the normal distribution

# Cumulative Distribution Function

The **cumulative distribution function** or c.d.f. of a random variable  $\mathbf{X}$  is a function defined for each real number  $x$  as

$$F(x) = P(\mathbf{X} \leq x), \quad -\infty < x < \infty.$$

Any c.d.f. of a random variable  $\mathbf{X}$  must satisfy

- $F(x)$  is a non-decreasing function, i.e., if  $x_1 < x_2$ , then  $F(x_1) \leq F(x_2)$ .
- $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ .
- $F(x)$  is continuous from the right, i.e.,  $\lim_{\epsilon \rightarrow 0, \epsilon > 0} F(x + \epsilon) = F(x)$ .

# Properties

1.  $P(X > x) = 1 - F(x)$ .
2.  $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$ .
3.  $P(X < x) = \lim_{\epsilon \rightarrow 0, \epsilon > 0} F(x - \epsilon)$ .
4.  $P(X = x) = \lim_{\epsilon \rightarrow 0, \epsilon > 0} F(x + \epsilon) - \lim_{\epsilon \rightarrow 0, \epsilon > 0} F(x - \epsilon)$ .
5.  $P(X = x) = 0$  if  $F(x)$  is continuous at  $x$ . This is always the case when  $X$  is a continuous random variable.
6. If  $X$  is a continuous random variable,

$$F(x) = \int_{-\infty}^x f(t) dt.$$

7. If  $F(x)$  is differentiable,  $\nabla_x F(x) = f(x)$ .

## Example: Continuous Distributions

1. Uniform distribution

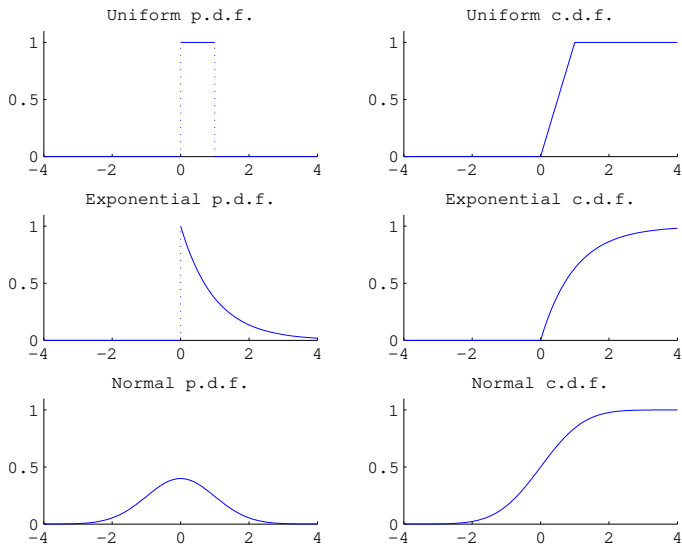
$$F(x) = \begin{cases} 0, & (x < a); \\ \frac{x-a}{b-a}, & (a \leq x \leq b); \\ 1, & (x > b). \end{cases}$$

2. Exponential distribution

$$F(x) = \begin{cases} 0, & (x < 0); \\ 1 - e^{-x/\theta}, & (x \geq 0). \end{cases}$$

3. Normal distribution

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(t - \mu)^2}{2\sigma^2} \right] dt.$$



**Figure 8:** The c.d.f. of continuous distributions

# Functions of a Random Variable

Suppose a random variable  $\mathbf{X}$  is transformed into a new random variable  $\mathbf{Y}$  with  $\mathbf{Y} = \mathbf{r}(\mathbf{X})$ . Then the probability distribution of  $\mathbf{Y}$  is given as follows.

1. If  $\mathbf{Y}$  is discrete, the p.f. of  $\mathbf{Y}$  is given by

$$g(y) = P(\mathbf{Y} = y) = P(\mathbf{r}(\mathbf{X}) = y) = \sum_{x: r(x)=y} f(x).$$

2. If  $\mathbf{Y}$  is continuous, the c.d.f. of  $\mathbf{Y}$  is given by

$$G(y) = P(\mathbf{Y} \leq y) = P(\mathbf{r}(\mathbf{X}) \leq y) = \int_{\{x: r(x) \leq y\}} f(x) dx,$$

and the p.d.f. of  $\mathbf{Y}$  is  $g(y) = \nabla_y G(y)$ .

## Example: Uniform Distribution

Suppose  $\mathbf{X}$  has a uniform distribution

$$f(x) = \begin{cases} \frac{1}{2}, & (-1 < x < 1); \\ 0, & (x \leq -1 \text{ or } x \geq 1). \end{cases}$$

and let us define  $\mathbf{Y} = \mathbf{X}^2$ . Since  $-1 < \mathbf{X} < 1$ ,  $0 \leq \mathbf{Y} < 1$ .

Then

$$\begin{aligned} G(y) &= P(\mathbf{Y} \leq y) = P(\mathbf{X}^2 \leq y) = P(-\sqrt{y} \leq \mathbf{X} \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} f(x) dx = \left. \frac{x}{2} \right|_{-\sqrt{y}}^{\sqrt{y}} = \frac{\sqrt{y}}{2} - \frac{-\sqrt{y}}{2} = \sqrt{y}. \end{aligned}$$

The p.d.f. of  $\mathbf{Y}$  is

$$g(y) = \nabla_y G(y) = \frac{1}{2\sqrt{y}}.$$

## Change of Variables Formula i

Suppose  $r(x)$  is a differentiable and strictly increasing function and let  $x = s(y)$  denote the inverse of  $y = r(x)$ . Then, as  $X$  varies over the interval  $(a, b)$ ,  $Y = r(X)$  will vary over  $(r(a), r(b))$ . Since  $r(x)$  is strictly increasing, each  $a$  in the interval  $(a, b)$  is uniquely matched with a value in  $(r(a), r(b))$  and  $y = r(x)$  if and only if  $x = s(y)$ . Thus, for  $r(a) < y < r(b)$ ,

$$G(y) = P(r(X) \leq y) = P(X \leq s(y)) = F(s(y)).$$

By applying the chain rule, the p.d.f. of  $Y$  is obtained as

$$g(y) = \nabla_y G(y) = \nabla_x F(s(y)) \nabla_y s(y) = f(s(y)) \nabla_y s(y).$$



## Change of Variables Formula ii

On the other hand, if  $r(x)$  is strictly decreasing, we have

$$G(y) = P(r(X) \leq y) = P(X \geq s(y)) = 1 - F(s(y)),$$

for  $r(b) < y < r(a)$ . Thus

$$g(y) = \nabla_y G(y) = -\nabla_x F(s(y)) \nabla_y s(y) = -f(s(y)) \nabla_y s(y).$$

In sum, either  $r(x)$  is strictly increasing or decreasing, we have

$$g(y) = \begin{cases} f(s(y)) |\nabla_y s(y)| & \text{for } \{y : y = r(x), a < x < b\}; \\ 0 & \text{otherwise.} \end{cases}$$

## Example: Uniform Distribution

Let us define  $\mathbf{Y} = \mathbf{X}^2$  where  $\mathbf{X}$  has a uniform distribution

$$f(x) = \begin{cases} \frac{1}{2}, & (-1 < x < 1); \\ 0, & (x \leq -1 \text{ or } x \geq 1). \end{cases}$$

The inverse of  $r(x) = x^2$  is

$$s(y) = \begin{cases} \sqrt{y}, & (x \geq 0); \\ -\sqrt{y}, & (x < 0). \end{cases}$$

By applying the change of variables formula, we obtain the p.d.f. of  $\mathbf{Y}$  as

$$g(y) = f(s(y)) |\nabla_y s(y)| = \frac{1}{2\sqrt{y}}.$$

# Probability Integral Transformation i

Let  $F(\mathbf{x})$  denote the c.d.f. of a continuous random variable  $\mathbf{X}$ . Now we define a new random variable  $\mathbf{Y} = F(\mathbf{X})$ , which is called **probability integral transformation**. Since  $F(\mathbf{x})$  is the c.d.f.,  $0 < y < 1$ . Furthermore,

$$\begin{aligned} G(y) &= P(\mathbf{Y} \leq y) = P(F(\mathbf{X}) \leq y) \\ &= P(\mathbf{X} \leq F^{-1}(y)) = F(F^{-1}(y)) = y, \end{aligned}$$

where  $\mathbf{x} = F^{-1}(y)$  is the inverse of  $y = F(\mathbf{x})$ . Obviously  $G(y)$  is the c.d.f. of a continuous uniform distribution over the interval  $[0, 1]$ .

## Probability Integral Transformation ii

Conversely, suppose  $\mathbf{Y}$  has a continuous uniform distribution over the interval  $[0, 1]$  and consider a new random variable  $\mathbf{Z} = \mathbf{F}^{-1}(\mathbf{Y})$ . Then

$$\begin{aligned} H(z) &= P(\mathbf{Z} \leq z) = P(\mathbf{F}^{-1}(\mathbf{Y}) \leq z) \\ &= P(\mathbf{Y} \leq \mathbf{F}(z)) = \mathbf{F}(z). \end{aligned}$$

Thus  $\mathbf{Z}$  is equivalent to  $\mathbf{X}$ .

The above result suggests that we can produce random numbers from an arbitrary probability distribution with the c.d.f.  $\mathbf{F}(\mathbf{x})$  once we obtain uniform random numbers in  $[0, 1]$ .

## Example: Exponential Distribution

The c.d.f. of an exponential distribution is

$$F(x) = \begin{cases} 0, & (x < 0); \\ 1 - e^{-x/\theta}, & (x \geq 0). \end{cases}$$

Hence the inverse of  $y = F(x)$  is

$$x = F^{-1}(y) = -\theta \log(1 - y),$$

for  $0 \leq y \leq 1$ .

# Expectation

The **expectation** or **expected value** of a random variable  $\mathbf{X}$  is defined as

## Definition: Expectation of a Random Variable

$$\mathbf{E}[\mathbf{X}] = \begin{cases} \sum_x x f(x) & \text{for discrete random variables;} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{for continuous random variables.} \end{cases}$$

The expectation of a random variable  $\mathbf{E}[\mathbf{X}]$  is often referred to as the mean of the distribution. This is due to the fact that in the discrete case  $\mathbf{E}[\mathbf{X}]$  is a weighted average of all possible values that  $\mathbf{X}$  would take.

**Table 5:** Expectation of Selected Distributions

Distribution	p.f./p.d.f	$E[X]$
Bernoulli	$p^x(1-p)^{1-x}$	$p$
Binomial	$\binom{n}{x} p^x(1-p)^{n-x}$	$np$
Neg. Binomial	$\binom{x+r-1}{x} p^r(1-p)^x$	$r \frac{1-p}{p}$
Poisson	$\frac{e^{-\lambda} \lambda^x}{x!}$	$\lambda$
Uniform	$\frac{1}{b-a}$	$\frac{a+b}{2}$
Exponential	$\frac{1}{\theta} e^{-\frac{x}{\theta}}$	$\theta$
Normal	$\frac{1}{\sqrt{2\pi}\sigma^2} \exp \left[ -\frac{(x-\mu)^2}{2\sigma^2} \right]$	$\mu$

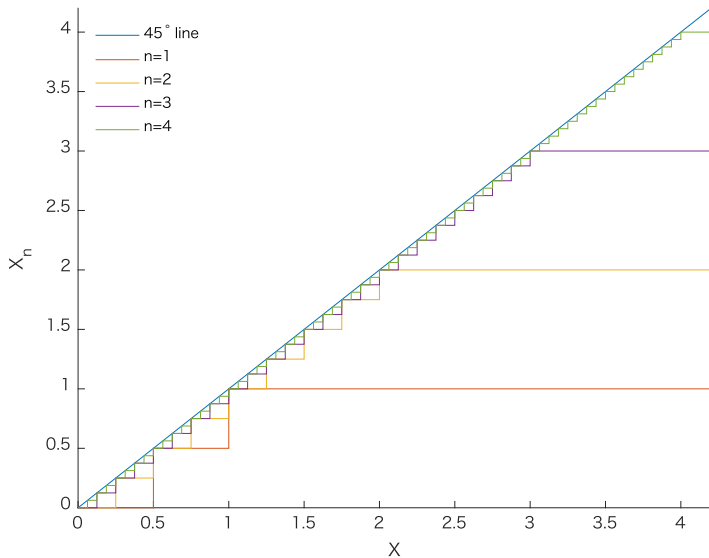
# Discretization of a Random Variable

Suppose  $\mathbf{X}$  is a continuous non-negative random variables and define a discrete non-negative random variable  $\mathbf{X}_n$  as

$$\mathbf{X}_n = \begin{cases} 0, & (0 \leq \mathbf{X} < \frac{1}{2^n}); \\ \frac{1}{2^n}, & (\frac{1}{2^n} \leq \mathbf{X} < \frac{2}{2^n}); \\ \vdots & \\ \frac{i-1}{2^n}, & (\frac{i-1}{2^n} \leq \mathbf{X} < \frac{i}{2^n}); \\ \vdots & \\ \frac{2^n n - 1}{2^n}, & (\frac{2^n n - 1}{2^n} \leq \mathbf{X} < n); \\ n, & (\mathbf{X} \geq n). \end{cases}$$

Note that (i)  $\mathbf{X}_n \leq \mathbf{X}$  for all  $n$ , and (ii)  $\lim_{n \rightarrow \infty} \mathbf{X}_n = \mathbf{X}$ .





**Figure 9:** Discretized Random Variable

# Formal Definition of Expectation i

Then the expectation of  $X_n$  is given by

$$\begin{aligned} E[X_n] &= \sum_{i=1}^{2^n} \frac{i-1}{2^n} P\left(X_n = \frac{i-1}{2^n}\right) + nP(X_n = n) \\ &= \sum_{i=1}^{2^n} \frac{i-1}{2^n} P\left(\frac{i-1}{2^n} \leq X < \frac{i}{2^n}\right) + nP(X \geq n) \\ &= \sum_{i=1}^{2^n} \frac{i-1}{2^n} \left\{ F\left(\frac{i}{2^n}\right) - F\left(\frac{i-1}{2^n}\right) \right\} + n\{1 - F(n)\} \\ &= \sum_{x=0, \dots, n-\epsilon_n} x \{F(x + \epsilon_n) - F(x)\} + n\{1 - F(n)\}, \end{aligned}$$

## Formal Definition of Expectation ii

where  $F(x)$  is the c.d.f. of  $X$  and  $\epsilon_n = \frac{1}{2^n}$ .

If the limit exists, it is expressed as

$$\begin{aligned}\lim_{n \rightarrow \infty} E[X_n] &= \lim_{n \rightarrow \infty} \overbrace{\sum_{x=0, \dots, n-\epsilon_n}^{\int_0^n} x \overbrace{\{F(x+\epsilon_n) - F(x)\}}^{dF(x)}} \\ &\quad + \overbrace{\lim_{n \rightarrow \infty} n \{1 - F(n)\}}^{=0} \\ &= \int_0^\infty x dF(x),\end{aligned}$$

## Formal Definition of Expectation iii

which is a type of **Lebesgue–Stieltjes integral**. If this limit exists, the expectation of  $\mathbf{X}$  is defined as

$$\mathbf{E}[\mathbf{X}] = \int_0^{\infty} x dF(x).$$

**Note 1:**  $\lim_{n \rightarrow \infty} \mathbf{E}[\mathbf{X}_n]$  is possibly non-convergent. If so,  $\mathbf{E}[\mathbf{X}]$  does not exist.

**Note 2:** For a continuous random variable, if  $\mathbf{E}[\mathbf{X}]$  exists,

$$\mathbf{E}[\mathbf{X}] = \int_0^{\infty} x dF(x) = \int_0^{\infty} x f(x) dx.$$

## Formal Definition of Expectation iv

**Note 3:**  $X$  can be decomposed as  $X = X^+ - X^-$  where

$$X^+ = \max\{X, 0\} \geq 0,$$

$$X^- = \max\{-X, 0\} \geq 0.$$

If both  $E[X^+]$  and  $E[X^-]$  exist, the expectation of a real-valued random variable  $X$  is defined as

$$E[X] = E[X^+] - E[X^-].$$

# Properties

$X, Y, X_1, \dots, X_n$ : random variables

$a, b, c, a_1, \dots, a_n$ : real numbers

1.  $E[X + c] = E[X] + c.$
2.  $E[aX] = aE[X].$
3.  $E[aX + c] = aE[X] + c.$
4.  $E[X + Y] = E[X] + E[Y].$
5.  $E[aX + bY + c] = aE[X] + bE[Y] + c.$
6.  $E[X_1 + \dots + X_n] = \sum_{i=1}^n E[X_i].$
7.  $E[a_1X_1 + \dots + a_nX_n + c] = \sum_{i=1}^n a_iE[X_i] + c.$

# Variance

The **variance** of a random variable  $X$  is defined as

## Definition: Variance of a Random Variable

$$\begin{aligned}\text{Var}[X] &= E[(X - \mu)^2] \\ &= \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{for discrete r.v.'s;} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{for continuous r.v.'s.} \end{cases}\end{aligned}$$

where  $\mu = E[X]$ .

The square root of the variance is called the **standard deviation**. The variance of a random variable is interpreted as a measurement of spread or dispersion of the distribution around the mean  $\mu$ .

**Table 6:** Variance of Selected Distributions

Distribution	p.f./p.d.f	Var[ $X$ ]
Bernoulli	$p^x(1-p)^{1-x}$	$p(1-p)$
Binomial	$\binom{n}{x} p^x(1-p)^{n-x}$	$np(1-p)$
Neg. Binomial	$\binom{x+r-1}{x} p^r(1-p)^x$	$r \frac{1-p}{p^2}$
Poisson	$\frac{e^{-\lambda} \lambda^x}{x!}$	$\lambda$
Uniform	$\frac{1}{b-a}$	$\frac{(b-a)^2}{12}$
Exponential	$\frac{1}{\theta} e^{-\frac{x}{\theta}}$	$\theta^2$
Normal	$\frac{1}{\sqrt{2\pi}\sigma^2} \exp \left[ -\frac{(x-\mu)^2}{2\sigma^2} \right]$	$\sigma^2$



# Properties

1.  $\text{Var}[X] = 0$  when  $P(X = c) = 1$  for a constant number  $c$ .
2.  $\text{Var}[X + c] = \text{Var}[X]$ .
3.  $\text{Var}[aX] = a^2 \text{Var}[X]$ .
4.  $\text{Var}[aX + c] = a^2 \text{Var}[X]$ .
5.  $\text{Var}[X] = E[X^2] - \mu^2$ .

# Skewness

The **skewness** of a random variable is defined as

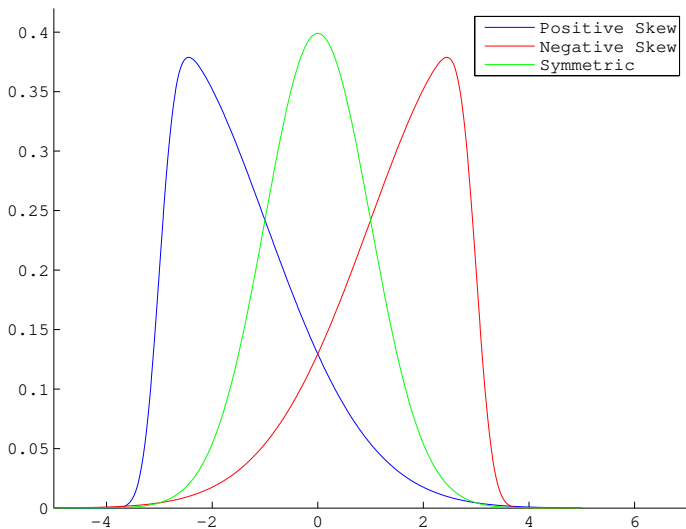
## Definition of Skewness

$$\beta_1 = \mathbb{E} \left[ \left( \frac{\mathbf{X} - \mu}{\sigma} \right)^3 \right],$$

where  $\mu = \mathbb{E}[\mathbf{X}]$  and  $\sigma^2 = \text{Var}[\mathbf{X}]$ .

The skewness  $\beta_1$  tells whether the distribution is symmetric around the mean  $\mu$  or not.

- If  $\beta_1 > 0$ , the distribution has a longer tail on the right.
- If  $\beta_1 < 0$ , the distribution has a longer tail on the left.
- If  $\beta_1 = 0$ , the distribution is symmetric around the mean  $\mu$ .



**Figure 10: Skewness**

# Kurtosis

The **kurtosis** of a random variable is defined as

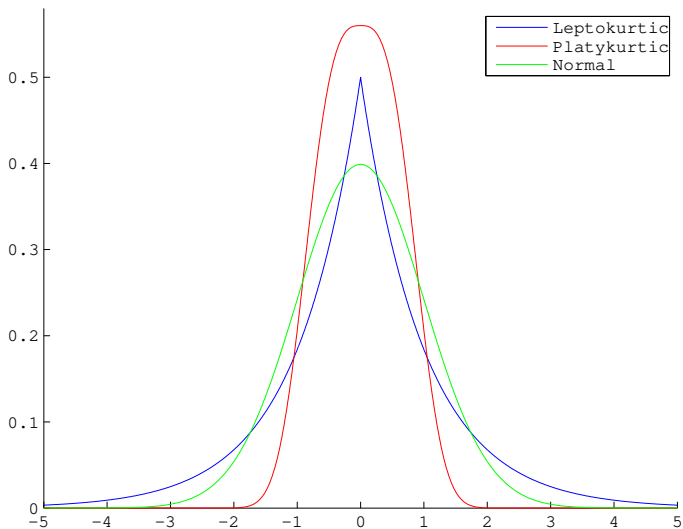
## Definition of Kurtosis

$$\beta_2 = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right],$$

where  $\mu = \mathbb{E}[X]$  and  $\sigma^2 = \text{Var}[X]$ .

The kurtosis  $\beta_2$  is a measurement of thickness/heaviness of the tail. Note that the kurtosis of the normal distribution is 3.

- If  $\beta_2 > 3$ , the distribution has a thicker tail (**leptokurtic**).
- If  $\beta_2 < 3$ , the distribution has a thinner tail (**platykurtic**).



**Figure 11: Kurtosis**

# Moment Generating Function i

For any random variable  $\mathbf{X}$  and any positive integer  $k$ , the expectation  $E[\mathbf{X}^k]$  is called the  $k$ -th **moment** of  $\mathbf{X}$ . The expectation of  $\mathbf{X}$  is the first moment of  $\mathbf{X}$ .

## Definition: Moment Generating Function

$$M_{\mathbf{X}}(t) = E[e^{t\mathbf{X}}].$$

By applying the Maclaurin series, we have

$$M_{\mathbf{X}}(t) = E[e^{t\mathbf{X}}] = E\left[\sum_{j=0}^{\infty} \frac{(t\mathbf{X})^j}{j!}\right] = \sum_{j=0}^{\infty} \frac{t^j}{j!} E[\mathbf{X}^j].$$

## Moment Generating Function ii

Since

$$\begin{aligned}\nabla_t^k M_X(t) &= \nabla_t^k \sum_{j=0}^{\infty} \frac{t^j}{j!} E[X^j] = \sum_{j=0}^{\infty} \frac{\nabla_t^k t^j}{j!} E[X^j] \\ &= \sum_{j=k}^{\infty} \frac{j(j-1)\cdots(j-k+1)t^{j-k}}{j!} E[X^j] \\ &= E[X^k] + \sum_{j=k+1}^{\infty} \frac{t^{j-k}}{(j-k)!} E[X^j],\end{aligned}$$

we have

$$\nabla_t^k M_X(0) = E[X^k].$$

Thus  $M_X(t)$  is called the **moment generating function**.

**Table 7:** Moment Generating Functions for Selected Distributions

Distribution	p.m.f or p.d.f.	m.g.f.
Bernoulli	$p^x q^{1-x} \mathbf{1}_{\{0,1\}}(x), \quad q = 1 - p$	$q + pe^t$
Binomial	$\binom{n}{x} p^x q^{n-x} \mathbf{1}_{\{0,1,\dots,n\}}(x)$	$(q + pe^t)^n$
Neg. Binomial	$\binom{x+r-1}{x} p^r q^x \mathbf{1}_{\{0,1,2,\dots\}}(x)$	$\left(\frac{p}{1-qe^t}\right)^r$
Poisson	$\frac{e^{-\lambda} \lambda^x}{x!} \mathbf{1}_{\{0,1,2,\dots\}}(x)$	$e^{\lambda(e^t-1)}$
Uniform	$\frac{1}{b-a} \mathbf{1}_{(a,b)}(x)$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$
Exponential	$\frac{1}{\theta} e^{-\frac{x}{\theta}} \mathbf{1}_{(0,\infty)}(x)$	$\frac{1}{1-t\theta}$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$e^{\mu t + \frac{\sigma^2}{2} t^2}$



# Properties

1.  $M_X(0) = 1$ .
2.  $M_Y(t) = M_X(ta)e^{tb}$  for  $Y = aX + b$ .
3. Suppose that  $\{X_i\}_{i=1}^n$  are mutually independent,  $M_{X_i}(t)$  is the m.g.f. of  $X_i$ , and  $S_n = \sum_{i=1}^n X_i$ . Then  $M_{S_n}(t) = \prod_{i=1}^n M_{X_i}(t)$ .
4. **Uniqueness of the Moment Generating Function**  
If two probability distributions have the same m.g.f., they must be the same. In other words, the m.g.f. is unique for every probability distribution.

# Bivariate Distribution i

In many cases, we need to consider the properties of two or more random variables simultaneously. The joint probability distribution of two random variables is called a **bivariate distribution**. For a pair of two discrete random variables  $(X, Y)$ , the **joint probability function** or joint p.f. is defined as

$$f(x, y) = P(X = x \text{ and } Y = y).$$

If the sequence  $\{(x_i, y_j)\}_{i,j=1}^{\infty}$  includes all possible values of  $(X, Y)$ ,

$$\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} f(x_i, y_j) = 1.$$

## Bivariate Distribution ii

**Table 8:** Example of Discrete Bivariate Distribution

X	Y				
		1	2	3	4
1		0.1	0	0.1	0
2		0.3	0	0.1	0.2
3		0	0.2	0	0

$$P(X = 1 \text{ and } Y = 3) = 0.1$$

$$P(X + Y = 5) = 0 + 0.1 + 0.2 = 0.3$$

$$P(X = 2) = 0.3 + 0 + 0.1 + 0.2 = 0.6$$

$$P(X \leq 2 \text{ and } Y \geq 3) = 0.1 + 0 + 0.1 + 0.2 = 0.4$$

# Marginal Distribution

Given the joint p.f.  $f(\mathbf{x}, \mathbf{y})$ , the **marginal probability function** or marginal p.f. of  $\mathbf{X}$  is defined as

$$f_X(x) = P(X = x) = \sum_y f(x, y).$$

In the same manner, the marginal p.f. of  $\mathbf{Y}$  is defined as

$$f_Y(y) = P(Y = y) = \sum_x f(x, y).$$

**Table 9:** Joint and Marginal Distributions

Y						
X		1	2	3	4	
1		0.1	0	0.1	0	0.2
2		0.3	0	0.1	0.2	0.6
3		0	0.2	0	0	0.2
		0.4	0.2	0.2	0.2	

$$E[X] = 0.2 \times 1 + 0.6 \times 2 + 0.2 \times 3 = 2$$

$$E[Y] = 0.4 \times 1 + 0.2 \times 2 + 0.2 \times 3 + 0.2 \times 4 = 2.2$$

$$\text{Var}[X] = 0.2 \times (1 - 2)^2 + 0.6 \times (2 - 2)^2 + 0.2 \times (3 - 2)^2 = 0.4$$

$$\begin{aligned}\text{Var}[Y] &= 0.4 \times (1 - 2.2)^2 + 0.2 \times (2 - 2.2)^2 \\ &\quad + 0.2 \times (3 - 2.2)^2 + 0.2 \times (4 - 2.2)^2 = 1.36\end{aligned}$$

# Conditional Distribution

The conditional probability of  $\mathbf{X} = \mathbf{x}$  given  $\mathbf{Y} = \mathbf{y}$  is

$$P(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}) = \frac{P(\mathbf{X} = \mathbf{x} \text{ and } \mathbf{Y} = \mathbf{y})}{P(\mathbf{Y} = \mathbf{y})} = \frac{f(\mathbf{x}, \mathbf{y})}{f_Y(\mathbf{y})} = f(\mathbf{x} | \mathbf{y}).$$

X	Y							
	1	$f(\mathbf{x} 1)$	2	$f(\mathbf{x} 2)$	3	$f(\mathbf{x} 3)$	4	$f(\mathbf{x} 4)$
1	0.1	1/4	0	0	0.1	1/2	0	0
2	0.3	3/4	0	0	0.1	1/2	0.2	1
3	0	0	0.2	1	0	0	0	0
	0.4	1	0.2	1	0.2	1	0.2	1

# Continuous Bivariate Distribution

A pair of continuous random variables  $(X, Y)$  has a bivariate distribution if the probability that a point  $(X, Y)$  is located in a region  $A$  is given by

$$P((X, Y) \in A) = \int_A \int f(x, y) dx dy.$$

where the function  $f$  is called a joint probability density function or joint p.d.f.

The joint p.d.f. should satisfy

1.  $f(x, y) \geq 0$ .
2.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ .

# Marginal and Conditional Distributions

Given the joint p.d.f.  $f(x, y)$ , the **marginal p.d.f.** of  $X$  and  $Y$  are defined as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

The **conditional p.d.f.** of  $X$  given  $Y = y$  is

$$f_X(x|y) = \begin{cases} \frac{f(x, y)}{f_Y(y)} & \text{if } f_Y(y) > 0; \\ 0 & \text{otherwise.} \end{cases}$$

**Bayes' theorem** for a bivariate distribution is given by

$$f_X(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_Y(y|x)}{f_Y(y)} f_X(x).$$



## Example: Bivariate Normal Distribution

The joint p.d.f. of a bivariate normal distribution is

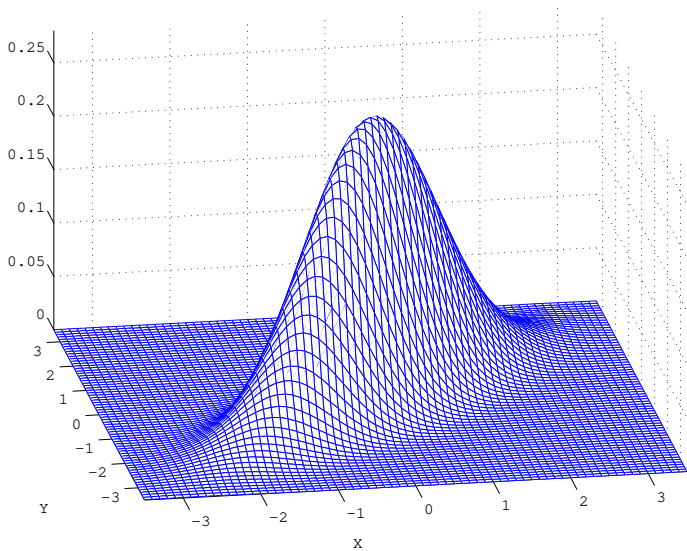
$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp \left[ -\frac{1}{2(1-\rho^2)} \left\{ \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right\} \right].$$

The marginal distribution of  $\mathbf{X}$  is

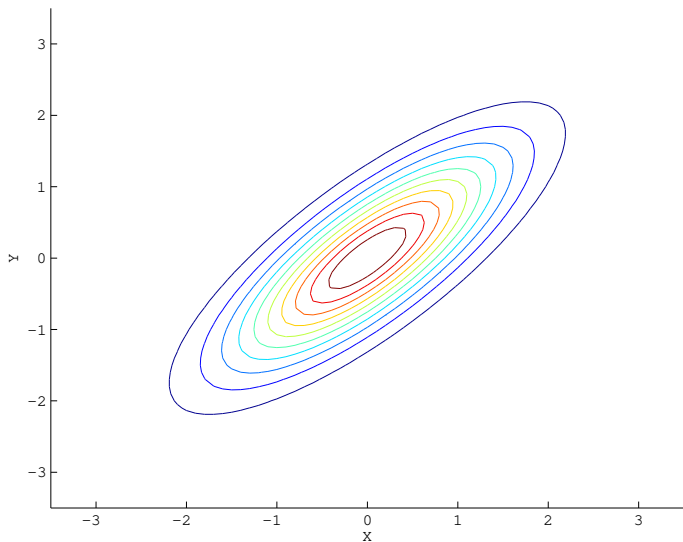
$$\mathbf{X} \sim \mathcal{N}(\mu_X, \sigma_X^2).$$

The conditional distribution of  $\mathbf{X}$  given  $\mathbf{Y} = \mathbf{y}$  is

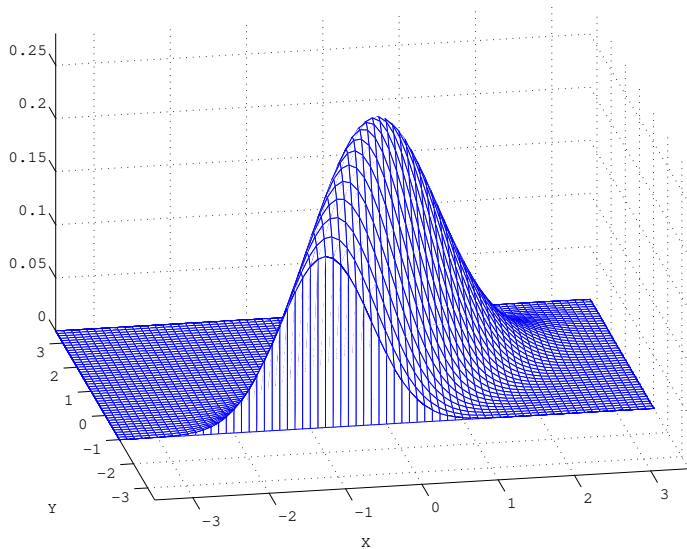
$$\mathbf{X}|\mathbf{Y} = \mathbf{y} \sim \mathcal{N} \left( \mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y), \sigma_X^2(1 - \rho^2) \right).$$



**Figure 12:** Bivariate Normal Distribution (Surface Plot)



**Figure 13:** Bivariate Normal Distribution (Contour Plot) 115/144



**Figure 14:** Bivariate Normal Distribution (Conditional Distribution)

# Covariance, Correlation, and Independence

The **covariance** of two random variables  $\mathbf{X}$  and  $\mathbf{Y}$  is

$$\text{Cov}[\mathbf{X}, \mathbf{Y}] = \text{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})], \quad \mu_{\mathbf{X}} = \text{E}[\mathbf{X}], \quad \mu_{\mathbf{Y}} = \text{E}[\mathbf{Y}].$$

The **correlation (coefficient)** of  $\mathbf{X}$  and  $\mathbf{Y}$  is

$$\rho_{\mathbf{X}\mathbf{Y}} = \frac{\text{Cov}[\mathbf{X}, \mathbf{Y}]}{\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}}, \quad \sigma_{\mathbf{X}}^2 = \text{Var}[\mathbf{X}], \quad \sigma_{\mathbf{Y}}^2 = \text{Var}[\mathbf{Y}].$$

$\mathbf{X}$  and  $\mathbf{Y}$  are **mutually independent** if and only if

$$f(x, y) = f_{\mathbf{X}}(x)f_{\mathbf{Y}}(y), \Leftrightarrow f_{\mathbf{X}}(x|y) = f_{\mathbf{X}}(x), \Leftrightarrow f_{\mathbf{Y}}(y|x) = f_{\mathbf{Y}}(y).$$

# Properties

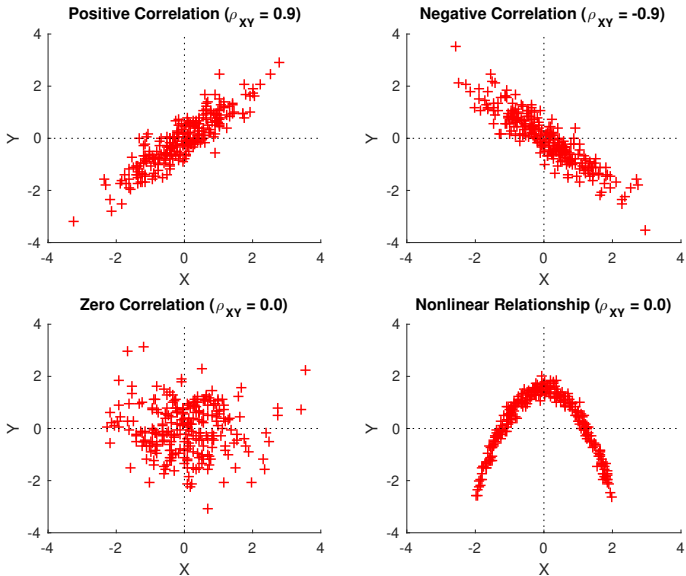
1.  $|\rho_{XY}| \leq 1$ .
2.  $\text{Cov}[X, Y] = E[XY] - E[X]E[Y]$ .
3. If  $X$  and  $Y$  are independent,  $\text{Cov}[X, Y] = 0$  and  $\rho_{XY} = 0$ .
4.  $\text{Cov}[X, Y] = 0$  does not imply that  $X$  and  $Y$  are independent.
5.  $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$ .
6.  $\text{Var}[aX + bY + c] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y]$ .

## Example: Discrete Bivariate Distribution

X	Y					
		1	2	3	4	
1		0.1	0	0.1	0	0.2
2		0.3	0	0.1	0.2	0.6
3		0	0.2	0	0	0.2
		0.4	0.2	0.2	0.2	

$$\begin{aligned}E[XY] &= 0.1 \times 1 \times 1 + 0.1 \times 1 \times 3 + 0.3 \times 2 \times 1 \\&\quad + 0.1 \times 2 \times 3 + 0.2 \times 2 \times 4 + 0.2 \times 3 \times 2 \\&= 0.1 + 0.3 + 0.6 + 0.6 + 1.6 + 1.2 = 4.4\end{aligned}$$

$$\text{Cov}[X, Y] = E[XY] - E[X]E[Y] = 4.4 - 2 \times 2.2 = 0$$



**Figure 15:** Scatter Plots for Illustration of Correlation



# Conditional Expectation

The **conditional expectation** of  $\mathbf{X}$  given  $\mathbf{Y} = \mathbf{y}$  is defined as

$$\mathbf{E}[\mathbf{X} | \mathbf{Y} = \mathbf{y}] = \begin{cases} \sum_x x f_{\mathbf{X}}(x | \mathbf{y}) & \text{for discrete r.v.'s;} \\ \int_{-\infty}^{\infty} x f_{\mathbf{X}}(x | \mathbf{y}) dx & \text{for continuous r.v.'s.} \end{cases}$$

where  $f(\mathbf{x} | \mathbf{y})$  is the conditional p.f. or p.d.f. of  $\mathbf{X}$  given  $\mathbf{Y} = \mathbf{y}$ .

Since  $f_{\mathbf{X}}(\mathbf{x} | \mathbf{y}) = \frac{f_{\mathbf{Y}}(\mathbf{y} | \mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y})} f_{\mathbf{X}}(\mathbf{x}) = m(\mathbf{y} | \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})$ , we have

$$\begin{aligned} \mathbf{E}[\mathbf{X} | \mathbf{Y} = \mathbf{y}] &= \mathbf{E}[\mathbf{X} m(\mathbf{y} | \mathbf{X})] \\ &= \begin{cases} \sum_x x m(\mathbf{y} | \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) & \text{for discrete r.v.'s;} \\ \int_{-\infty}^{\infty} x m(\mathbf{y} | \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) dx & \text{for continuous r.v.'s.} \end{cases} \end{aligned}$$

# Conditional Expectation as a Random Variable

If we do not fix the value of  $\mathbf{Y}$  at any specific value, the conditional expectation  $\mathbf{E}[\mathbf{X}|\mathbf{Y}]$  is regarded as a function of  $\mathbf{Y}$ . Thus  $\mathbf{E}[\mathbf{X}|\mathbf{Y}]$  is also a random variable.

The **conditional variance** of  $\mathbf{X}$  given  $\mathbf{Y}$  is defined as

$$\text{Var}[\mathbf{X}|\mathbf{Y}] = \mathbf{E}[(\mathbf{X} - \mathbf{E}[\mathbf{X}|\mathbf{Y}])^2 | \mathbf{Y}].$$

## Properties

1.  $\mathbf{E}[\mathbf{X}] = \mathbf{E}[\mathbf{E}[\mathbf{X}|\mathbf{Y}]]$ .
2.  $\text{Var}[\mathbf{X}] = \mathbf{E}[\text{Var}[\mathbf{X}|\mathbf{Y}]] + \text{Var}[\mathbf{E}[\mathbf{X}|\mathbf{Y}]]$ .

# Multivariate Distribution

A joint p.f. of  $n$  discrete random variables  $(\mathbf{X}_1, \dots, \mathbf{X}_n)$  is defined as

$$f(x_1, \dots, x_n) = P(\mathbf{X}_1 = x_1, \dots, \mathbf{X}_n = x_n).$$

A joint p.d.f of continuous  $(\mathbf{X}_1, \dots, \mathbf{X}_n)$  is

$$P(\mathbf{X}_1 \in \mathbf{A}_1, \dots, \mathbf{X}_n \in \mathbf{A}_n) = \int_{\mathbf{A}_1} \cdots \int_{\mathbf{A}_n} f(x_1, \dots, x_n) dx_1 \cdots dx_n.$$

A marginal p.d.f. of  $\mathbf{X}_1$  is

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_2 \cdots dx_n.$$

A conditional p.d.f. of  $(\mathbf{X}_1, \dots, \mathbf{X}_m)$  given  $\mathbf{X}_{m+1}, \dots, \mathbf{X}_n$  is

$$f(x_1, \dots, x_m | x_{m+1}, \dots, x_n) = \frac{f(x_1, \dots, x_m, x_{m+1}, \dots, x_n)}{f(x_{m+1}, \dots, x_n)}.$$

# Independence

If random variables  $(X_1, \dots, X_n)$  are mutually independent,

**Discrete random variables:**

$$\begin{aligned}f(x_1, \dots, x_n) &= P(X_1 = x_1, \dots, X_n = x_n) \\&= P(X_1 = x_1) \times \dots \times P(X_n = x_n) \\&= f_1(x_1) \times \dots \times f_n(x_n).\end{aligned}$$

**Continuous random variables:**

$$\begin{aligned}f(x_1, \dots, x_n) &= f_1(x_1) \times \dots \times f_n(x_n) \\P(X_1 \in A_1, \dots, X_n \in A_n) \\&= \int_{A_1} f_1(x_1) dx_1 \times \dots \times \int_{A_n} f_n(x_n) dx_n.\end{aligned}$$

# Properties

Suppose  $\mathbf{X}_1, \dots, \mathbf{X}_n$  are mutually independent.

1.  $f(\mathbf{x}_i | \mathbf{x}_j) = f(\mathbf{x}_i), (i \neq j).$
2.  $E[\mathbf{X}_i \mathbf{X}_j] = E[\mathbf{X}_i]E[\mathbf{X}_j], (i \neq j).$
3.  $E[\mathbf{X}_1 \times \dots \times \mathbf{X}_n] = E[\mathbf{X}_1] \times \dots \times E[\mathbf{X}_n].$
4.  $\text{Var}[\mathbf{X}_i + \mathbf{X}_j] = \text{Var}[\mathbf{X}_i] + \text{Var}[\mathbf{X}_j], (i \neq j).$
5.  $\text{Var}[\mathbf{X}_1 + \dots + \mathbf{X}_n] = \text{Var}[\mathbf{X}_1] + \dots + \text{Var}[\mathbf{X}_n].$
6.  $\text{Var}[a_1 \mathbf{X}_1 + \dots + a_n \mathbf{X}_n + b] = a_1^2 \text{Var}[\mathbf{X}_1] + \dots + a_n^2 \text{Var}[\mathbf{X}_n].$

# Variance-Covariance Matrix

$$\Sigma = \begin{bmatrix} \text{Var}[\mathbf{X}_1] & \text{Cov}[\mathbf{X}_1, \mathbf{X}_2] & \cdots & \text{Cov}[\mathbf{X}_1, \mathbf{X}_n] \\ \text{Cov}[\mathbf{X}_2, \mathbf{X}_1] & \text{Var}[\mathbf{X}_2] & \cdots & \text{Cov}[\mathbf{X}_2, \mathbf{X}_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[\mathbf{X}_n, \mathbf{X}_1] & \text{Cov}[\mathbf{X}_n, \mathbf{X}_2] & \cdots & \text{Var}[\mathbf{X}_n] \end{bmatrix},$$

is called the (variance-)covariance matrix.

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix}, \quad \rho_{ij} = \frac{\text{Cov}[\mathbf{X}_i, \mathbf{X}_j]}{\sqrt{\text{Var}[\mathbf{X}_i]\text{Var}[\mathbf{X}_j]}},$$

is called the correlation matrix.

# Properties

1. In the case of a bivariate distribution,

$$\begin{aligned}\text{Var}[aX + bY] &= a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y] \\ &= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \text{Var}[X] & \text{Cov}[X, Y] \\ \text{Cov}[X, Y] & \text{Var}[Y] \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}.\end{aligned}$$

2. In general,  $\text{Var}[\sum_{i=1}^n a_i X_i] = a' \Sigma a$ ,  $a = [a_1 \ \cdots \ a_n]'$ .
3.  $\Sigma = SRS$  or  $R = S^{-1}\Sigma S^{-1}$  where

$$S = \text{diag}\{\sigma_1, \dots, \sigma_n\} = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix}.$$

# Multivariate Normal Distribution

The joint p.d.f. of the **multivariate normal distribution** is given by

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n) \\ = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu) \right].$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \mathbf{E}[\mathbf{X}_1] \\ \vdots \\ \mathbf{E}[\mathbf{X}_n] \end{bmatrix},$$

and  $\Sigma$  is the covariance matrix.



# Markov Inequality

Suppose that a random variable is almost surely non-negative, i.e.,  $P(\mathbf{X} \geq 0) = 1$ . Then for any given  $t > 0$ , we have

$$P(\mathbf{X} \geq t) \leq \frac{E[\mathbf{X}]}{t}.$$

**Proof.** Without loss of generality, we suppose  $\mathbf{X}$  is continuous.

$$\begin{aligned} E[\mathbf{X}] &= \int_0^{\infty} xf(x)dx = \int_0^t xf(x)dx + \int_t^{\infty} xf(x)dx \\ &\geq \int_t^{\infty} xf(x)dx \geq \int_t^{\infty} tf(x)dx = tP(\mathbf{X} \geq t). \end{aligned}$$



# Chebyshev Inequality

Suppose that  $\mathbf{X}$  is a random variable for which  $\text{Var}[\mathbf{X}]$  exists. Then for any given  $t > 0$ , we have

$$\mathbf{P}(|\mathbf{X} - \mu| \geq t) \leq \frac{\text{Var}[\mathbf{X}]}{t^2}.$$

**Proof.** Let  $\mathbf{Y} = (\mathbf{X} - \mu)^2$ . Then  $\mathbf{P}(\mathbf{Y} \geq 0) = 1$  and

$\mathbf{E}[\mathbf{Y}] = \text{Var}[\mathbf{X}]$ . Thus we have the result by applying the Markov inequality to  $\mathbf{Y}$ , i.e.,

$$\mathbf{P}(|\mathbf{X} - \mu| \geq t) = \mathbf{P}(\mathbf{Y} \geq t^2) \leq \frac{\mathbf{E}[\mathbf{Y}]}{t^2} = \frac{\text{Var}[\mathbf{X}]}{t^2}.$$

# Properties of the Sample Mean

Suppose that we have data  $\mathbf{X}_1, \dots, \mathbf{X}_n$  which are drawn from the same population independently. This implies that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  follows the same probability distribution and are mutually independent. Further suppose that the expectation of the distribution is  $\mu$  and the variance is  $\sigma^2$ . Then we define the sample mean as

$$\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i.$$

The expectation and the variance of  $\bar{\mathbf{X}}_n$  are

$$\mathbb{E}[\bar{\mathbf{X}}_n] = \mu, \quad \text{Var}[\bar{\mathbf{X}}_n] = \frac{\sigma^2}{n}.$$

By apply the Chebyshev inequality to the sample mean, we have

$$\mathbf{P}(|\bar{\mathbf{X}}_n - \mu| \geq t) \leq \frac{\sigma^2}{nt^2}.$$

In particular, when  $t = 3\sigma/\sqrt{n}$ , we have

$$\mathbf{P}\left(|\bar{\mathbf{X}}_n - \mu| \geq 3\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{9} \approx 0.11$$

Note that if  $\mathbf{X}_i \sim \mathcal{N}(\mu, \sigma^2)$

$$\mathbf{P}\left(|\bar{\mathbf{X}}_n - \mu| \geq 3\frac{\sigma}{\sqrt{n}}\right) \approx 0.0027$$

and

$$\mathbf{P}\left(|\bar{\mathbf{X}}_n - \mu| \geq 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.05$$

## Example: Bernoulli Distribution

Suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  independently follow the Bernoulli distribution with the probability of success  $p$ . Since  $\mathbf{E}[\mathbf{X}_i] = p$  and  $\text{Var}[\mathbf{X}_i] = p(1 - p)$ ,  $i = 1, \dots, n$ ,

$$\mathbf{P} \left( |\bar{\mathbf{X}}_n - p| \geq 3 \sqrt{\frac{p(1 - p)}{n}} \right) \leq \frac{1}{9}.$$

Then  $3\sqrt{p(1 - p)/n}$  is interpreted as a margin of error.

$n$	Margin of Error ( $p = 0.5$ )
100	0.1500
10,000	0.0150
1,000,000	0.0015

# Law of Large Numbers

## Convergence in Probability

Suppose that  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots$  is a sequence of random variable. This sequence converges in probability to a real number  $\mathbf{a}$  if for any given number  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|\mathbf{X}_n - \mathbf{a}| \geq \epsilon) = 0.$$

The statement that  $\mathbf{X}_n$  converges to  $\mathbf{a}$  in probability is represented by the notation

$$\text{plim}_{n \rightarrow \infty} \mathbf{X}_n = \mathbf{a} \quad \text{or} \quad \mathbf{X}_n \xrightarrow{p} \mathbf{a}.$$

This property of a sequence of random variables is called the (weak) law of large numbers.

# Consistency of the Sample Mean

## Consistency

The sample mean  $\bar{X}_n$  converges in probability to  $\mu$  as  $n$  goes to infinity, i.e.,

$$\text{plim}_{n \rightarrow \infty} \bar{X}_n = \mu.$$

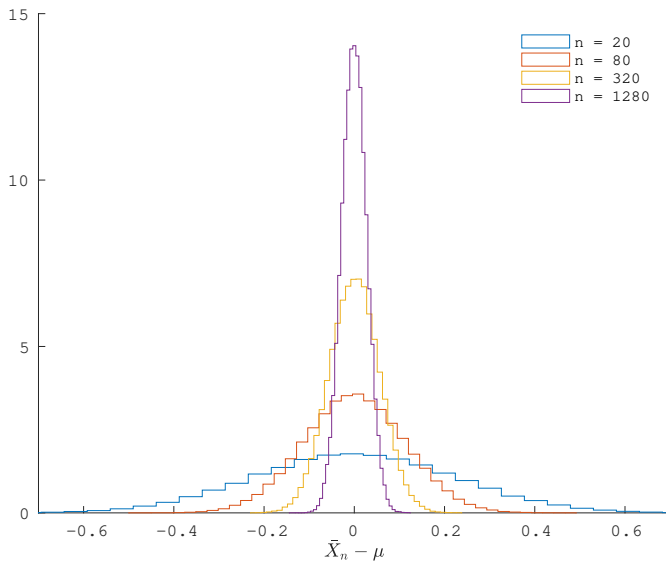
**Proof.** Applying the Chebyshev inequality, we have

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2},$$

for any  $\epsilon > 0$ . Thus

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0.$$





**Figure 16:** Consistency of the Sample Mean



# Central Limit Theorem

## Central Limit Theorem (Lindeberg and Levy)

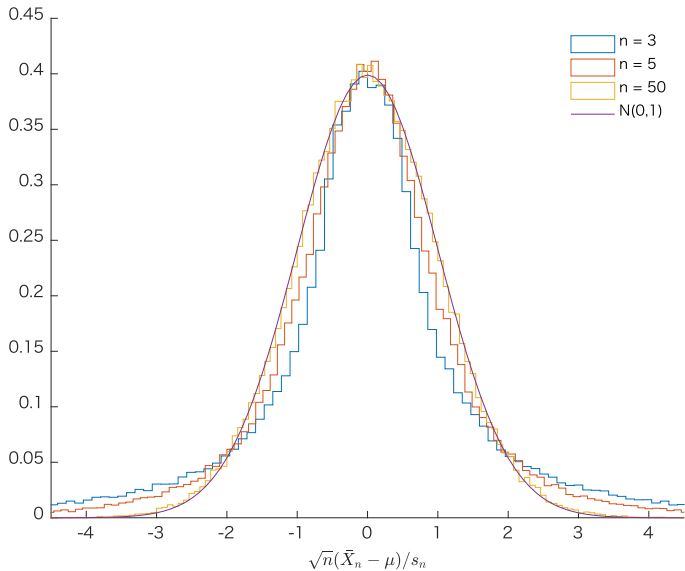
If  $X_1, \dots, X_n$  are independent random variables with mean  $\mu$  and variance  $\sigma^2$ ,

$$\lim_{n \rightarrow \infty} P \left( \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x \right) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

In other words, the central limit theorem states that the distribution of the sample mean  $\bar{X}_n$  is approximated by

$$\bar{X}_n \overset{a}{\sim} \mathcal{N} \left( \mu, \frac{\sigma^2}{n} \right),$$

if the sample size  $n$  is large enough.



**Figure 17:** Central Limit Theorem ( $s_n^2$  is the sample variance)<sub>138/144</sub>

# Convergence in Distribution

The result of the central limit theorem is a special case of **convergence in distribution** or **convergence in law**.

## Convergence in Distribution

Suppose that  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots$  is a sequence of random variables. This sequence converges in distribution to a random variable  $\mathbf{X}$  if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x) \quad \text{for every } x \in \mathbb{R},$$

where  $F_n(\cdot)$  is the c.d.f. of  $\mathbf{X}_n$  and  $F(\cdot)$  is the c.d.f. of  $\mathbf{X}$ .

The statement that  $\mathbf{X}_n$  converges in distribution to  $\mathbf{X}$  is represented by the notation

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X}.$$

Since there exists a one-to-one correspondence between the c.d.f. and the m.g.f, convergence in terms of the c.d.f. is equivalent to convergence in term of the m.g.f.

### **Convergence of the m.g.f.**

Suppose that  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots$  is a sequence of random variables. This sequence converges in distribution to a random variable  $\mathbf{X}$  if

$$\lim_{n \rightarrow \infty} M_n(t) = M(t) \quad \text{for every point around } t = 0,$$

where  $M_n(\cdot)$  is the m.g.f. of  $\mathbf{X}_n$  and  $M(\cdot)$  is the m.g.f. of  $\mathbf{X}$ .

# Proof of the Central Limit Theorem i

Define

$$Z_i = \frac{X_i - \mu}{\sigma}, \quad Y_i = \frac{Z_i}{\sqrt{n}}, \quad (i = 1, \dots, n).$$

Then

$$E[Z_i] = 0, \quad \text{Var}[Z_i] = 1,$$

and

$$S_n = \sum_{i=1}^n Y_i = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

We need to show that the m.g.f. of  $S_n$  converges to the m.g.f. of the standard normal distribution  $e^{\frac{t^2}{2}}$ .

## Proof of the Central Limit Theorem ii

Let  $M_Z(\cdot)$  be the m.g.f. of  $Z_i$  and  $M_Y(\cdot)$  be the m.g.f. of  $Y_i$ .

Then

$$M_Y(t) = M_Z\left(\frac{t}{\sqrt{n}}\right).$$

The Taylor expansion of  $M_Y(\cdot)$  around zero is

$$\begin{aligned} M_Y(t) &= M_Z\left(\frac{t}{\sqrt{n}}\right) = \sum_{j=0}^{\infty} \frac{1}{j!} E[Z_i^j] \left(\frac{t}{\sqrt{n}}\right)^j \\ &= 1 + \frac{1}{1!} E[Z_i] \left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2!} E[Z_i^2] \left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\left|\frac{t}{\sqrt{n}}\right|^2\right) \\ &= 1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right), \end{aligned}$$

## Proof of the Central Limit Theorem iii

where  $o(|t|^k)$  means that  $\lim_{t \rightarrow 0} o(|t|^k)|t|^{-k} = 0$ , i.e., the term goes to zero more rapidly than  $|t|^k$  as  $t \rightarrow 0$ .

Because  $Y_1, \dots, Y_n$  are independent,

$$M_{S_n}(t) = \prod_{i=1}^n M_{Y_i}(t) = \left\{ 1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \right\}^n.$$

Finally, we have

$$\lim_{n \rightarrow \infty} M_{S_n}(t) = \lim_{n \rightarrow \infty} \left\{ 1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \right\}^n = e^{\frac{t^2}{2}}.$$

Thus  $S_n \xrightarrow{d} \mathcal{N}(0, 1)$  is proved. □

## Example: Bernoulli Distribution

Suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_n$  independently follow the Bernoulli distribution with the probability of success  $\mathbf{p}$ . If  $\mathbf{n}$  is large enough,

$$\bar{\mathbf{X}}_n \overset{a}{\sim} \mathcal{N} \left( \mathbf{p}, \frac{\mathbf{p}(1 - \mathbf{p})}{n} \right),$$

since  $\mathbf{E}[\mathbf{X}_i] = \mathbf{p}$  and  $\text{Var}[\mathbf{X}_i] = \mathbf{p}(1 - \mathbf{p})$  ( $i = 1, \dots, N$ ).

Then

$$\mathbf{P} \left( |\bar{\mathbf{X}}_n - \mathbf{p}| \geq 1.96 \sqrt{\frac{\mathbf{p}(1 - \mathbf{p})}{n}} \right) \approx 0.05$$