Lecture Note: Foundations of Probability Theory

PROBABILITY AND STATISTICS B

Teruo Nakatsuma

Fall Semester 2021

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.

Python

- Python is a high-level programming language.
- Designed by Guido van Rossum
- Released in 1991
- Python is popular.
    - IEEE SPECTRUM
    - TIOBE

Why Python?

- It is free.
- It is slow in execution but highly manageable.
- Python codes are arguably more readable than other languages such as C/C++.
- Numerous packages have been developed for Python.
- Most of them are free and written in faster programming languages such as C/C++.

How To Obtain Python

- The official Python website
- Unfortunately, the plain Python does not include any useful tools for statistics / data science.
- Python distributions for scientific computing
    - Anaconda (we use this in the class)
    - ActivePython
    - Enthought Deployment Manager

Tools For Python Programming

- REPL (Read-Eval-Print-Loop)
  - Terminal-based REPL – **IPython**
  - Browser-based REPL – **Jupyter Notebook**
- An **integrated development environment (IDE)** is an application that consists of integrates an editor, a debugger, a profiler and other tools for developers.
  - **Spyder**
  - **PyCharm**
  - **Visual Studio Code**

Basic Packages

- **NumPy** – n-dimensional arrays and matrices
- **SciPy** – functions for scientific computing
- **Matplotlib** – 2D/3D plotting
- **Pandas** – data structure

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 $\boldsymbol{\Omega}$ denote the sample space of some experiment and let ω denote a possible outcome of the experiment. Since any outcome is a member of the sample space, the statement that ω is an outcome of the experiment is denoted by $\omega \in \boldsymbol{\Omega}$.
- An event is a collection of some outcomes. Thus the event A is a subset of the sample space $\boldsymbol{\Omega}$, i.e., $A \subseteq \boldsymbol{\Omega}$.

Example: Bankruptcy

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

$$\begin{cases} \omega_1: & HH; \\ \omega_2: & HT; \\ \omega_3: & TH; \\ \omega_4: & TT. \end{cases}$$

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

Example: Tossing a Coin ii

Let us define the following three events:

$$
\begin{aligned}
A &= \{\omega_1, \omega_2, \omega_3\} = \{HH, HT, TH\}, \\
B &= \{\omega_2, \omega_4\} = \{HT, TT\}, \\
C &= \{\omega_4\} = \{TT\}.
\end{aligned}
$$

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

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 $A \cap B$ is the event that both A and B occur. A union of two events $A \cup B$, on the other hand, is the one that either A or B occurs. The complement of an event A, denote by A^c, is the event that A does not occur. If $A \cap B = \varnothing$, we say that A and B are mutually exclusive or pairwise disjoint. Useful set operations are listed in the following table.

Table 1: Set Operations

$A \cap B$	intersection, $\{\omega : \omega \in A$ and $\omega \in B\}$ event that both A and B occur
$A \cup B$	union, $\{\omega : \omega \in A$ 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 occur
$A \setminus B$	difference, $\{\omega : \omega \in A$ and $\omega \notin B\} = A \cap B^c$ A occurs but B does not
$A \,\mathbf{\Delta}\, B$	symmetric difference, $(A \setminus B) \cup (B \setminus A)$
$A \subseteq B$	A is a subset of B, $^{\forall}\omega \in A, \omega \in B$ A occurs, then B occurs.
$A = B$	$A \subseteq B$ and $B \subseteq A$, i.e., A and B are equivalent.
$A \subset B$	$^{\forall}\omega \in A, \omega \in B$ but $^{\exists}\omega \in B$ such that $\omega \notin A$

Sequence of Events

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

$$\bigcap_{i=1}^{n} A_i = \{\omega \in \boldsymbol{\Omega} \ : \ {}^\forall i \in \{1, \ldots, n\}, \ \omega \in A_i\},$$

$$\bigcup_{i=1}^{n} A_i = \{\omega \in \boldsymbol{\Omega} \ : \ {}^\exists i \in \{1, \ldots, n\}, \ \omega \in A_i\}.$$

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

$$\begin{aligned} A &= \{\omega_1, \omega_2, \omega_3\} = \{HH, HT, TH\}, \\ B &= \{\omega_2, \omega_4\} = \{HT, TT\}, \\ C &= \{\omega_4\} = \{TT\}. \end{aligned}$$

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

$$\begin{aligned} C &= A^c, \\ C &\subset B, \\ A \cup B &= \boldsymbol{\Omega}, \\ A \cap C &= \varnothing. \end{aligned}$$

σ-Field

Field

A field (also algebra) $\mathcal{F}$ on Ω 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 $\mathcal{F} = \{\Omega, \varnothing\}$. #3 implies

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

When n can be infinite, such $\mathcal{F}$ is called the σ-field. A (σ-)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 $\boldsymbol{\Omega} = \{\omega_1, \omega_2\}$.

The simplest field on $\boldsymbol{\Omega}$ is

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

More exhaustive one is

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

Example: Asset Price i

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

Table 2: Asset Price Fluctuations

	Path			
Time	ω_1	ω_2	ω_3	ω_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

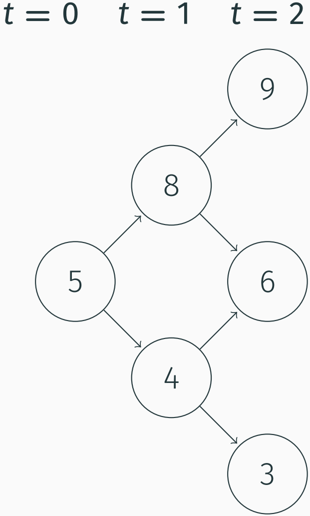


Figure 1: Tree of Price Fluctuations

Filtration

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

$$\begin{aligned}
\mathfrak{F}_0 &= \{\{\omega_1, \omega_2, \omega_3, \omega_4\}, \varnothing\}, \\
\mathfrak{F}_1 &= \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \{\omega_1, \omega_2, \omega_3, \omega_4\}, \varnothing\}, \\
\mathfrak{F}_2 &= \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \\
&\qquad \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \{\omega_1, \omega_4\}, \{\omega_2, \omega_3\}, \{\omega_2, \omega_4\}, \{\omega_3, \omega_4\}, \\
&\qquad \{\omega_1, \omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_4\}, \{\omega_1, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_4\}, \\
&\qquad \{\omega_1, \omega_2, \omega_3, \omega_4\}, \varnothing\}.
\end{aligned}$$

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

Axiomatic Definition of Probability

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

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

Axiom 2. $\mathbf{P}(\boldsymbol{\Omega}) = 1$.

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

$$\mathbf{P}\left(A_1 \cup \cdots \cup A_n\right) = \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n).$$

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

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

Properties

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

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

Interpretations of Probability

- Classical Interpretation

  Suppose every outcome is equally likely. The probability of an event A is defined as

  $$\mathbf{P}(A) = \frac{\text{\# of all outcomes in } A}{\text{\# of all outcomes in } \boldsymbol{\Omega}}.$$

- Frequentist Interpretation

  The probability of A is the limit of the ratio of occurrence in infinitely repeated trials, i.e.,

  $$\mathbf{P}(A) = \lim_{n \to \infty} \frac{\text{\# of occurrence of } A \text{ up to the } n\text{-th trial}}{n}.$$

- Bayesian Interpretation

  $\mathbf{P}(A)$ is a researcher's degree of belief in A.

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

$$\mathbf{P}(HH) = \mathbf{P}(HT) = \mathbf{P}(TH) = \mathbf{P}(TT).$$

Since

$$\mathbf{P}(HH) + \mathbf{P}(HT) + \mathbf{P}(TH) + \mathbf{P}(TT) = 1,$$

we have

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

Subjective Probability

- Bayesian statistics relies on the concept called subjective probability.
- The subjective probability still satisfies all mathematical properties of probability.
- Bruno de Finetti proved that probability could be derived as a subjective measure of uncertainty in relation to bookmaking.
- His founding is called the Dutch Book Theorem.

Dutch Book Theorem

Consider the following bet on event A.

- The amount of money the bettor bets is $\mathbf{P}(A)$.
- If A occurs, the bookmaker will pay 1 to the bettor.
- If A does not occur, the bettor will receive nothing.

If either bettor or bookmaker gains for sure, the bet is called a Dutch book. de Finitte's Dutch Book Theorem claims that a subjective measure of uncertainty must be a probability measure; otherwise it leads to a Dutch book.

Definition of Probability

Axiomatic Definition of Probability

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

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

Axiom 2. $\mathbf{P}(\Omega) = 1$.

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

$$\mathbf{P}(A_1 \cup \cdots \cup A_n) = \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n).$$

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

Payoff Function

The payoff function of the bet is defined as follows.

$$G_A = \begin{cases} 1 - \mathbf{P}(A), & \text{if } A \text{ occurs;} \\ -\mathbf{P}(A), & \text{otherwise.} \end{cases}$$

G_A is rearraged as

$$\begin{aligned} G_A(\omega) &= (1 - \mathbf{P}(A))1_A(\omega) - \mathbf{P}(A)(1 - 1_A(\omega)) \\ &= 1_A(\omega) - \mathbf{P}(A), \end{aligned}$$

where

$$1_A(\omega) = \begin{cases} 1, & (\omega \in A), \\ 0, & (\omega \notin A). \end{cases}$$

Proof of Axiom 1

Suppose $\mathbf{P}(A) < 0$. Then

$$G_A(\omega) = \begin{cases} 1 - \mathbf{P}(A) > 0, & (\omega \in A); \\ -\mathbf{P}(A) > 0, & (\omega \notin A). \end{cases}$$

Therefore this bet is a Dutch book. Hence $\mathbf{P}(A) \geqq 0$.

Proof of Axiom 2

Suppose we can bet on $\boldsymbol{\Omega}$. From the definition of the payoff function G_A, we have

$$G_{\Omega}(\omega) = 1_{\Omega}(\omega) - \mathbf{P}(\Omega),$$

Since any ω belongs to Ω, $1_{\Omega}(\omega) = 1$. Thus

$$G_{\Omega}(\omega) = 1 - \mathbf{P}(\Omega).$$

- If $\mathbf{P}(\Omega) < 1$, $G_{\Omega}(\omega) = 1 - \mathbf{P}(\Omega) > 0$. So the bettor will always gain.
- If $\mathbf{P}(\Omega) > 1$, $G_{\Omega}(\omega) = 1 - \mathbf{P}(\Omega) < 0$. So the bettor will always lose.

Therefore $\mathbf{P}(\Omega) = 1$.

Proof of Axiom 3 i

It is sufficient to show $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$ if n is finite in Axiom 3. Define $C = A \cup B$ and consider the following payoff function:

$$\begin{aligned} G(\omega) &= G_A(\omega) + G_B(\omega) - G_C(\omega) \\ &= 1_A(\omega) - \mathbf{P}(A) + 1_B(\omega) - \mathbf{P}(B) - (1_C(\omega) - \mathbf{P}(C)). \end{aligned}$$

Define further

- $\omega_1 \in A$: A occurs
- $\omega_2 \in B$: B occurs
- $\omega_3 \in C^c$: either A or B does not occur

Proof of Axiom 3 ii

Suppose $\mathbf{P}(A \cup B) > \mathbf{P}(A) + \mathbf{P}(B)$. Then

$$\begin{aligned} G(\omega_1) &= \mathbf{P}(C) - \mathbf{P}(A) - \mathbf{P}(B) > 0, \\ G(\omega_2) &= \mathbf{P}(C) - \mathbf{P}(A) - \mathbf{P}(B) > 0, \\ G(\omega_3) &= \mathbf{P}(C) - \mathbf{P}(A) - \mathbf{P}(B) > 0. \end{aligned}$$

Therefore the bettor will gain for sure.

If $\mathbf{P}(A \cup B) < \mathbf{P}(A) + \mathbf{P}(B)$, on the other hand, the bettor will lose for sure.

Thus $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

Independence

Two events are called independent if and only if

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

In general, the n events $A_1, \ldots, A_n$ are independent if, for any subset $A_{i_1}, \ldots, A_{i_m}$,

$$\mathbf{P}(A_{i_1} \cap \cdots \cap A_{i_m}) = \mathbf{P}(A_{i_1}) \times \cdots \times \mathbf{P}(A_{i_m}).$$

Conditional Probability

The conditional probability is defined as follows:

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

If A and B are independent,

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

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

Properties

1. $\mathbf{P}(A \cap B) = \mathbf{P}(A|B)\mathbf{P}(B)$.
2. $\mathbf{P}(A|B)\mathbf{P}(B) = \mathbf{P}(B|A)\mathbf{P}(A)$.
3. $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A|B \cap C)\mathbf{P}(B|C)\mathbf{P}(C)$.
4. $\mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right) =$
   $\mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1 \cap A_2) \cdots \mathbf{P}(A_n| \bigcap_{i=1}^{n-1} A_i)$.
5. $\mathbf{P}(A) = \mathbf{P}(A|B)\mathbf{P}(B) + \mathbf{P}(A|B^c)\mathbf{P}(B^c)$.
6. Suppose $B_1, \ldots, B_n$ are pairwise disjoint and $\bigcup_{i=1}^{n} B_i = \Omega$. Then $\mathbf{P}(A) = \sum_{i=1}^{n} \mathbf{P}(A|B_i)\mathbf{P}(B_i)$.

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 3: Balls in the Urn

		Picked Color	
	Initial State	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

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

Example: Urn iii

The probability to pick a red ball in the second experiment

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

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

Bayes' Theorem

Bayes' theorem is defined as

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

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

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

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 4: Diagnosis Test

	Does the patient have the disease?	
Reaction	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?

$$\begin{aligned}
\mathbf{P}(A|B) &= \frac{\mathbf{P}(B|A)P(A)}{\mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B|A^c)\mathbf{P}(A^c)} \\
&= \frac{\frac{99}{100} \times \frac{1}{100{,}000}}{\frac{99}{100} \times \frac{1}{100{,}000} + \frac{1}{100} \times \frac{99{,}999}{100{,}000}} \\
&= \frac{99}{100,098} \approx 9.89 \times 10^{-4}.
\end{aligned}$$

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

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

Since

$$\mathbf{P}(A) = \mathbf{P}(A|B)\mathbf{P}(B) + \mathbf{P}(A|B^c)\mathbf{P}(B^c),$$
$$\mathbf{P}(A^c) = \mathbf{P}(A^c|B)\mathbf{P}(B) + \mathbf{P}(A^c|B^c)\mathbf{P}(B^c),$$

we have

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

Let A_t denote the event that A occur at time t and we replace A and B with A_{t+1} and A_t respectively. Then

$$\begin{bmatrix} \mathbf{P}(A_{t+1}) \\ \mathbf{P}(A_{t+1}^c) \end{bmatrix} = \begin{bmatrix} \mathbf{P}(A_{t+1}|A_t) & \mathbf{P}(A_{t+1}|A_t^c) \\ \mathbf{P}(A_{t+1}^c|A_t) & \mathbf{P}(A_{t+1}^c|A_t^c) \end{bmatrix} \begin{bmatrix} \mathbf{P}(A_t) \\ \mathbf{P}(A_t^c) \end{bmatrix},$$

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 $\mathbf{P}(A_{t+1}|A_t^c)$ is interpreted as the probability that the chain moves from state A^c at t to state A at $t+1$.

Markov Chain iii

The matrix

$$\begin{bmatrix} \mathbf{P}(A_{t+1}|A_t) & \mathbf{P}(A_{t+1}|A_t^c) \\ \mathbf{P}(A_{t+1}^c|A_t) & \mathbf{P}(A_{t+1}^c|A_t^c) \end{bmatrix}$$

is called the transition matrix. When the transition matrix does not change over time, the Markov chain is time-homogeneous. The vector of probabilities at time $t = 0$ is called the initial probability vector.

Let $p_{1t} = \mathbf{P}(A_t)$, $p_{2t} = \mathbf{P}(A_t^c)$, and let π_{ij} $(i, j = 1, 2)$ denote the (i, j) element of the time-homogeneous transition matrix.

Then the Markov chain is rewritten as

$$\begin{bmatrix} p_{1,t+1} \\ p_{2,t+1} \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} p_{1,t} \\ p_{2,t} \end{bmatrix}.$$

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

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

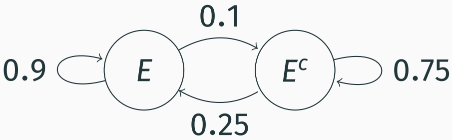
where $P_{i,t} = \mathbf{P}(A_t = A_i)$ and $\pi_{ij} = \mathbf{P}(A_{t+1} = A_i | A_t = A_j)$, $(i, j = 1, \ldots, k)$.

Example: Business Cycle

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

Table 5: Transition Matrix $\mathbf{\Pi}$

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



Properties

1. $\boldsymbol{p}_{t+k} = \boldsymbol{\Pi}^k \boldsymbol{p}_t \ (k = 1, 2, \dots)$.
2. In particular, $\boldsymbol{p}_t = \boldsymbol{\Pi}^t \boldsymbol{p}_0$. Thus all $\boldsymbol{p}_t$ in a Markov chain is determined by the initial probability vector $\boldsymbol{p}_0$ and the transition matrix $\boldsymbol{\Pi}$.
3. The probability vector $\boldsymbol{p}^*$ satisfies
$$\boldsymbol{p}^* = \boldsymbol{\Pi} \boldsymbol{p}^*,$$
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

For the transition matrix

$$\mathbf{\Pi} = \begin{bmatrix} 0.9 & 0.25 \\ 0.1 & 0.75 \end{bmatrix},$$

the stationary distribution is given by

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

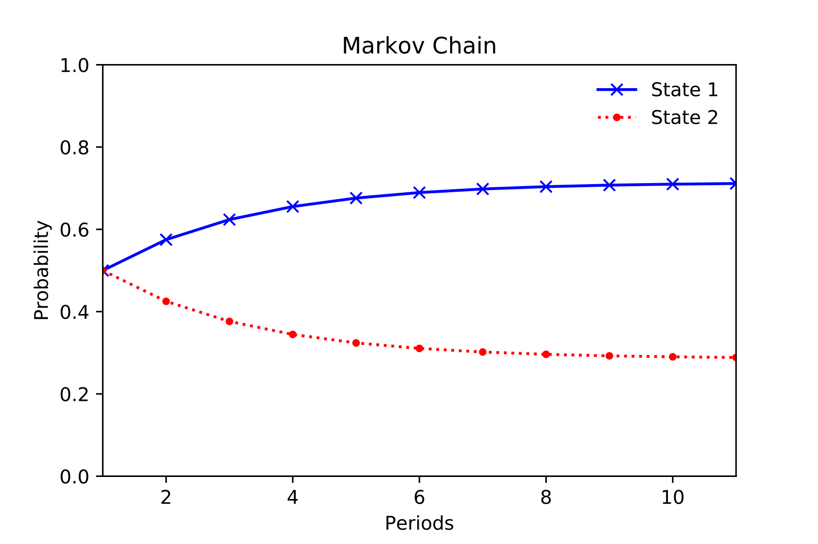


Figure 2: Convergence to the Stationary Distribution

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

Definition of a Random Variable

A r.v. X is a function from Ω 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 σ-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 σ-field on $\mathbb{R}$ is the σ-field that contains all Borel sets in $\mathbb{R}$.

Random Variable ii

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

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

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

$$\begin{aligned} \{\omega : \ X(\omega) = 1\} &= \{\omega_1\}, \\ \{\omega : \ X(\omega) = 0\} &= \{\omega_2\}, \\ \{\omega : \ X(\omega) \in \{0, 1\}\} &= \{\omega_1, \omega_2\} = \Omega, \\ \{\omega : \ X(\omega) \in B\} &= \varnothing \text{ for any } B \subseteq \mathbb{R} \backslash \{0, 1\}. \end{aligned}$$

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

Example: Tossing a Coin

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

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

Alternatively, if we define a binary random variable as

$$Y_i = \begin{cases} 1 & \text{if a head is obtained;} \\ 0 & \text{if a tail is obtained,} \end{cases} \qquad (i = 1, \ldots, 10).$$

X is obtained by $X = Y_1 + \cdots + Y_{10}$.

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 $(1, 2, 3, \ldots)$. Such X is often called a discrete random variable.

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

$$f(x) = \mathbf{P}(X = x).$$

Properties

Suppose that X is a discrete random variable and $f(x)$ is its probability mass function.

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

Example: Discrete Uniform Distribution

The probability mass function of the discrete uniform distribution is

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

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

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.m.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.m.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 $0 < p < 1$ and

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

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.

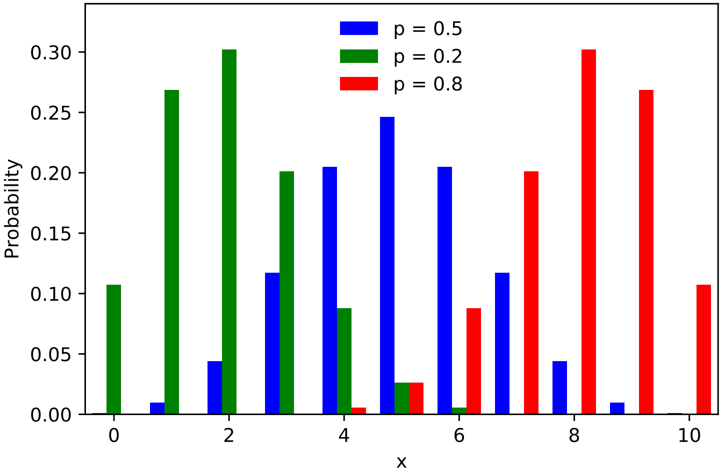


Figure 3: The p.m.f. of the binomial distribution

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 $\mathbf{P}(Y_i = 1) = p$ 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

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

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.m.f. is

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

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

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

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

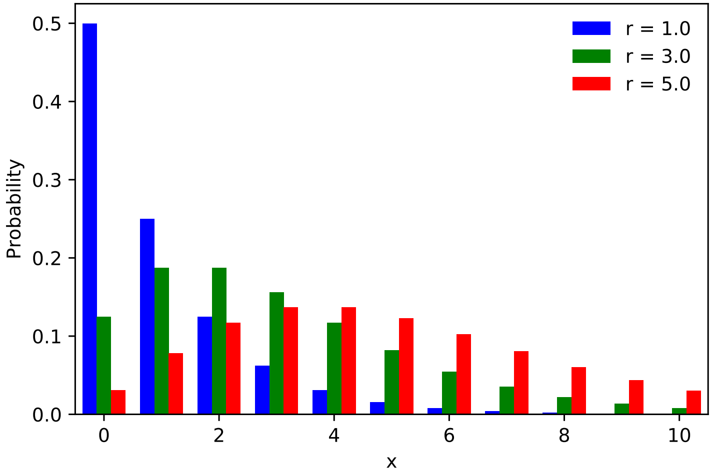


Figure 4: The p.m.f. of the negative binomial distribution

Example: Poisson Distribution i

The p.m.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 $n \to \infty$ while we keep λ constant (so p is negligibly small). Then

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

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n} = e^{-\lambda}$$

Thus we have

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

which is the p.m.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

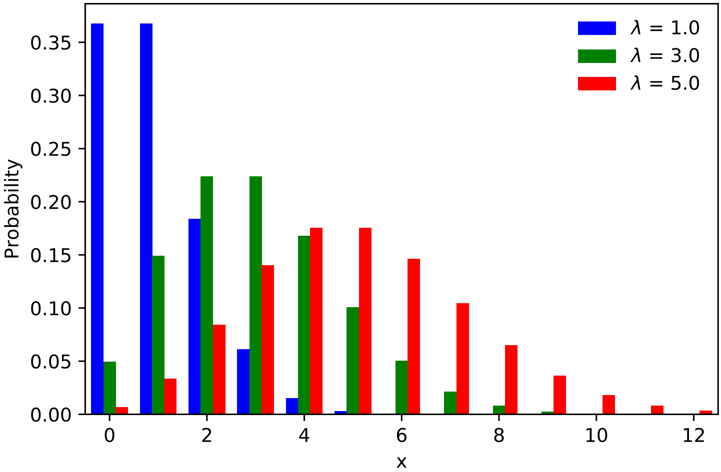


Figure 5: The p.m.f. of the Poisson distribution

Continuous Distribution

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

$$\mathbf{P}(X \in A) = \int_A f(x)dx,$$

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

1. $f(x) \geqq 0$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

Example: Continuous Uniform Distribution

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

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

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

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

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

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

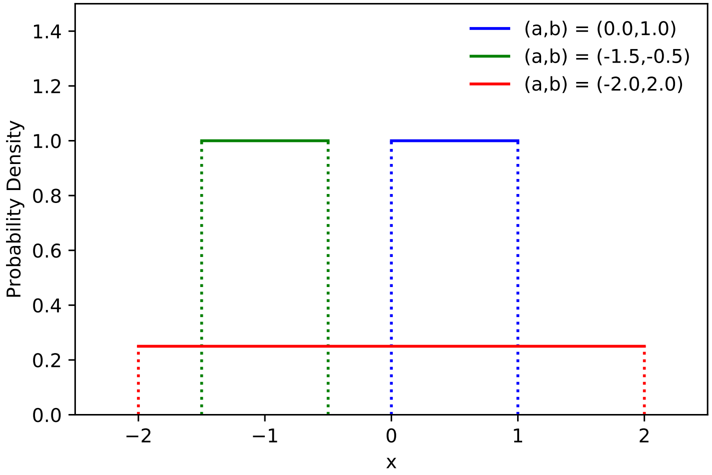


Figure 6: The p.d.f. of the uniform distribution

Example: Exponential Distribution

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

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

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

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

An exponential distribution is used for longevity or duration.

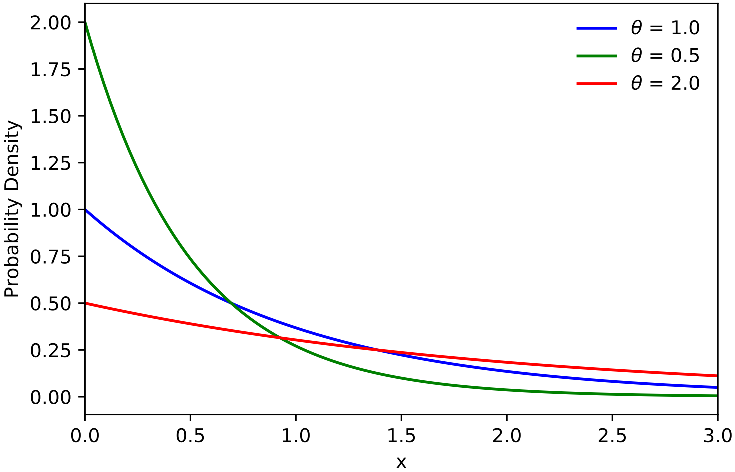


Figure 7: The p.d.f. of the exponential distribution

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.

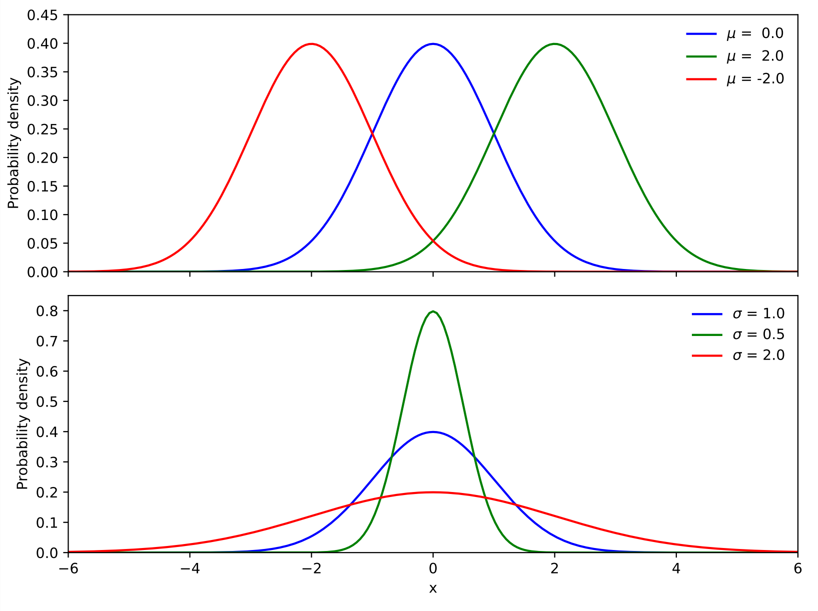


Figure 8: The p.d.f. of the normal distribution

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

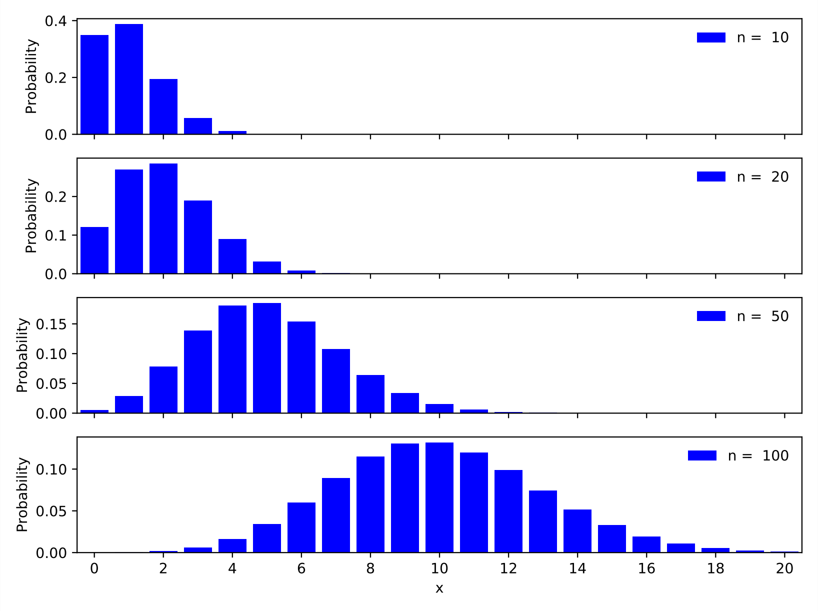


Figure 9: The binomial distribution converges to the normal distribution

Cumulative Distribution Function

The cumulative distribution function or c.d.f. of a random variable X is a function defined for each real number x as

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

Any c.d.f. of a random variable X must satisfy

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

Properties

1. $\mathbf{P}(X > x) = 1 - F(x)$.
2. $\mathbf{P}(x_1 < X \leqq x_2) = F(x_2) - F(x_1)$.
3. $\mathbf{P}(X < x) = \lim_{\epsilon \to 0, \epsilon > 0} F(x - \epsilon)$.
4. $\mathbf{P}(X = x) = \lim_{\epsilon \to 0, \epsilon > 0} F(x + \epsilon) - \lim_{\epsilon \to 0, \epsilon > 0} F(x - \epsilon)$.
5. $\mathbf{P}(X = x) = 0$ 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

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

2. Exponential distribution

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

3. Normal distribution

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

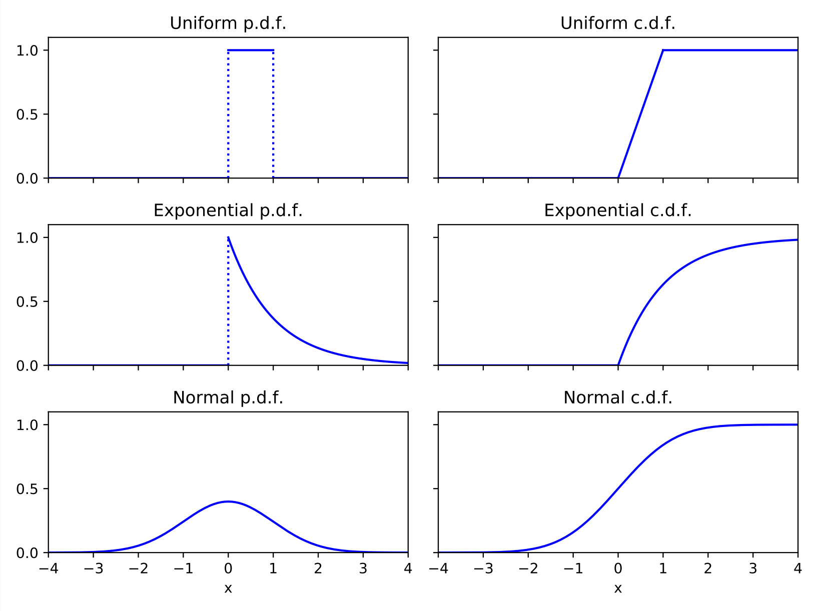


Figure 10: The c.d.f. of continuous distributions

Functions of a Random Variable

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

1. If Y is discrete, the p.m.f. of Y is given by

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

2. If Y is continuous, the c.d.f. of Y is given by

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

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

Example: Uniform Distribution

Suppose X has a uniform distribution

$$f(x) = \begin{cases} 1, & (0 \leqq x \leqq 1); \\ 0, & (x < 0 \text{ or } x > 1). \end{cases}$$

and let us define $Y = -\log(1 - X)$. Since $0 \leqq X \leqq 1$, $0 \leqq Y < \infty$. Then

$$\begin{aligned} G(y) &= \mathrm{P}(Y \leqq y) = \mathrm{P}(-\log(1 - X) \leqq y) = \mathrm{P}(X \leqq 1 - e^{-y}) \\ &= \int_0^{1-e^{-y}} f(x)dx = 1\big|_0^{1-e^{-y}} = 1 - e^{-y}. \end{aligned}$$

The p.d.f. of Y is

$$g(y) = \nabla_y G(y) = e^{-y}.$$

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 $r(a) < y < r(b)$,

$$G(y) = \mathrm{P}(r(X) \leqq y) = \mathrm{P}(X \leqq s(y)) = F(s(y)).$$

Change of Variables Formula ii

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

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

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

$$G(y) = \mathrm{P}(r(X) \leqq y) = \mathrm{P}(X \geqq s(y)) = 1 - F(s(y)),$$

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

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

Change of Variables Formula iii

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

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

Example: Uniform Distribution

Let us define $Y = -\log(1 - X)$ where X has a uniform distribution on $[0, 1]$. Note that $r(x) = -\log(1 - x)$ is an strictly increasing function of x. The inverse of $r(x)$ is

$$s(y) = 1 - e^{-y}.$$

By applying the change of variables formula, we obtain the p.d.f. of Y as

$$g(y) = f(s(y))|\nabla_y s(y)| = 1 \times \left|e^{-y}\right| = e^{-y},$$

which is the same p.d.f. as before.

Probability Integral Transformation i

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

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

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

Probability Integral Transformation ii

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

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

Thus Z is equivalent to X.

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

Example: Exponential Distribution

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

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

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

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

for $0 \leqq y \leqq 1$. Therefore we can generate an exponential random number as follows.

Step 1. Generate a random number y from the uniform distribution over $[0, 1]$.

Step 2. Compute $x = -\theta \log(1 - y)$.

Expectation

The expectation or expected value of a random variable X is defined as

Definition: Expectation of a Random Variable

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

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

Table 6: Expectation of Selected Distributions

Distribution	p.m.f. / p.d.f	$\mathbb{E}[X]$
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	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ
Uniform	$\frac{1}{b-a}$	$\frac{a+b}{2}$
Exponential	$\frac{1}{\theta}e^{-\frac{x}{\theta}}$	θ
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ

Discretization of a Random Variable

Suppose X is a continuous non-negative random variables and define a discrete non-negative random variable X_n as

$$X_n = \begin{cases} 0, & \left(0 \leqq X < \frac{1}{2^n}\right); \\ \frac{1}{2^n}, & \left(\frac{1}{2^n} \leqq X < \frac{2}{2^n}\right); \\ \quad\vdots & \\ \frac{i-1}{2^n}, & \left(\frac{i-1}{2^n} \leqq X < \frac{i}{2^n}\right); \\ \quad\vdots & \\ \frac{2^n n-1}{2^n}, & \left(\frac{2^n n-1}{2^n} \leqq X < n\right); \\ n, & (X \geqq n). \end{cases}$$

Note that (i) $X_n \leqq X$ for all n, and (ii) $\lim_{n\to\infty} X_n = X$.

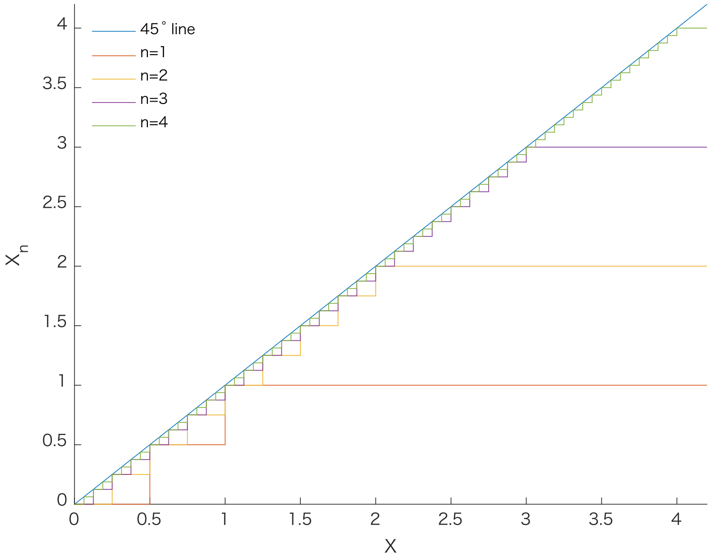


Figure 11: Discretized Random Variable

Then the expectation of X_n is given by

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

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

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

which is a type of Lebesgue–Stieltjes integral. If this limit exists, the expectation of X is defined as

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

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

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

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

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

$$\begin{aligned} X^+ &= \max\{X, 0\} \geqq 0, \\ X^- &= \max\{-X, 0\} \geqq 0. \end{aligned}$$

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

$$\mathbf{E}[X] = \mathbf{E}[X^+] - \mathbf{E}[X^-].$$

Properties

$X, Y, X_1, \ldots, X_n$: random variables
$a, b, c, a_1, \ldots, a_n$: real numbers

1. $\mathbf{E}[X + c] = \mathbf{E}[X] + c$.
2. $\mathbf{E}[aX] = a\mathbf{E}[X]$.
3. $\mathbf{E}[aX + c] = a\mathbf{E}[X] + c$.
4. $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.
5. $\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c$.
6. $\mathbf{E}[X_1 + \cdots + X_n] = \sum_{i=1}^{n} \mathbf{E}[X_i]$.
7. $\mathbf{E}[a_1X_1 + \cdots + a_1X_n + c] = \sum_{i=1}^{n} a_i\mathbf{E}[X_i] + c$.

Variance

The variance of a random variable X is defined as

Definition: Variance of a Random Variable

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

where $\mu = \mathbf{E}[X]$.

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 μ.

Table 7: Variance of Selected Distributions

Distribution	p.m.f. / p.d.f	$\mathbf{Var}[X]$
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	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ
Uniform	$\frac{1}{b-a}$	$\frac{(b-a)^2}{12}$
Exponential	$\frac{1}{\theta}e^{-\frac{x}{\theta}}$	θ^2
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	σ^2

Properties

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

Skewness

The skewness of a random variable is defined as

Definition of Skewness

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

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

The skewness β_1 tells whether the distribution is symmetric around the mean μ 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.

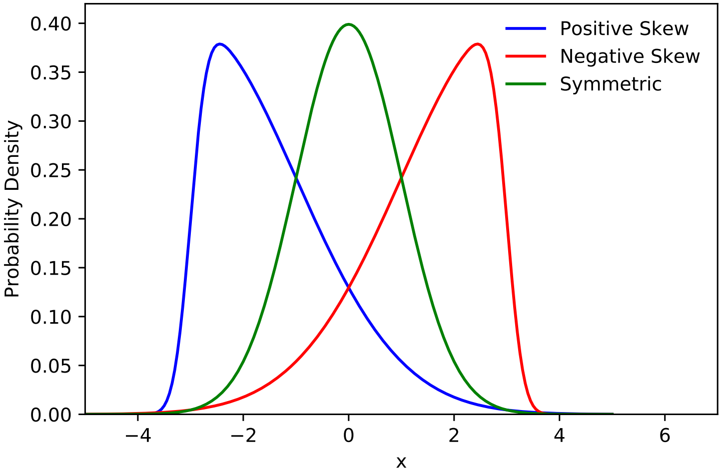


Figure 12: Skewness

Kurtosis

The kurtosis of a random variable is defined as

Definition of Kurtosis

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

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

The kurtosis β_2 is a measurement of thickness/heaveiness 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).

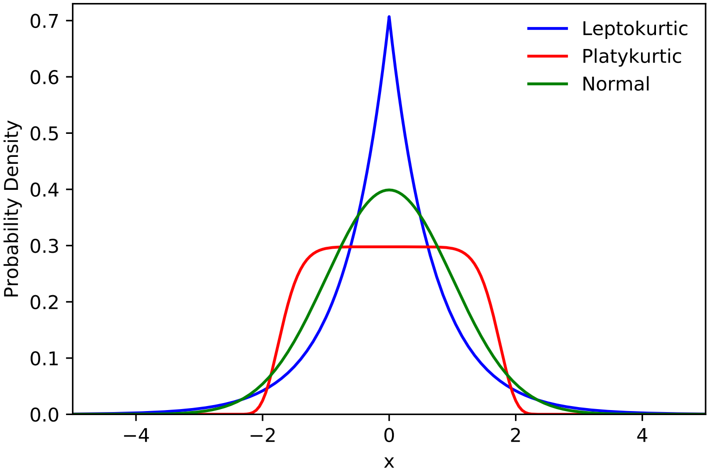


Figure 13: Kurtosis

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

Definition: Moment Generating Function

$$M_x(t) = \mathbf{E}[e^{tX}].$$

By applying the Maclaurin series, we have

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

Moment Generating Function ii

Since

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

we have

$$\nabla_t^k M_x(0) = \mathbf{E}[X^k].$$

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

Table 8: Moment Generating Functions for Selected Distributions

Distribution	p.m.f / p.d.f.	m.g.f.
Bernoulli	$p^x q^{1-x},\ q = 1 - p$	$q + pe^t$
Binomial	$\binom{n}{x} p^x q^{n-x}$	$(q + pe^t)^n$
Neg. Binomial	$\binom{x+r-1}{x} p^r q^x$	$\left(\frac{p}{1-qe^t}\right)^r$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	$e^{\lambda(e^t-1)}$
Uniform	$\frac{1}{b-a}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Exponential	$\frac{1}{\theta} e^{-\frac{x}{\theta}}$	$\frac{1}{1-\theta t}$
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.

Example: Binomial Distribution

We can prove the sum of i.i.d. Bernoulli random variables follows a binomial distribution with the moment generating function.

Suppose $X_1, \ldots, X_n$ independently follow the same Bernoulli distribution with the probability of success p. Since the m.g.f. of X_i is $q + pe^t$, $(q = 1 - p)$, the m.g.f. of $S_n = \sum_{i=1}^{n} X_i$ is given as

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(q + pe^t\right)^n,$$

which is the m.g.f. of the binomial distribution. Because of the uniqueness of the m.g.f., we conclude that the distribution of S_n is binomial.

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 mass function or joint p.m.f. is defined as

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

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 9: 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

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

$$\mathbf{P}(X + Y = 5) = 0 + 0.1 + 0.2 = 0.3$$

$$\mathbf{P}(X = 2) = 0.3 + 0 + 0.1 + 0.2 = 0.6$$

$$\mathbf{P}(X \leqq 2 \text{ and } Y \geqq 3) = 0.1 + 0 + 0.1 + 0.2 = 0.4$$

Marginal Distribution

Given the joint p.m.f. $f(x, y)$, the marginal probability mass function or marginal p.m.f. of X is defined as

$$f_x(x) = \mathbf{P}(X = x) = \sum_y f(x, y).$$

In the same manner, the marginal p.m.f. of Y is defined as

$$f_y(y) = \mathbf{P}(Y = y) = \sum_x f(x, y).$$

Table 10: Joint and Marginal Distributions

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	

$$\begin{aligned}
\mathbf{E}[X] &= 0.2 \times 1 + 0.6 \times 2 + 0.2 \times 3 = 2 \\
\mathbf{E}[Y] &= 0.4 \times 1 + 0.2 \times 2 + 0.2 \times 3 + 0.2 \times 4 = 2.2 \\
\mathbf{Var}[X] &= 0.2 \times (1-2)^2 + 0.6 \times (2-2)^2 + 0.2 \times (3-2)^2 = 0.4 \\
\mathbf{Var}[Y] &= 0.4 \times (1-2.2)^2 + 0.2 \times (2-2.2)^2 \\
&\quad + 0.2 \times (3-2.2)^2 + 0.2 \times (4-2.2)^2 = 1.36
\end{aligned}$$

Conditional Distribution

The conditional probability of $X = x$ given $Y = y$ is

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

Y X	1	$f(x\|1)$	2	$f(x\|2)$	3	$f(x\|3)$	4	$f(x\|4)$
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 varaibles (X, Y) has a bivariate distribution if the probability that a point (X, Y) is located in a region A is given by

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

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

The joint p.d.f. should satisfy

1. $f(x, y) \geqq 0$.
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 X is

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2).$$

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

$$X|Y = y \sim \mathcal{N}\left(\mu_x + \frac{\rho\sigma_x}{\sigma_y}(y-\mu_y), \sigma_x^2(1-\rho^2)\right).$$

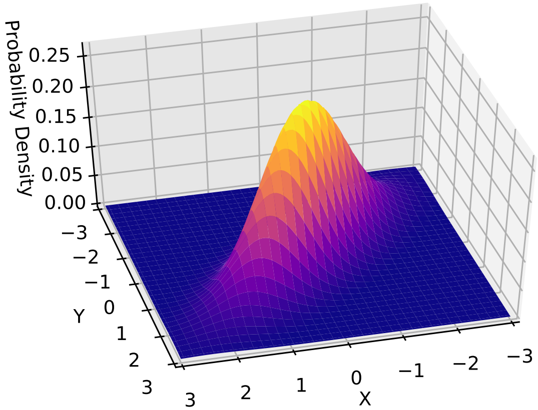


Figure 14: Bivariate Normal Distribution (Surface Plot)

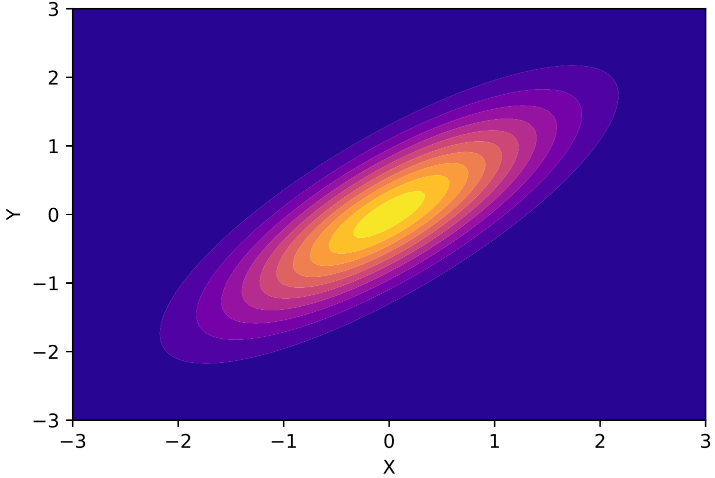


Figure 15: Bivariate Normal Distribution (Contour Plot)

Covariance, Correlation, and Independence

The covariance of two random variables X and Y is

$$\mathbf{Cov}[X,Y] = \mathbf{E}[(X-\mu_x)(Y-\mu_y)],\ \mu_x = \mathbf{E}[X],\ \mu_y = \mathbf{E}[Y].$$

The correlation (coefficient) of X and Y is

$$\rho_{XY} = \frac{\mathbf{Cov}[X,Y]}{\sigma_x \sigma_y}, \quad \sigma_x^2 = \mathbf{Var}[X], \quad \sigma_y^2 = \mathbf{Var}[Y].$$

X and Y are mutually independent if and only if

$$f(x,y) = f_x(x) f_y(y), \Leftrightarrow f_x(x|y) = f_x(x), \Leftrightarrow f_y(y|x) = f_y(y).$$

Properties

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

Example: Discrete Bivariate Distribution

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	

$$\begin{aligned}
\mathbf{E}[XY] &= 0.1 \times 1 \times 1 + 0.1 \times 1 \times 3 + 0.3 \times 2 \times 1 \\
&\quad + 0.1 \times 2 \times 3 + 0.2 \times 2 \times 4 + 0.2 \times 3 \times 2 \\
&= 0.1 + 0.3 + 0.6 + 0.6 + 1.6 + 1.2 = 4.4 \\
\mathbf{Cov}[X, Y] &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = 4.4 - 2 \times 2.2 = 0
\end{aligned}$$

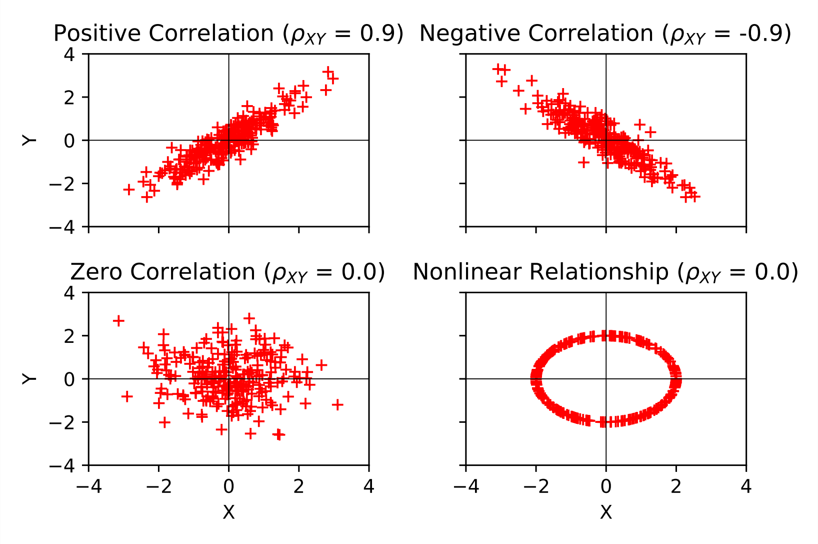


Figure 16: Scatter Plots for Illustration of Correlation

Conditional Expectation

The conditional expectation of X given $Y = y$ is defined as

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

If Y is not fixed at any specific value, $\mathbf{E}[X|Y]$ is also a random variable sicne it is a function of Y.

The conditional variance of X given Y is defined as

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

Properties

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

A joint p.m.f. of n discrete random variables $X_1, \ldots, X_n$ is defined as

$$f(x_1, \ldots, x_n) = \mathbf{P}(X_1 = x_1, \ldots, X_n = x_n).$$

A joint p.d.f of continuous $X_1, \ldots, X_n$ is

$$\mathbf{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{A_1} \cdots \int_{A_n} f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

A marginal p.d.f. of X_1 is

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

A conditional p.d.f. of $X_1, \ldots, X_m$ given $X_{m+1}, \ldots, X_n$ is

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

Independence

If random variables $X_1, \ldots, X_n$ are mutually independent,

Discrete random variables:

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

Continuous random variables:

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

Properties

Suppose $X_1, \ldots, X_n$ are mutually independent.

1. $f_i(x_i | x_j) = f_i(x_i)$, $(i \neq j)$.
2. $\mathbf{E}[X_i X_j] = \mathbf{E}[X_i]\mathbf{E}[X_j]$, $(i \neq j)$.
3. $\mathbf{E}[X_1 \times \cdots \times X_n] = \mathbf{E}[X_1] \times \cdots \times \mathbf{E}[X_n]$.
4. $\mathbf{Var}[X_i + X_j] = \mathbf{Var}[X_i] + \mathbf{Var}[X_j]$, $(i \neq j)$.
5. $\mathbf{Var}[X_1 + \cdots + X_n] = \mathbf{Var}[X_1] + \cdots + \mathbf{Var}[X_n]$.
6. $\mathbf{Var}[a_1 X_1 + \cdots + a_n X_n + b] =$ $a_1^2 \mathbf{Var}[X_1] + \cdots + a_n^2 \mathbf{Var}[X_n]$.

Independently And Identically Distributed R.V.'s

Random variables $X_1, \ldots, X_n$ are said to be independently and identically distributed (i.i.d.) when

1. they are mutually independent.
2. the marginal distribution of X_i is identical for all $i = 1, \ldots, n$, that is,

$$f_1(x) = \cdots = f_n(x) = f(x), \quad {}^{\forall} x \in \mathbb{R}.$$

If $X_1, \ldots, X_n$ are i.i.d., the joint p.m.f. or p.d.f. becomes

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

Examples

1. Bernoulli distribution
$$f(x_1,\dots,x_n)=\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}=\theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{n-\sum_{i=1}^{n}x_i}.$$
2. Exponential Distribution
$$f(x_1,\dots,x_n)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-\frac{x_i}{\theta}}=\theta^{-n}e^{-\frac{\sum_{i=1}^{n}x_i}{\theta}}.$$
3. Normal Distribution
$$\begin{aligned}f(x_1,\dots,x_n)&=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\,\mathbf{exp}\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\\&=(2\pi\sigma^2)^{-\frac{n}{2}}\,\mathbf{exp}\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right].\end{aligned}$$

Variance-Covariance Matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{Var}[X_1] & \mathbf{Cov}[X_1, X_2] & \cdots & \mathbf{Cov}[X_1, X_n] \\ \mathbf{Cov}[X_2, X_1] & \mathbf{Var}[X_2] & \cdots & \mathbf{Cov}[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Cov}[X_n, X_1] & \mathbf{Cov}[X_n, X_2] & \cdots & \mathbf{Var}[X_n] \end{bmatrix},$$

is called the (variance-)covariance matrix.

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

is called the correlation matrix.

Properties

1. In the case of a bivariate distribution,

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

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

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

Example: Portfolio Analysis

Portfolio = a collection of assets that the investor holds.

The return on a portfolio of assets is a weighted average of the return on the individual assets.

Return of a portfolio

$$R_P = \sum_{i=1}^{n} w_i R_i.$$

R_P: the return on the portfolio

R_i: the return on the i-th asset

w_i: the fraction of the funds invested in the i-th asset

Measure Of Average Outcome

Expected return

$$\mu_P = \mathbf{E}(R_P) = \mathbf{E}\left[\sum_{i=1}^{n} w_i R_i\right] = \sum_{i=1}^{n} w_i \mathbf{E}[R_i]$$
$$= \sum_{i=1}^{n} w_i \mu_i = w^\intercal \mu,$$

where

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \mathbf{E}[R_1] \\ \vdots \\ \mathbf{E}[R_n] \end{bmatrix}.$$

Measure Of Dispersion

Variance

$$
\begin{aligned}
\sigma_P^2 = \mathbf{E}[(R_P - \mu_P)^2] &= \mathbf{E}\left[\left\{\sum_{i=1}^{n} w_i (R_i - \mu_i)\right\}^2\right] \\
&= \mathbf{E}\left[\{w^\intercal (R - \mu)\}^2\right] = \mathbf{E}\left[w^\intercal (R - \mu)(R - \mu)^\intercal w\right] \\
&= w^\intercal \mathbf{E}\left[(R - \mu)(R - \mu)^\intercal\right] w \\
&= w^\intercal \mathbf{\Sigma} w,
\end{aligned}
$$

where $R = \begin{bmatrix} R_1 & \cdots & R_n \end{bmatrix}^\intercal$.

Optimal Portfolio Selection Problem

We can form an optimal portfolio by finding the *sweet spot* that balances the risk $\boldsymbol{\sigma_P}$ and the return $\boldsymbol{\mu_P}$.

Mean-variance approach

$$\begin{aligned} \max_{w} \quad & w^\top \boldsymbol{\mu} - \gamma w^\top \boldsymbol{\Sigma} w, \\ \text{subject to} \quad & w^\top \boldsymbol{\iota} = 1, \ \text{(budget constraint)}; \\ & w \geqq \mathbf{0}, \ \text{(no shortselling)}, \end{aligned} \tag{1}$$

where $\boldsymbol{\iota}$ is the $n \times 1$ vector of 1's.

γ is the degree of risk aversion. A large γ leads to a more conservative (less risky but less profitable) portfolio. The optimal portfolio selection method based on (1) is called the mean-variance approach.

Example: Multivariate Normal Distribution

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

$$
\begin{aligned}
& f(x_1, \ldots, x_n) \\
& \quad = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu)'\boldsymbol{\Sigma}^{-1}(x-\mu)\right],
\end{aligned}
$$

where $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\intercal$.

The bivariate normal distribution is a special case of the multivariate normal distribution with $n = 2$.

Markov Inequality

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

$$\mathbf{P}(X \geqq t) \leqq \frac{\mathbf{E}[X]}{t}.$$

Proof. Without loss of generality, we suppose X is

continuous.

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

□

Chebyshev Inequality

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

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

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

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

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

Properties of the Sample Mean

Suppose that we have data $X_1, \ldots, X_n$ which are drawn from the same population independently. This implies that $X_1, \ldots, X_n$ follows the same probability distribution and are mutually independent. Further suppose that the expectation of the distribution is μ and the variance is σ^2. Then we define the sample mean as

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

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

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

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

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

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

$$\mathbf{P}\left(|\bar{X}_n - \mu| \geqq 3\frac{\sigma}{\sqrt{n}}\right) \leqq \frac{1}{9} \approx 11\%.$$

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

$$\mathbf{P}\left(|\bar{X}_n - \mu| \geqq 3\frac{\sigma}{\sqrt{n}}\right) \approx 0.27\%,$$

and

$$\mathbf{P}\left(|\bar{X}_n - \mu| \geqq 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 5\%.$$

Example: Bernoulli Distribution

Suppose that $X_1, \ldots, X_n$ independently follow the Bernoulli distribution with the probability of success p. Since $\mathbf{E}[X_i] = p$ and $\mathbf{Var}[X_i] = p(1-p)$, $i = 1, \ldots, n$,

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

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 $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables. This sequence converges in probability to a real number a if for any given number $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbf{P}(|X_n - a| \geqq \epsilon) = 0.$$

The statement that X_n converges to a in probability is represented by the notation

$$\operatorname*{plim}_{n \to \infty} X_n = a \quad \text{or} \quad X_n \xrightarrow{p} a.$$

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 μ as n goes to infinity, i.e.,

$$\operatorname*{plim}_{n\to\infty} \bar{X}_n = \mu.$$

Proof. Applying the Chebyshev inequality, we have

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

for any $\epsilon > 0$. Thus

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

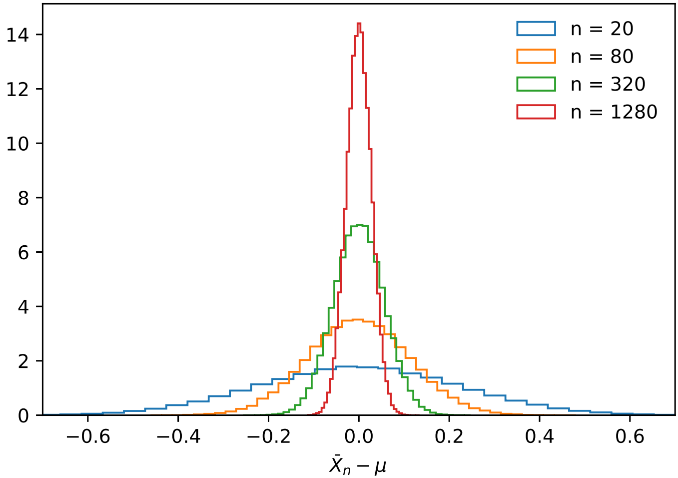


Figure 17: Consistency of the Sample Mean

Another mode of convergence, convergence in distribution or convergence in law, is defined as follows.

Convergence in Distribution

Suppose that $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables. This sequence converges in distribution to a random variable X if

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

where $F_n(\cdot)$ is the c.d.f. of X_n and $F(\cdot)$ is the c.d.f. of X.

Convergence in Distribution ii

The statement that X_n converges in distribution to X is represented by the notation

$$X_n \rightsquigarrow X \quad \text{or} \quad X_n \xrightarrow{d} X.$$

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 in Distribution iii

Convergence of the m.g.f.

Suppose that $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables. This sequence converges in distribution to a random variable X if

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

where $M_n(\cdot)$ is the m.g.f. of X_n and $M(\cdot)$ is the m.g.f. of X.

Central Limit Theorem (Lindeberg and Levy)

If $X_1, \ldots, X_n$ are independent random variables with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$,

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

In other words, the central limit theorem implies

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow \mathcal{N}(0, 1).$$

Central Limit Theorem ii

Asymptotic Normality with Unknown σ^2

Suppose $X_1, \ldots, X_n$ are independent random variables with mean μ and variance σ^2, but σ^2 is unknown. We estimate σ^2 with the sample variance:

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \rightsquigarrow \mathcal{N}(0, 1).$$

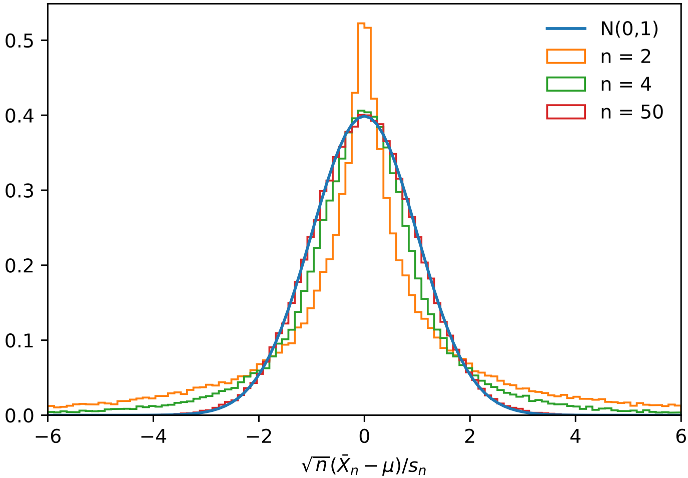


Figure 18: Asymptotic Normality

Proof of the Central Limit Theorem i

Define

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

Then

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

and

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

Proof of the Central Limit Theorem ii

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

Let $M_Z(\cdot)$ be the m.g.f. of Z_i and $M_y(\cdot)$ be the m.g.f. of Y_i. The subscript i is omitted because $M_Z(\cdot)$ and $M_y(\cdot)$ are identical for all i. Then

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

Thus the m.g.f. of S_n is given as

$$M_{S_n}(t) = \prod_{i=1}^{n} M_Y(t) = \prod_{i=1}^{n} M_Z\left(\frac{t}{\sqrt{n}}\right) = \left\{M_Z\left(\frac{t}{\sqrt{n}}\right)\right\}^n.$$

Next, we introduce the cumulant generating function:

$$K_x(t) = \log M_x(t),$$

which can be expanded as

$$K_x(t) = \sum_{j=1}^{\infty} \frac{\kappa_j}{j!} t^j = \mu t + \frac{\sigma^2}{2} t^2 + \frac{\kappa_3}{3!} t^3 + \cdots ,$$

where $\kappa_1 = \mu = \mathbf{E}[X]$ and $\kappa_2 = \sigma^2 = \mathbf{Var}[X]$. κ_j is called the j-th cumulant. Note that

$$M_x(t) = \exp\left(\sum_{j=1}^{\infty} \frac{\kappa_j}{j!} t^j \right).$$

Proof of the Central Limit Theorem iv

The cumulant generating function of S_n is

$$\begin{aligned}
K_{S_n}(t) &= \log M_{S_n}(t) \\
&= \log\left(\left\{M_Z\left(\frac{t}{\sqrt{n}}\right)\right\}^n\right) = n \times K_Z\left(\frac{t}{\sqrt{n}}\right) \\
&= n \times \left\{0 \times \frac{t}{\sqrt{n}} + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + \sum_{j=3}^{\infty} \frac{\kappa_j}{j!}\left(\frac{t}{\sqrt{n}}\right)^j\right\} \\
&= \frac{t^2}{2} + \sum_{j=3}^{\infty} \frac{\kappa_j}{j!}\frac{t^j}{n^{\frac{j}{2}-1}},
\end{aligned}$$

where $\mu = 0$ and $\sigma^2 = 1$.

Proof of the Central Limit Theorem v

Since

$$\lim_{n\to\infty} \frac{\kappa_j}{j!} \frac{t^j}{n^{\frac{j}{2}-1}} = 0,$$

we have

$$\lim_{n\to\infty} K_{S_n}(t) = \frac{t^2}{2}.$$

Therefore

$$\lim_{n\to\infty} M_{S_n}(t) = e^{\frac{t^2}{2}}.$$

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

Example: Bernoulli Distribution

Suppose that $X_1, \ldots, X_n$ independently follow the Bernoulli distribution with the probability of success p. If n is large enough,

$$\frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{p(1-p)}} \rightsquigarrow \mathcal{N}(0, 1),$$

since $\mathbf{E}[X_i] = p$ and $\mathbf{Var}[X_i] = p(1-p)$ $(i = 1, \ldots, n)$. Then

$$\mathbf{P}\left(|\bar{X}_n - p| \geqq 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 5\%.$$