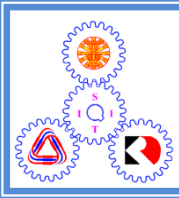

Sirindhorn International Institute of Technology
Thammasat University
School of Information, Computer and Communication Technology

ECS 315: Probability and Random Processes

Prapun Suksompong, Ph.D.
prapun@siit.tu.ac.th

September 26, 2011

This note covers fundamental concepts in probability and random processes for undergraduate students in electronics and communication engineering.

Despite my best endeavors, this note will not be error-free. I hope that none of the mistakes are misleading. But since probability can be a subtle subject, if I have made some slips of principle, do not hesitate to tell me about them. Greater minds than mine have erred in this field.

Contents

ECS315 2011/1 Part I Dr.Prapun

1 Motivation

Whether you like it or not, probabilities rule your life. If you have ever tried to make a living as a gambler, you are painfully aware of this, but even those of us with more mundane life stories are constantly affected by these little numbers.

Example 1.1. Some examples from daily life where probability calculations are involved are the determination of insurance premiums, the introduction of new medications on the market, opinion polls, weather forecasts, and DNA evidence in courts. Probabilities also rule who you are. Did daddy pass you the X or the Y chromosome? Did you inherit grandma's big nose?

Meanwhile, in everyday life, many of us use probabilities in our language and say things like "I'm 99% certain" or "There is a one-in-a-million chance" or, when something unusual happens, ask the rhetorical question "What are the odds?". [17, p 1]

1.1 Randomness

1.2. Many clever people have thought about and debated what randomness really is, and we could get into a long philosophical discussion that could fill up a whole book. Let's not. The French mathematician Laplace (1749–1827) put it nicely:

> "Probability is composed partly of our ignorance, partly of our knowledge."

Inspired by Laplace, let us agree that you can use probabilities whenever you are faced with uncertainty. [17, p 2]

1.3. Random phenomena arise because of [12]:

(a) our partial ignorance of the generating mechanism

(b) the laws governing the phenomena may be fundamentally random (as in quantum mechanics)

(c) our unwillingness to carry out exact analysis because it is not worth the trouble

Example 1.4. Communication Theory [24]: The essence of communication is randomness.

(a) **Random Source**: The transmitter is connected to a random source, the output of which the receiver cannot predict with certainty.

- If a listener knew in advance exactly what a speaker would say, and with what intonation he would say it, there would be no need to listen!

(b) **Noise**: There is no communication problem unless the transmitted signal is disturbed during propagation or reception in a random way.

(c) Probability theory is used to *evaluate the performance* of communication systems.

Example 1.5. Random numbers are used directly in the transmission and security of data over the airwaves or along the Internet.

(a) A radio transmitter and receiver could switch transmission frequencies from moment to moment, seemingly at random, but nevertheless in synchrony with each other.

(b) The Internet data could be credit-card information for a consumer purchase, or a stock or banking transaction secured by the clever application of random numbers.

Example 1.6. Randomness is an essential ingredient in games of all sorts, computer or otherwise, to make for unexpected action and keen interest.

Example 1.7. On a more profound level, quantum physicists teach us that everything is governed by the laws of probability. They toss around terms like the Schrödinger wave equation and Heisenberg's uncertainy principle, which are much too difficult for most of us to understand, but one thing they do mean is that the fundamental laws of physics can only be stated in terms of probabilities. And the fact that Newton's deterministic laws of physics are still useful can also be attributed to results from the theory of probabilities. [17, p 2]

1.8. Most people have preconceived notions of randomness that often differ substantially from true randomness. Truly random data sets often have unexpected properties that go against intuitive thinking. These properties can be used to test whether data sets have been tampered with when suspicion arises. [22, p 191]

- [13, p 174]: "people have a very poor conception of randomness; they do not recognize it when they see it and they cannot produce it when they try"

Example 1.9. Apple ran into an issue with the random shuffling method it initially employed in its iPod music players: true randomness sometimes produces repetition, but when users heard the same song or songs by the same artist played back-to-back, they believed the shuffling wasn't random. And so the company made the feature "less random to make it feel more random," said Apple founder Steve Jobs. [13, p 175]

1.2 Background on some frequently used examples

Probabilists love to play with coins and dice. We like the idea of tossing coins, rolling dice, and drawing cards as experiments that have equally likely outcomes.

1.10. ***Coin flipping*** or ***coin tossing*** is the practice of throwing a coin in the air to observe the outcome.

When a **coin** is tossed, it does not necessarily fall heads or tails; it can roll away or stand on its edge. Nevertheless, we shall agree to regard "**head**" (**H**) and "**tail**" (**T**) as the only possible outcomes of the experiment. [4, p 7]

- Typical experiment includes
  - "Flip a coin N times. Observe the sequence of heads and tails" or "Observe the number of heads."

1.11. Historically, ***dice*** is the plural of ***die***, but in modern standard English dice is used as both the singular and the plural. [Excerpted from Compact Oxford English Dictionary.]

- Usually assume six-sided dice
- Usually observe the number of dots on the side facing upwards.

1.12. A complete set of **cards** is called a pack or **deck**.

(a) The subset of cards held at one time by a player during a game is commonly called a **hand**.

(b) For most games, the cards are assembled into a deck, and their order is randomized by **shuffling**.

(c) A standard deck of 52 cards in use today includes thirteen ranks of each of the four French suits.

- The four suits are called spades (♠), clubs (♣), hearts (♡), and diamonds (♢). The last two are red, the first two black.

(d) There are thirteen face values ($2, 3, \ldots, 10$, jack, queen, king, ace) in each suit.

- Cards of the same face value are called of the same **kind**.
- "court" or face card: a king, queen, or jack of any suit.

(e) For our purposes, playing bridge means distributing the cards to four players so that each receives thirteen cards. Playing poker, by definition, means selecting five cards out of the pack.

1.3 A Glimpse at Probability

1.13. Probabilities are used in situations that involve ***randomness***. A ***probability*** is a number used to describe how likely something is to occur, and ***probability*** (without indefinite article) is the study of probabilities. It is the art of ***being certain of how uncertain you are***. [17, p 2–4] If an event is certain to happen, it is given a probability of 1. If it is certain not to happen, it has a probability of 0. [7, p 66]

1.14. Probabilities can be expressed as fractions, as decimal numbers, or as percentages. If you toss a coin, the probability to get heads is 1/2, which is the same as 0.5, which is the same as 50%. There are no explicit rules for when to use which notation.

- In daily language, proper fractions are often used and often expressed, for example, as "one in ten" instead of 1/10 ("one tenth"). This is also natural when you deal with equally likely outcomes.
- **Decimal numbers** are more common in technical and scientific reporting when probabilities are calculated from data. Percentages are also common in daily language and often with "chance" replacing "probability."
- Meteorologists, for example, typically say things like "there is a 20% chance of rain." The phrase "the probability of rain is 0.2" means the same thing.
- When we deal with probabilities from a theoretical viewpoint, we always think of them as numbers between 0 and 1, not as percentages.
- See also 3.7.

[17, p 10]

Definition 1.15. Important terms [12]:

(a) An activity or procedure or observation is called a **random experiment** if its outcome cannot be predicted precisely because the conditions under which it is performed cannot be predetermined with sufficient accuracy and completeness.

- The term "experiment" is to be construed loosely. We do not intend a laboratory situation with beakers and test tubes.
- Tossing/flipping a coin, rolling a die, and drawing a card from a deck are some examples of random experiments.

(b) A random experiment may have several separately identifiable **outcomes**. We define the **sample space** Ω as a collection of all possible (separately identifiable) outcomes/results/measurements of a random experiment. Each outcome (ω) is an element, or sample point, of this space.

- Rolling a dice has six possible identifiable outcomes $(1, 2, 3, 4, 5,$ and $6)$.

(c) **Events** are sets (or classes) of outcomes meeting some specifications.

- Any event is a subset of Ω.
- Intuitively, an event is a statement about the outcome(s) of an experiment.
- For our class, it may be less confusing to allow event A to be any collection of outcomes (, i.e. any subset of Ω).
  - In more advanced courses, when we deal with uncountable Ω, we limit our interest to only some subsets of Ω. Technically, the collection of these subsets must form a σ-algebra.

The goal of probability theory is to compute the probability of various events of interest. Hence, we are talking about a set function which is defined on (some class of) subsets of Ω.

Example 1.16. The statement "when a coin is tossed, the probability to get heads is l/2 (50%)" is a precise statement.

(a) It tells you that you are as likely to get heads as you are to get tails.

(b) Another way to think about probabilities is in terms of **average long-term behavior**. In this case, if you toss the coin repeatedly, in the long run you will get *roughly* 50% heads and 50% tails.

Of this you can be certain. We can make an even more precise statement by calculating the relative frequency of heads in n tosses of a fair coin. (See 1.18 and Example 1.19 below.) In any case, although the outcome of a random experiment is unpredictable, there is a **statistical regularity** about the outcomes. What you cannot be certain of is how the next toss will come up. [17, p 4]

1.17. ***Long-run frequency interpretation***: If the probability of an event A in some actual physical experiment is p, then we believe that if the experiment is repeated independently over and over again, then in the long run the event A will happen $100p\%$ of the time. In other words, If we repeat an experiment a large number of times then the fraction of times the event A occurs will be close to $P(A)$.

Definition 1.18. Let A be one of the events of a random experiment. If we conduct a sequence of n independent trials of this experiment, and if the event A occurs in $N(A, n)$ out of these n trials, then the fraction

$$\frac{N(A,n)}{n} := r_n(A)$$

is called the **relative frequency** of the event A in these n trials. Later on in Section 14, we will arrive at an important fact:

$$P(A) \text{ “=” } \lim_{n\to\infty} r_n(A)$$

as a theorem called the law of large numbers (LLN).

Example 1.19. Probability theory predicts that relative frequency of heads in n tosses of a fair coin that we considered in Example 1.16 will converge to $1/2$ as n tends to infinity. Furthermore, probability theory (LLN) also tells us that for each k, the relative

frequency of having exactly k heads in 100 tosses should be close to

$$\frac{100!}{k!(100-k)!}\frac{1}{2^{100}}.$$

Using the central limit theorem which will be presented in Section 14.2, we will see that the above expression is approximately equal to

$$\frac{1}{5\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{k-50}{5}\right)^2\right)$$

1.20. In terms of practical range, probability theory is comparable with ***geometry***; both are branches of applied mathematics that are directly linked with the problems of daily life. But while pretty much anyone can call up a natural feel for geometry to some extent, many people clearly have trouble with the development of a good intuition for probability.

- Probability and intuition do not always agree. ***In no other branch of mathematics is it so easy to make mistakes as in probability theory.***
- Students facing difficulties in grasping the concepts of probability theory might find comfort in the idea that even the genius Leibniz, the inventor of differential and integral calculus along with Newton, had difficulties in calculating the probability of throwing 11 with one throw of two dice.

[22, p 4]

2 Review of Set Theory

2.1. If ω is a member of a set A, we write $\omega \in A$.

Definition 2.2. Basic set operations (set algebra)

- Complementation: $A^c = \{\omega : \omega \notin A\}$.
- Union: $A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$
  - Here "or"is inclusive; i.e., if $\omega \in A$, we permit ω to belong either to A or to B or to both.
- Intersection: $A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$
  - Hence, $\omega \in A$ if and only if ω belongs to both A and B.
  - $A \cap B$ is sometimes written simply as AB.
- The ***set difference*** operation is defined by $B \setminus A = B \cap A^c$.
  - $B \setminus A$ is the set of $\omega \in B$ that do not belong to A.
  - When $A \subset B$, $B \setminus A$ is called the complement of A in B.

2.3. Basic Set Identities:

- Idempotence: $(\mathrm{A}^c)^c = \mathrm{A}$
- Commutativity (symmetry):

$$A \cup B = B \cup A \,,\; A \cap B = B \cap A$$

- Associativity:

- $A \cap (B \cap C) = (A \cap B) \cap C$
- $A \cup (B \cup C) = (A \cup B) \cup C$

- Distributivity
  - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

- **de Morgan laws**
  - $(A \cup B)^c = A^c \cap B^c$
  - $(A \cap B)^c = A^c \cup B^c$

2.4. Venn diagram is very useful in set theory. Many identities can be read out simply by examining Venn diagrams.

2.5. Disjoint Sets:

- Sets A and B are said to be ***disjoint*** ($A \perp B$) if and only if $A \cap B = \emptyset$. (They do not share member(s).)
- A collection of sets $(A_i : i \in I)$ is said to be ***pairwise disjoint*** or mutually exclusive [9, p. 9] if and only if $\mathrm{A}_i \cap \mathrm{A}_j = \emptyset$ when $i \neq j$.

2.6. For a set of sets, to avoid the repeated use of the word "set", we will call it a **collection/class/family** of sets.

Definition 2.7. Given a set S, a collection $\Pi = (A_\alpha : \alpha \in I)$ of subsets[1] of S is said to be a ***partition*** of S if

(a) $S = \bigcup A_{\alpha \in I}$ and

(b) For all $i \neq j$, $A_i \perp A_j$ (pairwise disjoint).

Remarks:

- The subsets A_α, $\alpha \in I$ are called the **parts** of the partition.
- A part of a partition may be empty, but usually there is no advantage in considering partitions with one or more empty parts.

[1] In this case, the subsets are indexed or labeled by α taking values in an index or label set I

- If a collection $\Pi = (A_\alpha : \alpha \in I)$ is a partition of Ω, then for any set B, the collection $(B \cap A_\alpha : \alpha \in I)$ is a partition of B. In other words, any set B can be expressed as $B = \bigcup_\alpha (B \cap A_i)$ where the union is a disjoint union.

Example 2.8 (slide:maps)**.**

Example 2.9. Let E be the set of students taking ECS315 in 2011.

Definition 2.10. The ***cardinality*** (or size) of a collection or set A, denoted $|A|$, is the number of elements of the collection. This number may be finite or infinite.

- An infinite set A is said to be ***countable*** if the elements of A can be enumerated or listed in a sequence: $a_1, a_2, \ldots$.
  - Empty set and finite sets are also said to be countable.
- By a ***countably infinite*** set, we mean a countable set that is not finite. Examples of such sets include
  - the set $\mathbb{N} = \{1, 2, , 3, \ldots\}$ of natural numbers,
  - the set $\{2k : k \in \mathbb{N}\}$ of all even numbers,
  - the set $\{2k + 1 : k \in \mathbb{N}\}$ of all odd numbers,
  - the set $\mathbb{Z}$ of integers,

- the set $\mathbb{Q}$ of all rational numbers,
- the set $\mathbb{Q}^+$ of positive rational numbers,
- the set of all finite-length sequences of natural numbers,
- the set of all finite subsets of the natural numbers.

- A ***singleton*** is a set with exactly one element.
  - Ex. $\{1.5\}$, $\{.8\}$, $\{\pi\}$.
  - *Caution*: Be sure you understand the difference between the outcome -8 and the event $\{-8\}$, which is the set consisting of the single outcome -8.

2.11. We can categorize sets according to their cardinality:

Example 2.12. Example of uncountable sets[2]:

- $\mathbb{R} = (-\infty, \infty)$
- interval $[0, 1]$
- interval $(0, 1]$
- $(2, 3) \cup [5, 7)$
- power set of $\mathbb{N}$ (denoted by $2^{\mathbb{N}}$) = a collection of all subsets of $\mathbb{N}$

Definition 2.13. Probability theory renames some of the terminology in set theory. See Table 1 and Table 2.

- Sometimes, ω's are called states, and Ω is called the state space.

Set Theory	Probability Theory
Set	Event
Universal set	Sample Space (Ω)
Element	Outcome (ω)

Table 1: The terminology of set theory and probability theory

	Event Language
A	A occurs
A^c	A does not occur
$A \cup B$	Either A or B occur
$A \cap B$	Both A and B occur

Table 2: Event Language

2.14. Because of the mathematics required to determine probabilities, probabilistic methods are divided into two distinct types, discrete and continuous. A discrete approach is used when the number of experimental outcomes is finite (or infinite but countable). A continuous approach is used when the outcomes are continuous (and therefore infinite). It will be important to keep in mind which case is under consideration since otherwise, certain paradoxes may result.

[2]We use a technique called diagonal argument to prove that a set is not countable and hence uncountable.

3 Classical Probability

Classical probability, which is based upon the ratio of the number of outcomes favorable to the occurrence of the event of interest to the total number of possible outcomes, provided most of the probability models used prior to the 20th century. It is the first type of probability problems studied by mathematicians, most notably, Frenchmen Fermat and Pascal whose 17th century correspondence with each other is usually considered to have started the systematic study of probabilities. [17, p 3] Classical probability remains of importance today and provides the most accessible introduction to the more general theory of probability.

Definition 3.1. Given a finite sample space Ω, the ***classical probability*** of an event A is

$$P(A) = \frac{\|A\|}{\|\Omega\|} \tag{1}$$

[6, Defn. 2.2.1 p 58]. In traditional language, a probability is a fraction in which the bottom represents the number of possible outcomes, while the number on top represents the number of outcomes in which the event of interest occurs.

- Assumptions: When the following are not true, do not calculate probability using (1).
  - Finite Ω: The number of possible outcomes is finite.
  - Equipossibility: The outcomes have equal probability of occurrence.
- The bases for identifying equipossibility were often
  - physical symmetry (e.g. a well-balanced die, made of homogeneous material in a cubical shape) or
  - a balance of information or knowledge concerning the various possible outcomes.
- Equipossibility is meaningful only for finite sample space, and, in this case, the evaluation of probability is accomplished through the definition of classical probability.

- We will NOT use this definition beyond this section. We will soon introduce a formal definition in Section 5.

Example 3.2 (Slide)**.** In drawing a card from a deck, there are 52 equally likely outcomes, 13 of which are diamonds. This leads to a probability of 13/52 or 1/4.

3.3. Basic properties of classical probability: From Definition 3.1, we can easily verified the properties below. Because we will not rely on Definition 3.1 beyond this section, we will not worry about how to prove these properties. We will prove the same properties in a more general setting later in Section 5.

- $P(A) \geq 0$
- $P(\Omega) = 1$
- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ which comes directly from
$$|A \cup B| = |A| + |B| - |A \cap B|.$$
- $A \perp B \Rightarrow P(A \cup B) = P(A) + P(B)$
- Suppose $\Omega = \{\omega_1, \ldots, \omega_n\}$ and $P(\{\omega_i\}) = \frac{1}{n}$. Then $P(A) = \sum_{\omega \in A} P(\{\omega\})$.
    - The probability of an event is equal to the sum of the probabilities of its component outcomes because outcomes are mutually exclusive

3.4. In classical probability theory defined by Definition 3.1,

$$A \perp B \text{ is equivalent to } P(A \cap B) = 0.$$

However, In general probability theory, the above statement is NOT true.

Example 3.5 (Slides)**.** When rolling two dice, there are 36 (equiprobable) possibilities.

$$P[\text{sum of the two dice} = 5] = 4/36.$$

Example 3.6. *Chevalier de Mere's Scandal of Arithmetic*:

> Which is more likely, obtaining at least one six in 4 tosses of a fair die (event A), or obtaining at least one double six in 24 tosses of a pair of dice (event B)?

We have

$$P(A) = \frac{6^4 - 5^4}{6^4} = 1 - \left(\frac{5}{6}\right)^4 \approx .518$$

and

$$P(B) = \frac{36^{24} - 35^{24}}{36^{24}} = 1 - \left(\frac{35}{36}\right)^{24} \approx .491.$$

Therefore, the first case is more probable.

Remark 1: Probability theory was originally inspired by gambling problems. In 1654, Chevalier de Mere invented a gambling system which bet even money[3] on event B above. However, when he began losing money, he asked his mathematician friend Pascal to analyze his gambling system. Pascal discovered that the Chevalier's system would lose about 51 percent of the time. Pascal became so interested in probability and together with another famous mathematician, Pierre de Fermat, they laid the foundation of probability theory. [U-X-L Encyclopedia of Science]

Remark 2: de Mere originally claimed to have discovered a *contradiction in arithmetic*. De Mere correctly knew that it was advantageous to wager on occurrence of event A, but his experience as gambler taught him that it was not advantageous to wager on occurrence of event B. He calculated $P(A) = 1/6 + 1/6 + 1/6 + 1/6 = 4/6$ and similarly $P(B) = 24 \times 1/36 = 24/36$ which is the same as $P(A)$. He mistakenly claimed that this evidenced a contradiction to the arithmetic law of proportions, which says that $\frac{4}{6}$ should be the same as $\frac{24}{36}$. Of course we know that he could not

[3]Even money describes a wagering proposition in which if the bettor loses a bet, he or she stands to lose the same amount of money that the winner of the bet would win.

simply add up the probabilities from each tosses. (By De Meres logic, the probability of at least one head in two tosses of a fair coin would be $2 \times 0.5 = 1$, which we know cannot be true). [22, p 3]

Definition 3.7. In the world of gambling, probabilities are often expressed by ***odds***. To say that the odds are n:1 *against* the event A means that it is n times as likely that A does not occur than that it occurs. In other words, $P(A^c) = nP(A)$ which implies $P(A) = \frac{1}{n+1}$ and $P(A^c) = \frac{n}{n+1}$.

"Odds" here has nothing to do with even and odd numbers. The odds also mean what you will win, in addition to getting your stake back, should your guess prove to be right. If I bet \$1 on a horse at odds of 7:1, I get back \$7 in winnings plus my \$1 stake. The bookmaker will break even in the long run if the probability of that horse winning is 1/8 (not 1/7). Odds are "even" when they are 1:1 - win \$1 and get back your original \$1. The corresponding probability is 1/2.

3.8. It is important to remember that classical probability relies on the assumption that the outcomes are ***equally likely***.

Example 3.9. ***Mistake*** made by the famous French mathematician Jean Le Rond d'Alembert (18th century) who is an author of several works on probability:

"The number of heads that turns up in those two tosses can be 0, 1, or 2. Since there are three outcomes, the chances of each must be 1 in 3."

4 Enumeration / Combinatorics / Counting

There are many probability problems, especially those concerned with gambling, that can ultimately be reduced to questions about cardinalities of various sets. ***Combinatorics*** is the *study of systematic counting methods*, which we will be using to find the cardinalities of various sets that arise in probability.

4.1. ***Addition Principle (Rule of sum)***:

- When there are m cases such that the ith case has n_i options, for $i = 1, \ldots, m$, and no two of the cases have any options in common, the total number of options is $n_1 + n_2 + \cdots + n_m$.
- In set-theoretic terms, suppose that a finite set S can be partitioned into (pairwise disjoint parts) $S_1, S_2, \ldots, S_m$. Then,

$$|S| = |S_1| + |S_2| + \cdots + |S_m|.$$

- The art of applying the addition principle is to partition the set S to be counted into "manageable parts"; that is, parts which we can readily count. But this statement needs to be qualified. If we partition S into too many parts, then we may have defeated ourselves. For instance, if we partition 8 into parts each containing only one element, then applying the addition principle is the same as counting the number of parts, and this is basically the same as listing all the objects of S. Thus, a more appropriate description is that the art of applying the addition principle is to partition the set S into not too many manageable parts.[1, p 28]

Example 4.2. [1, p 28] Suppose we wish to find the number of different courses offered by SIIT. We partition the courses according to the department in which they are listed. Provided there is no cross-listing (cross-listing occurs when the same course is listed by more than one department), the number of courses offered by SIIT equals the sum of the number of courses offered by each department.

Example 4.3. [1, p 28] A student wishes to take either a mathematics course or a biology course, but not both. If there are four mathematics courses and three biology courses for which the student has the necessary prerequisites, then the student can choose a course to take in $4 + 3 = 7$ ways.

4.4. ***Multiplication Principle (Rule of product)***:

- When a procedure can be broken down into m steps, such that there are n_1 options for step 1, and such that after the completion of step $i - 1$ $(i = 2, \ldots, m)$ there are n_i options for step i, the number of ways of performing the procedure is $n_1 n_2 \cdots n_m$.
- In set-theoretic terms, if sets $S_1, \ldots, S_m$ are finite, then $|S_1 \times S_2 \times \cdots \times S_m| = |S_1| \cdot |S_2| \cdot \cdots \cdot |S_m|$.
- For k finite sets $A_1, ..., A_k$, there are $|A_1| \cdots |A_k|$ k-tuples of the form $(a_1, \ldots, a_k)$ where each $a_i \in A_i$.

Example 4.5. Let A, B, and C be finite sets. How many triples are there of the form (a,b,c), where $a \in A$, $b \in B$, $c \in C$?

Example 4.6. Suppose that a deli offers three kinds of bread, three kinds of cheese, four kinds of meat, and two kinds of mustard. How many different meat and cheese sandwiches can you make?

First choose the bread. For each choice of bread, you then have three choices of cheese, which gives a total of $3 \times 3 = 9$ bread/cheese combinations (rye/swiss, rye/provolone, rye/cheddar, wheat/swiss, wheat/provolone ... you get the idea). Then choose among the four kinds of meat, and finally between the

two types of mustard or no mustard at all. You get a total of $3 \times 3 \times 4 \times 3 = 108$ different sandwiches.

Suppose that you also have the choice of adding lettuce, tomato, or onion in any combination you want. This choice gives another 2 x 2 x 2 = 8 combinations (you have the choice "yes" or "no" three times) to combine with the previous 108, so the total is now $108 \times 8 = 864$.

That was the multiplication principle. In each step you have several choices, and to get the total number of combinations, multiply. It is fascinating how quickly the number of combinations grow. Just add one more type of bread, cheese, and meat, respectively, and the number of sandwiches becomes 1,920. It would take years to try them all for lunch. [17, p 33]

Example 4.7. In 1961, Raymond Queneau, a French poet and novelist, wrote a book called *One Hundred Thousand Billion Poems*. The book has ten pages, and each page contains a sonnet, which has 14 lines. There are cuts between the lines so that each line can be turned separately, and because all lines have the same rhyme scheme and rhyme sounds, any such combination gives a readable sonnet. The number of sonnets that can be obtained in this way is thus 10^{14} which is indeed a hundred thousand billion. Somebody has calculated that it would take about 200 million years of nonstop reading to get through them all. [17, p 34]

Example 4.8. [1, p 29–30] Determine the number of positive integers that are factors of the number

$$3^4 \times 5^2 \times 11^7 \times 13^8.$$

The numbers 3,5,11, and 13 are prime numbers. By the fundamental theorem of arithmetic, each factor is of the form

$$3^i \times 5^j \times 11^k \times 13^\ell,$$

where $0 \leq i \leq 4$, $0 \leq j \leq 2$, $0 \leq k \leq 7$, and $0 \leq \ell \leq 8$. There are five choices for i, three for j, eight for k, and nine for ℓ. By the multiplication principle, the number of factors is

$$5 \times 3 \times 8 \times 9 = 1080.$$

4.9. *Subtraction Principle*: Let A be a set and let S be a larger set containing S. Then

$$|A| = |S| - |S \setminus A|$$

- When S is the same as Ω, we have $|A| = |S| - |A^c|$
- Using the subtraction principle makes sense only if it is easier to count the number of objects in S and in $S \setminus A$ than to count the number of objects in A.

4.10. *Division Principle (Rule of quotient)*: When a finite set S is partitioned into equal-sized parts of m elements each, there are $\frac{|S|}{m}$ parts.

4.1 Four kinds of counting problems

4.11. Choosing objects from a collection is also called **sampling**, and the chosen objects are known as a **sample**. The four kinds of counting problems are [9, p 34]:

(a) ordered sampling of r out of n items with replacement: n^r;

(b) ordered sampling of $r \leq n$ out of n items without replacement: $(n)_r$;

(c) unordered sampling of $r \leq n$ out of n items without replacement: $\binom{n}{r}$;

(d) unordered sampling of r out of n items with replacement: $\binom{n+r-1}{r}$.

- See 4.18 for "bars and stars" argument.

4.12. Given a set of n distinct items/objects, select a distinct **ordered**[4] sequence (word) of length r drawn from this set.

(a) ***Ordered Sampling with replacement***: $\mu_{n,r} = n^r$

- Meaning

[4]Different sequences are distinguished by the order in which we choose objects.

- Ordered sampling of r out of n items with replacement.
  - An object can be chosen repeatedly.

- $\mu_{n,1} = n$
- $\mu_{1,r} = 1$
- Examples:
  - From a deck of n cards, we draw r cards with replacement; i.e., we draw each card, make a note of it, put the card back in the deck and re-shuffle the deck before choosing the next card. How many different sequences of r cards can be drawn in this way?

  - Suppose A is a finite set, then the cardinality of its power set is $\left|2^A\right| = 2^{|A|}$.
  - There are 2^r binary strings/sequences of length r.

(b) ***Ordered Sampling without replacement***:

$$\begin{aligned}(n)_r &= \prod_{i=0}^{r-1}(n-i) = \frac{n!}{(n-r)!} \\ &= \underbrace{n \cdot (n-1) \cdots (n-(r-1))}_{r \text{ terms}}; \quad r \leq n\end{aligned}$$

- Meaning
  - Ordered sampling of $r \leq n$ out of n items without replacement.
    - Once we choose an object, we remove that object from the collection and we cannot choose it again.
  - "the number of possible r-**permutations** of n distinguishable objects"

- the number of sequences[5] of size r drawn from an alphabet of size n without replacement.

- Example: $(3)_2 = 3 \times 2 = 6 =$ the number of sequence of size 2 drawn from an alphabet of size $= 3$ without replacement.

  Suppose the alphabet set is {A, B, C}. We can list all sequences of size 2 drawn from {A, B, C} without replacement:

  A B
  A C
  B A
  B C
  C A
  C B

- Example: From a deck of 52 cards, we draw a hand of 5 cards without replacement (drawn cards are not placed back in the deck). How many hands can be drawn in this way?

- For integers r, n such that $r > n$, we have $(n)_r = 0$.
- Extended definition: The definition in product form

$$(n)_r = \prod_{i=0}^{r-1} (n-i) = \underbrace{n \cdot (n-1) \cdots (n-(r-1))}_{\text{r terms}}$$

  can be extended to *any real number* n and a non-negative integer r. We define $(n)_0 = 1$. (This makes sense because we usually take the empty product to be 1.)

[5] Elements in a sequence are ordered.

- $(n)_1 = n$
- $(n)_r = (n-(r-1))(n)_{r-1}$. For example, $(7)_5 = (7-4)(7)_4$.
- $(1)_r = \begin{cases} 1, & \text{if r} = 1 \\ 0, & \text{if r} > 1 \end{cases}$

4.13. ***Factorial and Permutation***: The number of arrangements (permutations) of $n \geq 0$ distinct items is $(n)_n = n!$.

- For any integer n greater than 1, the symbol $n!$, pronounced "n factorial," is defined as the product of all positive integers less than or equal to n.
- $0! = 1! = 1$
- $n! = n(n-1)!$
- $n! = \int_0^\infty e^{-t} t^n dt$
- Computation:
  (a) `MATLAB`: Use `factorial(n)`. Since double precision numbers only have about 15 digits, the answer is only accurate for $n \leq 21$. For larger n, the answer will have the right magnitude, and is accurate for the first 15 digits.
  (b) Google's web search box built-in calculator: `n!`
- Meaning: The number of ways that n distinct objects can be ordered.
  - A special case of ordered sampling without replacement where $r = n$.
- In `MATLAB`, use `perms(v)`, where v is a row vector of length n, to creates a matrix whose rows consist of all possible permutations of the n elements of v. (So the matrix will contain $n!$ rows and n columns.)
- Example: In `MATLAB`, `perms([3 4 7])` gives

7 4 3
7 3 4
4 7 3
4 3 7
3 4 7
3 7 4

Similarly, `perms('abcd')` gives

dcba dcab dbca dbac dabc dacb
cdba cdab cbda cbad cabd cadb
bcda bcad bdca bdac badc bacd
acbd acdb abcd abdc adbc adcb

- Approximation: Stirling's Formula [5, p. 52]:

$$n! \approx \sqrt{2\pi n} n^n e^{-n} = \left(\sqrt{2\pi e}\right) e^{\left(n+\frac{1}{2}\right)\ln\left(\frac{n}{e}\right)}. \tag{2}$$

  - The sign $\approx$ can be replaced by $\sim$ to emphasize that the ratio of the two sides converges to unity as $n \to \infty$.
  - $\ln n! = n \ln n - n + o(n)$

4.14. Ratio:

$$\begin{aligned}\frac{(n)_r}{n^r} &= \frac{\prod_{i=0}^{r-1}(n-i)}{\prod_{i=0}^{r-1}(n)} = \prod_{i=0}^{r-1}\left(1-\frac{i}{n}\right) \\ &\approx \prod_{i=0}^{r-1}\left(e^{-\frac{i}{n}}\right) = e^{-\frac{1}{n}\sum_{i=0}^{r-1} i} = e^{-\frac{r(r-1)}{2n}} \\ &\approx e^{-\frac{r^2}{2n}}\end{aligned}$$

In fact, when $r - 1 < \frac{n}{2}$, (A.11) gives

$$e^{\frac{1}{2}\frac{r(r-1)}{n}\frac{3n+2r-1}{3n}} \le \prod_{i=1}^{r-1}\left(1-\frac{i}{n}\right) \le e^{\frac{1}{2}\frac{r(r-1)}{n}}.$$

See also 4.20.

Example 4.15. (Slides) See Example 4.25.

4.16. ***Binomial coefficient***:

$$\binom{n}{r} = \frac{(n)_r}{r!} = \frac{n!}{(n-r)!r!}$$

(a) Read "n choose r".

(b) Meaning:

(i) **Unordered sampling** of $r \leq n$ out of n items **without replacement**

(ii) The number of subsets of size r that can be formed from a set of n elements (without regard to the order of selection).

(iii) The number of combinations of n objects selected r at a time.

(iv) the number of k**-combinations** of n objects.

(v) The number of (unordered) sets of size r drawn from an alphabet of size n without replacement.

(c) Computation:

(i) `MATLAB`:

- `nchoosek(n,r)`, where n and r are nonnegative integers, returns $\binom{n}{r}$.
- `nchoosek(v,r)`, where v is a row vector of length n, creates a matrix whose rows consist of all possible combinations of the n elements of v taken r at a time. The matrix will contains $\binom{n}{r}$ rows and r columns.

- Example: `nchoosek('abcd',2)` gives

  ab
  ac
  ad
  bc
  bd
  cd

(ii) Use `combin(n,r)` in `Mathcad`. However, to do symbolic manipulation, use the factorial definition directly.

(iii) In `Maple`, use $\binom{n}{r}$ directly.

(iv) Google's web search box built-in calculator: `n choose k`

(d) Reflection property: $\binom{n}{r} = \binom{n}{n-r}$.

(e) $\binom{n}{n} = \binom{n}{0} = 1$.

(f) $\binom{n}{1} = \binom{n}{n-1} = n$.

(g) $\binom{n}{r} = 0$ if $n < r$ or r is a negative integer.

(h) $\max\limits_{r} \binom{n}{r} = \binom{n}{\lfloor \frac{n+1}{2} \rfloor}$.

Example 4.17. In bridge, 52 cards are dealt to four players; hence, each player has 13 cards. The order in which the cards are dealt is not important, just the final 13 cards each player ends up with. How many different bridge games can be dealt? (Answer: 53,644,737,765,488,792,839,237,440,000)

4.18. The **bars and stars argument**:

- Example: Find all nonnegative integers x_1, x_2, x_3 such that

$$x_1 + x_2 + x_3 = 3.$$

$0+0+3$	1 1 1
$0+1+2$	1 1 1
$0+2+1$	1 1 1
$0+3+0$	1 1 1
$1+0+2$	1 1 1
$1+1+1$	1 1 1
$1+2+0$	1 1 1
$2+0+1$	1 1 1
$2+1+0$	1 1 1
$2+0+0$	1 1 1

- There are $\binom{n+r-1}{r} = \binom{n+r-1}{n-1}$ distinct vector $x = x_1^n$ of nonnegative integers such that $x_1 + x_2 + \cdots + x_n = r$. We use $n-1$ bars to separate r 1's.

  (a) Suppose we further require that the x_i are strictly positive ($x_i \geq 1$), then there are $\binom{r-1}{n-1}$ solutions.

  (b) ***Extra Lower-bound Requirement***: Suppose we further require that $x_i \geq a_i$ where the a_i are some given nonnegative integers, then the number of solution is

$$\binom{r - (a_1 + a_2 + \cdots + a_n) + n - 1}{n-1}.$$

  Note that here we work with equivalent problem: $y_1 + y_2 + \cdots + y_n = r - \sum_{i=1}^n a_i$ where $y_i \geq 0$.

- Consider the distribution of $r = 10$ indistinguishable balls into $n = 5$ distinguishable cells. Then, we only concern with the number of balls in each cell. Using $n - 1 = 4$ bars, we can divide $r = 10$ stars into $n = 5$ groups. For example, ****|***||**|* would mean (4,3,0,2,1). In general, there are $\binom{n+r-1}{r}$ ways of arranging the bars and stars.

4.19. ***Unordered sampling with replacement***: There are n items. We sample r out of these n items with replacement.

Because the order in the sequences is not important in this kind of sampling, two samples are distinguished by the number of each item in the sequence. In particular, Suppose r letters are drawn with replacement from a set $\{a_1, a_2, \ldots, a_n\}$. Let x_i be the number of a_i in the drawn sequence. Because we sample r times, we know that, for every sample, $x_1 + x_2 + \cdots x_n = r$ where the x_i are non-negative integers. Hence, there are $\binom{n+r-1}{r}$ possible unordered samples with replacement.

4.20. A random sample of size r with replacement is taken from a population of n elements. The probability of the event that in the sample no element appears twice (that is, no repetition in our sample) is

$$\frac{(n)_r}{n^r}.$$

From 4.14, we have

$$\frac{(n)_r}{n^r} \approx e^{-\frac{r(r-1)}{2n}}.$$

Therefore, the probability that at least one element appears twice is

$$p_u(n, r) = 1 - \frac{(n)_r}{n^r} \approx 1 - e^{-\frac{r(r-1)}{2n}}.$$

- From the approximation, to have $p_u(n, r) = p$, we need

$$r \approx \frac{1}{2} + \frac{1}{2}\sqrt{1 - 8n \ln(1-p)}.$$

4.2 Binomial Theorem and Multinomial Theorem

4.21. ***Binomial theorem***: Sometimes, the number $\binom{n}{r}$ is called a ***binomial coefficient*** because it appears as the coefficient of $x^r y^{n-r}$ in the expansion of the binomial $(x+y)^n$. More specifically, for any positive integer n, we have,

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r} \tag{3}$$

To see how we get (3), let's consider a smaller case of $n = 3$. The expansion of $(x+y)^3$ can be found using combinatorial reasoning

instead of multiplying the three terms out. When $(x+y)^3 = (x+y)(x+y)(x+y)$ is expanded, all products of a term in the first sum, a term in the second sum, and a term in the third sum are added. Terms of the form x^3, x^2y, xy^2, and y^3 arise. To obtain a term of the form x^3, an x must be chosen in each of the sums, and this can be done in only one way. Thus, the x^3 term in the product has a coefficient of 1. To obtain a term of the form x^2y, an x must be chosen in two of the three sums (and consequently a y in the other sum). Hence. the number of such terms is the number of 2-combinations of three objects, namely, $\binom{3}{2}$. Similarly, the number of terms of the form xy^2 is the number of ways to pick one of the three sums to obtain an x (and consequently take a y from each of the other two terms). This can be done in $\binom{3}{1}$ ways. Finally, the only way to obtain a y^3 term is to choose the y for each of the three sums in the product, and this can be done in exactly one way. Consequently. it follows that

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3.$$

Now, let's state a combinatorial proof of the binomial theorem (3). The terms in the product when it is expanded are of the form $x^r y^{n-r}$ for $r = 0, 1, 2, \ldots, n$. To count the number of terms of the form $x^r y^{n-r}$, note that to obtain such a term it is necessary to choose r xs from the n sums (so that the other $n-r$ terms in the product are ys). Therefore. the coefficient of $x^r y^{n-r}$ is $\binom{n}{r}$.

From (3), if we let $x = y = 1$, then we get another important identity:

$$\sum_{r=0}^{n} \binom{n}{r} = 2^n. \tag{4}$$

4.22. ***Multinomial Counting***: The ***multinomial coefficient*** $\binom{n}{n_1\, n_2 \cdots n_r}$ is defined as

$$\prod_{i=1}^{r} \binom{n - \sum_{k=0}^{i-1} n_k}{n_i} = \binom{n}{n_1} \cdot \binom{n-n_1}{n_2} \cdot \binom{n-n_1-n_2}{n_3} \cdots \binom{n_r}{n_r} = \frac{n!}{\prod_{i=1}^{r} n!}.$$

It is the number of ways that we can arrange $n = \sum_{i=1}^{r} n_i$ tokens when having r types of symbols and n_i indistinguishable copies/tokens of a type i symbol.

4.23. ***Multinomial Theorem***:

$$(x_1 + \ldots + x_r)^n = \sum \frac{n!}{i_1! i_2! \cdots i_r!} x_1^{i_1} x_2^{i_2} \cdots x_r^{i_r},$$

where the sum ranges over all ordered r-tuples of integers $i_1, \ldots, i_r$ satisfying the following conditions:

$$i_1 \geq 0, \ldots, i_r \geq 0, \quad i_1 + i_2 + \cdots + i_r = n.$$

More specifically, $(x_1 + \ldots + x_r)^n$ can be written as

$$\sum_{i_1=0}^{n} \sum_{i_2=0}^{n-i_1} \cdots \sum_{i_{r-1}=0}^{n-\sum_{j<r-1} i_j} \frac{n!}{\left(n - \sum_{k<n} i_k\right)! \prod_{k<n} i_k!} x_r^{n-\sum_{j<r} i_j} \prod_{k=1}^{r-1} x_k^{i_k}$$

When $r = 2$ this reduces to the binomial theorem.

4.3 Famous Examples

Example 4.24. *Probability of coincidence birthday*: Probability that there is at least two people who have the same birthday[6] in a group of r persons:

$$= \begin{cases} 1, & \text{if } r \geq 365, \\ 1 - \left(\underbrace{\frac{365}{365} \cdot \frac{364}{365} \cdot \cdots \cdot \frac{365-(r-1)}{365}}_{\text{r terms}} \right), & \text{if } 0 \leq r \leq 365 \end{cases}$$

[6]We ignore February 29 which only comes in leap years.

Example 4.25. It is surprising to see how quickly the probability in Example 4.24 approaches 1 as r grows larger.

Birthday Paradox: In a group of 23 randomly selected people, the probability that at least two will share a birthday (assuming birthdays are equally likely to occur on any given day of the year[7]) is about 0.5.

- At first glance it is surprising that the probability of 2 people having the same birthday is so large[8], since there are only 23 people compared with 365 days on the calendar. Some of the surprise disappears if you realize that there are $\binom{23}{2} = 253$ pairs of people who are going to compare their birthdays. [3, p. 9]

Remark: The group size must be at least 253 people if you want a probability > 0.5 that someone will have the same birthday as you. [3, Ex. 1.13] (The probability is given by $1 - \left(\frac{364}{365}\right)^r$.)

- A naive (but incorrect) guess is that $\lceil 365/2 \rceil = 183$ people will be enough. The "problem" is that many people in the group will have the same birthday, so the number of different birthdays is smaller than the size of the group.

Example 4.26. On late-night television's The Tonight Show with Johnny Carson, Carson was discussing the birthday problem in one of his famous monologues. At a certain point, he remarked to his audience of approximately 100 people: "Great! There must be someone here who was born on my birthday!" He was off by a long shot. Carson had confused two distinctly different probability problems: (1) the probability of one person out of a group of 100 people having the same birth date as Carson himself, and (2) the probability of any two or more people out of a group of 101 people having birthdays on the same day. [22, p 76]

[7] In reality, birthdays are not uniformly distributed. In which case, the probability of a match only becomes larger for any deviation from the uniform distribution. This result can be mathematically proved. Intuitively, you might better understand the result by thinking of a group of people coming from a planet on which people are always born on the same day.

[8] In other words, it was surprising that the size needed to have 2 people with the same birthday was so small.

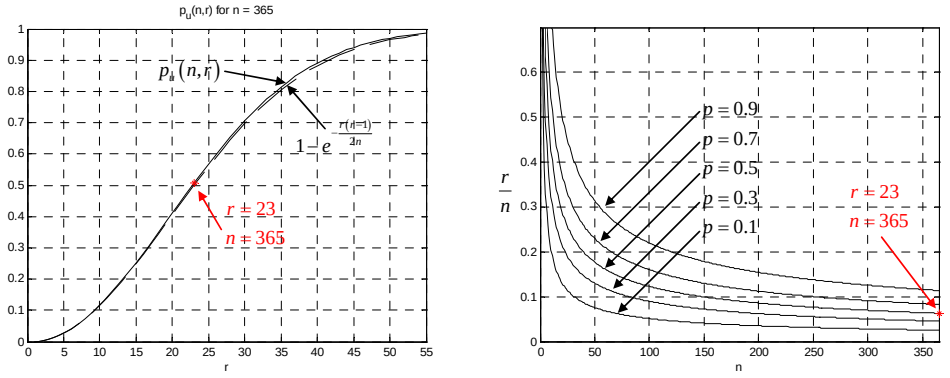

Figure 1: $p_u(n, r)$: The probability of the event that at least one element appears twice in random sample of size r with replacement is taken from a population of n elements.

Example 4.27. When you toss three dice, the chance of the sum being 10 is greater than the chance of the sum being 9.

- The Grand Duke of Tuscany "ordered" Galileo to explain a paradox arising in the experiment of tossing three dice [2]:

  > "Why, although there were an equal number of 6 partitions of the numbers 9 and 10, did experience state that the chance of throwing a total 9 with three fair dice was less than that of throwing a total of 10?"

- Partitions of sums 11, 12, 9 and 10 of the game of three fair dice:

1+4+6=11	1+5+6=12	3+3+3=9	1+3+6=10
2+3+6=11	2+4+6=12	1+2+6=9	1+4+5=10
2+4+5=11	3+4+5=12	1+3+5=9	2+2+6=10
1+5+5=11	2+5+5=12	1+4+4=9	2+3+5=10
3+3+5=11	3+3+6=12	2+2+5=9	2+4+4=10
3+4+4=11	4+4+4=12	2+3+4=9	3+3+3=10

  The partitions above are not equivalent. For example, from the addenda 1, 2, 6, the sum 9 can come up in $3! = 6$ different ways; from the addenda 2, 2, 5, the sum 9 can come up in $\frac{3!}{2!1!} = 3$ different ways; the sum 9 can come up in only one way from 3, 3, 3.

- ***Remarks***: Let X_i be the outcome of the ith dice and S_n be the sum $X_1 + X_2 + \cdots + X_n$.
  (a) $P[S_3 = 9] = P[S_3 = 12] = \frac{25}{6^3} < \frac{27}{6^3} = P[S_3 = 10] = P[S_3 = 11]$. Note that the difference between the two probabilities is only $\frac{1}{108}$.
  (b) The range of S_n is from n to $6n$. So, there are $6n-n+1 = 5n+1$ possible values.
  (c) The pmf of S_n is symmetric around its expected value at $\frac{n+6n}{2} = \frac{7n}{2}$.
    - $P[S_n = m] = P[S_n = 7n - m]$.

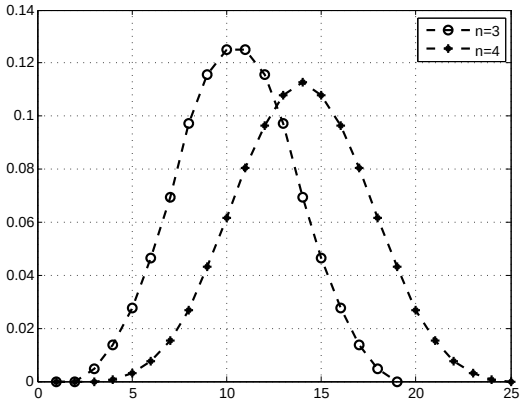

Figure 2: pmf of S_n for $n = 3$ and $n = 4$.

4.28. Further reading on combinatorial ideas: the pigeon-hole principle, inclusion-exclusion principle, generating functions and recurrence relations, and flows in networks.

4.4 Application: Success runs

Example 4.29. We are all familiar with "success runs" in many different contexts. For example, we may be or follow a tennis player and count the number of consecutive times the player's first serve is good. Or we may consider a run of forehand winners. A basketball player may be on a "**hot streak**" and hit his or her shots perfectly for a number of plays in row.

In all the examples, whether you should or should not be amazed by the observation depends on a lot of other information. There may be perfectly reasonable explanations for any particular success run. But we should be curious as to whether randomness could also be a perfectly reasonable explanation. Could the hot streak of a player simply be a snapshot of a random process, one that we particularly like and therefore pay attention to?

In 1985, cognitive psychologists Amos Taversky and Thomas Gilovich examined[9] the shooting performance of the Philadelphia 76ers, Boston Celtics and Cornell University's men's basketball team. They sought to discover whether a player's previous shot had any predictive effect on his or her next shot. Despite basketball fans' and players' widespread belief in hot streaks, the researchers found no support for the concept. (No evidence of nonrandom behavior.) [13, p 178]

4.30. Academics call the mistaken impression that a random streak is due to extraordinary performance the **hot-hand fallacy**. Much of the work on the hot-hand fallacy has been done in the context of sports because in sports, performance is easy to define and measure. Also, the rules of the game are clear and definite, data are plentiful and public, and situations of interest are replicated repeatedly. Not to mention that the subject gives academics a way to attend games and pretend they are working. [13, p 178]

Example 4.31. Suppose that two people are separately asked to toss a fair coin 120 times and take note of the results. Heads is noted as a "one" and tails as a "zero". The following two lists of compiled zeros and ones result

1 1 0 0 1 0 0 1 0 1 1 0 0 1 0 0 0 1 1 0 1 0 1 0 0 1 1 0 1 0
0 1 0 1 0 1 1 0 1 1 0 0 1 1 0 1 1 1 0 1 0 0 1 0 0 1 1 0 1 0
0 1 1 0 1 0 0 1 1 0 1 0 1 1 0 0 1 1 1 0 0 1 0 1 0 1 0 0 0 1
0 1 0 1 0 1 0 1 0 1 1 0 0 1 0 0 1 0 1 1 0 0 1 0 0 1 1 0 1 1

and

[9] "The Hot Hand in Basketball: On the Misperception of Random Sequences"

1 1 1 0 0 0 1 1 1 0 1 0 1 1 1 1 1 1 0 1 0 0 0 1 1 0 0 1 1 0
1 0 1 0 0 0 1 1 0 1 0 0 1 1 1 0 1 0 0 0 0 1 0 1 1 1 0 1 1 0
0 1 1 1 0 1 1 0 0 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 0 0 0 0 0 0
0 0 1 1 0 1 1 1 0 1 1 1 1 0 1 1 1 1 0 1 0 1 1 0 1 1 0 1 0 1

One of the two individuals has cheated and has fabricated a list of numbers without having tossed the coin. Which list is more likely be the fabricated list? [22, Ex. 7.1 p 42–43]

The answer is later provided in Example 4.37.

Definition 4.32. A **run** is a sequence of more than one consecutive identical outcomes, also known as a **clump**.

Definition 4.33. Let R_n represent the length of the longest run of heads in n independent tosses of a fair coin. Let $\mathcal{A}_n(x)$ be the set of (head/tail) sequences of length n in which the longest run of heads does not exceed x. Let $a_n(x) = \|\mathcal{A}_n(x)\|$.

Example 4.34. If a fair coin is flipped, say, three times, we can easily list all possible sequences:

HHH, HHT, HTH, HTT, THH, THT, TTH, TTT

and accordingly derive:

x	$P[R_3 = x]$	$a_3(x)$
0	1/8	1
1	4/8	4
2	2/8	7
3	1/8	8

4.35. Consider $a_n(x)$. Note that if $n \leq x$, then $a_n(x) = 2^n$ because any outcome is a favorable one. (It is impossible to get more than three heads in three coin tosses). For $n > x$, we can partition $\mathcal{A}_n(x)$ by the position k of the first tail. Observe that k must be $\leq x + 1$ otherwise we will have more than x consecutive heads in the sequence which contradicts the definition of $\mathcal{A}_n(x)$. For each $k \in \{1, 2, \ldots, x + 1\}$, the favorable sequences are in the form

$$\underbrace{\text{HH}\ldots\text{H}}_{k-1 \text{ heads}}\text{T}\ \underbrace{\text{XX}\ldots\text{X}}_{n-k \text{ positions}}$$

where, to keep the sequences in $\mathcal{A}_n(x)$, the last $n-k$ positions[10] must be in $\mathcal{A}_{n-k}(x)$. Thus,

$$a_n(x) = \sum_{k=1}^{x+1} a_{n-k}(x) \text{ for } n > x.$$

In conclusion, we have

$$a_n(x) = \begin{cases} \sum_{j=0}^{x} a_{n-j-1}(x), & n > x, \\ 2^n & n \leq x \end{cases}$$

[21]. The following MATLAB function calculates $a_n(x)$

```
function a = a_nx(n,x)
a = [2.^(1:x) zeros(1,n-x)];
a(x+1) = 1+sum(a(1:x));
for k = (x+2):n
    a(k) = sum(a((k-1-x):(k-1)));
end
a = a(n);
```

4.36. Similar technique can be used to contract $\mathcal{B}_n(x)$ defined as the set of sequences of length n in which the longest run of heads and the longest run of tails do not exceed x. To check whether a sequence is in $\mathcal{B}_n(x)$, first we convert it into sequence of S and D by checking each adjacent pair of coin tosses in the original sequence. S means the pair have same outcome and D means they are different. This process gives a sequence of length $n-1$. Observe that a string of $x-1$ consecutive S's is equivalent to a run of length x. This put us back to the earlier problem of finding $a_n(x)$ where the roles of H and T are now played by S and D, respectively. (The length of the sequence changes from n to $n-1$ and the max run length is $x-1$ for S instead of x for H.) Hence, $b_n(x) = \|\mathcal{B}_n(x)\|$ can be found by

$$b_n(x) = 2a_{n-1}(x-1)$$

[21].

[10]Strictly speaking, we need to consider the case when $n = x+1$ separately. In such case, when $k = x+1$, we have $\mathcal{A}_0(x)$. This is because the sequence starts with x heads, then a tail, and no more space left. In which case, this part of the partition has only one element; so we should define $a_0(x) = 1$. Fortunately, for $x \geq 1$, this is automatically satisfied in $a_n(x) = 2^n$.

Example 4.37. Continue from Example 4.31. We can check that in 120 tosses of a fair coin, there is a very large probability that at some point during the tossing process, a sequence of five or more heads or five or more tails will naturally occur. The probability of this is

$$\frac{2^{120} - b_{120}(4)}{2^{120}} \approx 0.9865.$$

0.9865. In contrast to the second list, the first list shows no such sequence of five heads in a row or five tails in a row. In the first list, the longest sequence of either heads or tails consists of three in a row. In 120 tosses of a fair coin, the probability of the longest sequence consisting of three or less in a row is equal to

$$\frac{b_{120}(3)}{2^{120}} \approx 0.000053,$$

which is extremely small indeed. Thus, the first list is almost certainly a fake. Most people tend to avoid noting long sequences of consecutive heads or tails. Truly random sequences do not share this human tendency! [22, Ex. 7.1 p 42–43]

ECS315 2011/1 Part II.1 Dr.Prapun

5 Probability Foundations

To study formal definition of probability, we start with the ***probability space*** $(\Omega, \mathcal{A}, P)$. Let Ω be an arbitrary space or set of points ω. Recall (from Definition 1.15) that, viewed probabilistically, a subset of Ω is an ***event*** and an element ω of Ω is a ***sample point***. Each event is a collection of outcomes which are elements of the sample space Ω.

The theory of probability focuses on collections of events, called event ***σ-algebras***, typically denoted by $\mathcal{A}$ (or $\mathcal{F}$), that contain all the events of interest[11] (regarding the random experiment $\mathcal{E}$) to us, and are such that we have knowledge of their likelihood of occurrence. The probability P itself is defined as a number in the range $[0, 1]$ associated with each event in $\mathcal{A}$.

Constructing the mathematical foundations of probability theory has proven to be a long-lasting process of trial and error. *The approach consisting of defining probabilities as relative frequencies in cases of repeatable experiments leads to an unsatisfactory theory.* The frequency view of probability has a long history that goes back to **Aristotle**. It was not until 1933 that the great Russian mathematician A. N. **Kolmogorov** (1903-1987) laid a satisfactory mathematical foundation of probability theory. He did this by taking a number of axioms as his starting point, as had been done in other fields of mathematics. [22, p 223]

[11] The class 2^{Ω} of all subsets can be too large for us to define probability measures with consistency, across all member of the class. (There is no problem when Ω is countable.)

Definition 5.1. Kolmogorov's Axioms for Probability [11]: A **probability measure** defined on a σ-algebra $\mathcal{A}$ of Ω is a real-valued (set) function[12] that satisfies[13]:

P1 ***Nonnegativity***:

$$\forall A \in \mathcal{A}, \quad P(A) \geq 0.$$

P2 ***Unit normalization***:

$$P(\Omega) = 1.$$

P3 ***Countable additivity*** or ***σ-additivity***: For every countable sequence $(A_n)_{n=1}^{\infty}$ of disjoint events in $\mathcal{A}$,

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P\left(A_n\right).$$

- The number $P\left(A\right)$ is called the ***probability*** of the event A
- The *triple* $(\Omega, \mathcal{A}, P)$ is called a ***probability measure space***, or simply a ***probability space***
- The entire sample space Ω is called the ***sure event*** or the ***certain event***.
- If an event A satisfies $P(A) = 1$, we say that A is an ***almost-sure event***.
- A ***support*** of P is any set A for which $P\left(A\right) = 1$.

From the three axioms above, we can derive many more properties of probability measure. These basic rules are useful for calculating probabilities.

[12] A real-valued set function is a function the maps sets to real numbers.

[13] The axioms provided here are not exactly the same as those suggested by Kolmogorov. However, it can be shown that they are equivalent to Kolmogorov's axioms.

5.2. $P(\emptyset) = 0$.

5.3. Finite additivity: If $A_1, \ldots, A_n$ are disjoint events, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right).$$

Remark:

(a) It is not possible to go backwards and use finite additivity to derive countable additivity (P3).

(b) Special case when $n = 2$:
Addition rule (Additivity):

$$\text{If } A \cap B = \emptyset, \text{ then } P\left(A \cup B\right) = P\left(A\right) + P\left(B\right). \qquad (5)$$

5.4. If A is countable, then A can be written as a countable disjoint union of singletons:

$$A = \{a_1, a_2, \ldots\} = \bigcup_{n=1}^{\infty} \{a_n\}.$$

By countable additivity (P3), we have

$$P(A) = \sum_{n=1}^{\infty} P(\{a_n\}).$$

Similarly, if A is finite, then A can be written as a disjoint union of singletons:

$$A = \{a_1, a_2, \ldots, a_{|A|}\} = \bigcup_{n=1}^{|A|} \{a_n\}.$$

By countable additivity (P3), we have

$$P(A) = \sum_{n=1}^{|A|} P(\{a_n\}).$$

5.5. ***Monotonicity***: If $A \subset B$, then $P(A) \leq P(B)$

Example 5.6. Let A be the event to roll a 6 and B the event to roll an even number. Whenever A occurs, B must also occur. However, B can occur without A occurring if you roll 2 or 4.

5.7. If $A \subset B$, then $P(B \setminus A) = P(B) - P(A)$

5.8. $P(A) \in [0, 1]$.

5.9. $P(A \cap B)$ can not exceed $P(A)$ and $P(B)$. In other words, "the composition of two events is always less probable than (or at most equally probable to) each individual event."

Example 5.10. Let us consider Mrs. Boudreaux and Mrs. Thibodeaux who are chatting over their fence when the new neighbor walks by. He is a man in his sixties with shabby clothes and a distinct smell of cheap whiskey. Mrs.B, who has seen him before, tells Mrs. T that he is a former Louisiana state senator. Mrs. T finds this very hard to believe. "Yes," says Mrs.B, "he is a former state senator who got into a scandal long ago, had to resign, and started drinking." "Oh," says Mrs. T, "that sounds more likely." "No," says Mrs. B, "I think you mean less likely."

Strictly speaking, Mrs. B is right. Consider the following two statements about the shabby man: "He is a former state senator" and "He is a former state senator who got into a scandal long ago, had to resign, and started drinking." It is tempting to think that the second is more likely because it gives a more exhaustive explanation of the situation at hand. However, this reason is precisely why it is a less likely statement. Note that whenever somebody satisfies the second description, he must also satisfy the first but not vice versa. Thus, the second statement has a lower probability (from Mrs. Ts subjective point of view; Mrs. B of course knows who the man is).

This example is a variant of examples presented in the book *Judgment under Uncertainty* by Economics Nobel laureate Daniel Kahneman and co-authors Paul Slovic and Amos Tversky. They show empirically how people often make similar mistakes when they are asked to choose the most probable among a set of statements. It certainly helps to know the rules of probability. A more discomforting aspect is that the more you explain something in detail, the more likely you are to be wrong. If you want to be credible, be vague. [17, p 11–12]

5.11. Complement Rule:

$$P\left(A^c\right) = 1 - P\left(A\right).$$

- "The probability that something does not occur can be computed as one minus the probability that it does occur."
- Named "probability's Trick Number One" in *Taking Chances: Winning with Probability*, by British probabilist Haigh.

5.12. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- $P(A \cup B) \leq P(A) + P(B)$. See 5.14 for generalization of this subadditivity.
- Approximation: If $P(A) \gg P(B)$ then we may approximate $P(A \cup B)$ by $P(A)$.

Example 5.13. In his bestseller *Innumeracy*, John Allen Paulos tells the story of how he once heard a local weatherman claim that there was a 50% chance of rain on Saturday and a 50% chance of rain on Sunday and thus a 100% chance of rain during the weekend. Clearly absurd, but what is the error?

Answer: Faulty use of the addition rule (5)!

If we let A denote the event that it rains on Saturday and B the event that it rains on Sunday, in order to use $P(A \cup B) = P(A) + P(B)$, we must first confirm that A and B cannot occur at the same time ($P(A \cap B) = 0$). More generally, the formula that is always holds regardless of whether $P(A \cap B) = 0$ is given by 5.12:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

The event "$A \cap B$" describes the case in which it rains both days. To get the probability of rain over the weekend, we now add 50% and 50%, which gives 100%, but we must then subtract the probability that it rains both days. Whatever this is, it is certainly more than 0 so we end up with something less than 100%, just like common sense tells us that we should.

You may wonder what the weatherman would have said if the chances of rain had been 75% each day. [17, p 12]

5.14. Two bounds:

(a) **Subadditivity or Boole's Inequality**: If $A_1, \ldots, A_n$ are events, not necessarily disjoint, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P\left(A_i\right).$$

(b) **σ-subadditivity** or **countable subadditivity**: If $A_1, A_2, \ldots$ is a sequence of measurable sets, not necessarily disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P\left(A_i\right)$$

- This formula is known as the ***union bound*** in engineering.

5.15. If a (finite) collection $\{B_1, B_2, \ldots, B_n\}$ is a partition of Ω, then

$$P(A) = \sum_{i=1}^{n} P\left(A \cap B_i\right)$$

Similarly, if a (countable) collection $\{B_1, B_2, \ldots\}$ is a partition of Ω, then

$$P(A) = \sum_{i=1}^{\infty} P\left(A \cap B_i\right)$$

5.16. If $P(A) = 1$, A is not necessary Ω.

5.17. If $A \subset B$ and $P(B) = 0$, then $P(A) = 0$.

5.18. Connection to classical probability theory: Consider an experiment with **finite** sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ in which each outcome ω_i is **equally likely**. Note that $n = |\Omega|$.

We must have

$$P(\{\omega_i\}) = \frac{1}{n}, \quad \forall i.$$

Now, given any event A, we can write A as a disjoint union of singletons:

$$A = \bigcup_{\omega \in A} \{\omega\}.$$

After applying finite additivity from 5.3, we have

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} \frac{1}{n} = \frac{|A|}{n} = \frac{|A|}{|\Omega|}.$$

We can then say that the probability theory we are working on right now is an extension of the classical probability theory. When the conditons/assumptions of classical probability theory are met, then we get back the defining definition of classical classical probability. The extended part gives us ways to deal with situation where assumptions of classical probability theory are not satisfied.

Definition 5.19. ***Discreet Sample Space***: A sample space Ω is discrete if it is countable[14].

Recall that a probability measure P is a (set) function that assigns number (probability) to all set (event) in $\mathcal{A}$. Recall also that when Ω is countable, we may let $\mathcal{A} = 2^{\Omega}$ = the power set of the sample space. In other words, it is possible to assign[15] probability value to all subsets of Ω.

[14]Recall that Ω is countable if it is either finite or countably infinite. It is countably infinite if its elements can be enumerated as, say, $\omega_1, \omega_2, \ldots$.

[15]Again, this is not true for uncountable Ω.

To define P, it seems that we need to specify a large number of values. Recall that to define a function $g(x)$ you usually specify (in words or as a formula) the value of $g(x)$ at all possible x in the domain of g. The same task must be done here because we have a function that maps sets in $\mathcal{A}$ to real numbers (or, more specifically, the interval $[0, 1]$). It seems that we will need to explicitly specify $P(A)$ for each set A in $\mathcal{A}$. Fortunately, axiom P3 for probability measure implies that we only need to define P for all the singletons when Ω is countable.

To see why this is true, consider any event A. Because Ω is countable, the set $A \subset \Omega$ is also countable. Therefore, we can write A as a disjoint union of singletons:

$$A = \bigcup_{\omega \in A} \{\omega\}.$$

After applying Axiom P3, we have

$$P(A) = P\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} P\left(\{\omega\}\right). \tag{6}$$

In other words, we can find the probability of any event A by adding the probability of the individual outcomes ω in it. Therefore, there is no need to explicitly define $P(A)$ for all A. We only need to define $P\left(\{\omega\}\right)$ for all ω in Ω. In particular, for Ω of size n, formula (6) reduces the task of defining a probability measure from 2^n to n assignments.

Of course, writing $P\left(\{\omega\}\right)$ is tedious. For concisenese, we will introduce the probability mass function (pmf) in 5.21

Example 5.20. For $\Omega = \{1, 2, 3, 4, 5, 6\}$, there are $2^6 = 64$ possible events. Instead of explicitly assigning the probability for all 64 events, by the discussion above, we can simply specify

$$P\left(\{1\}\right), P\left(\{2\}\right), P\left(\{3\}\right), P\left(\{4\}\right), P\left(\{5\}\right), P\left(\{6\}\right)$$

and then use (6) to evaluate $P(A)$ for non-singleton A.For example, when $A = \{1, 2\}$, we can calculate $P(A)$ by

$$P(A) = P\left(\{1, 2\}\right) = P\left(\{1\} \cup \{2\}\right) = P\left(\{1\}\right) + P\left(\{2\}\right).$$

Definition 5.21. When Ω is countable, a ***probability mass function*** (pmf) is any function $p : \Omega \to [0, 1]$ such that

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

When the elements of Ω are enumerated, then it is common to abbreviate $p(\omega_i) = p_i$.

5.22. Every pmf p defines a probability measure P and conversely. Their relationship is given by

$$p(\omega) = P(\{\omega\}), \tag{7}$$

$$P(A) = \sum_{\omega \in A} p(\omega). \tag{8}$$

Again, the convenience of a specification by pmf becomes clear when Ω is a finite set of, say, n elements. Specifying P requires specifying 2^n values, one for each event in $\mathcal{A}$, and doing so in a manner that is consistent with the Kolmogorov axioms. However, specifying p requires only providing n values, one for each element of Ω, satisfying the simple constraints of nonnegativity and addition to 1. The probability measure P satisfying (8) automatically satisfies the Kolmogorov axioms.

Definition 5.23. ***Discrete probability measure*** P is a discrete probability measure if $\exists$ finitely or countably many points ω_k and nonnegative masses p_k such that

$$\forall A \in \mathcal{A}, \quad P(A) = \sum_{k:\omega_k \in A} p_k = \sum_k p_k 1_A(\omega_k).$$

If there is just one of these points, say ω_0, with mass $p_0 = 1$, then P is a ***unit mass*** at ω_0. In this case,

$$\forall A \in \mathcal{A}, \quad P(A) = 1_A(\omega_0).$$

Notation: $P = \delta_{\omega_0}$

- Here, Ω can be ***un***countable.

ECS315 2011/1 Part II.2 Dr.Prapun

6 Event-based Independence and Conditional Probability

Example 6.1. Diagnostic Tests.

6.1 Event-based Conditional Probability

Definition 6.2. ***Conditional Probability***: The conditional probability $P(A|B)$ of event A, given that event $B \neq \emptyset$ occurred, is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{9}$$

- Read "the (conditional) probability of A given B".
- Defined only when $P(B) > 0$.
- If $P(B) = 0$, then it is illogical to speak of $P(A|B)$; that is $P(A|B)$ is not defined.

6.3. ***Interpretation***: Sometimes, we refer to $P(A)$ as

- **a priori probability** , or
- the **prior probability** of A, or
- the **unconditional probability** of A.

It is sometimes useful to interpret $P(A)$ as our knowledge of the occurrence of event A *before* the experiment takes place. Conditional probability $P(A|B)$ is the ***updated probability*** of the event A given that we now know that B occurred (but we still do not know which particular outcome in the set B occurred).

The term **a posteriori** is often used for $P(A|B)$ when we refer to $P(A)$ as a priori.

Example 6.4. In the diagnostic tests example, we learn whether we have the disease from test result. Originally, before taking the test, the probability of having the disease is 0.01%. Being tested positive from the 99%-accurate test ***updates*** the probability of having the disease to about 1%.

More specifically, let D be the event that the testee has the disease and T_P be the event that the test returns positive result.

- Before taking the test, the probability of having the disease is $P(D) = 0.01\%$.
- Using 99%-accurate test means
$$P(T_P|D) = 0.99 \text{ and } P(T_P^c|D^c) = 0.99.$$
- Our calculation shows that $P(D|T_P) \approx 0.01$.

Note also that although the symbol $P(A|B)$ itself is practical, it phrasing in words can be so unwieldy that in practice, less formal descriptions are used. For example, we refer to "the probability that a tested-positive person has the disease" instead of saying "the conditional probability that a randomly chosen person has the disease given that the test for this person returns positive result."

Example 6.5. In communication, as a receiver, you can not "see" what is transmitted to you directly. It is almost always corrupted by noise.

6.6. If the occurrence of B does not give you more information about A, then

$$P(A|B) = P(A) \tag{10}$$

and we say that A and B are ***independent***.

- Meaning: "learning that event B has occurred does not change the probability that event A occurs."
- Interpretation: "the occurrence of event A is not contingent on the occurrence (or nonoccurrence) of event B."

We will soon define "independence". Property (10) can be regarded as a "practical" definition for independence. However, there are some "technical" issues that we need to deal with when we actually define independence.

6.7. Similar properties to the three probability axioms:

(a) Nonnegativity: $P(A|B) \geq 0$

(b) Unit normalization: $P(\Omega|B) = 1$.

In fact, for any event A such that $B \subset A$, we have $P(A|B) = 1$.

This implies

$$P(\Omega|B) = P(B|B) = 1.$$

(c) Countable additivity: For every countable sequence $(A_n)_{n=1}^{\infty}$ of disjoint events,

$$P\left(\bigcup_{n=1}^{\infty} A_n \middle| B\right) = \sum_{n=1}^{\infty} P\left(A_n|B\right).$$

- In particular, if $A_1 \perp A_2$, $P(A_1 \cup A_2 | B) = P(A_1 | B) + P(A_2 | B)$

6.8. Properties:

- $P(A|\Omega) = P(A)$
- If $B \subset A$ and $P(B) \neq 0$, then $P(A|B) = 1$.
- If $A \cap B = \emptyset$ and $P(B) \neq 0$, then $P(A|B) = 0$
- $P(A^c|B) = 1 - P(A|B)$
- $P(A \cap B|B) = P(A|B)$
- $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$.
- $P(A \cap B) \leq P(A|B)$

6.9. When Ω is finite and all outcomes have equal probabilities,

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

This formula can be regarded as the classical version of conditional probability.

Example 6.10. Someone has rolled a fair die twice. You know that one of the rolls turned up a face value of six. The probability that the other roll turned up a six as well is $\frac{1}{11}$ (not $\frac{1}{6}$). [22, Example 8.1, p. 244]

Example 6.11. You know that roughly 5% of all used cars have been flood-damaged and estimate that 80% of such cars will later develop serious engine problems, whereas only 10% of used cars that are not flood-damaged develop the same problems. Of course, no used car dealer worth his salt would let you know whether your car has been flood damaged, so you must resort to probability calculations. What is the probability that your car will later run into trouble?

You might think about this problem in terms of proportions. Out of every 1,000 cars sold, 50 are previously flood-damaged,

and of those, 80%, or 40 cars, develop problems. Among the 950 that are not flood-damaged, we expect 10%, or 95 cars, to develop the same problems. Hence, we get a total of 40+95 = 135 cars out of a thousand, and the probability of future problems is 13.5%.

If you solved the problem in this way, congratulations. You have just used the law of total probability.

6.12. ***Total Probability Theorem***: If a (finite or infinitely) countable collection of events $\{B_1, B_2, \ldots\}$ is a partition of Ω, then

$$P(A) = \sum_i P(A|B_i)P(B_i). \tag{11}$$

To see this, recall, from 5.15, that

$$P(A) = \sum_i P\left(A \cap B_i\right).$$

This is a formula for computing the probability of an event that can occur in different ways.

Example 6.13. The probability that a cell-phone call goes through depends on which tower handles the call.

The probability of internet packets being dropped depends on which route they take through the network.

6.14. Special case:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

This gives exactly the same calculation as what we discussed in Example 6.11.

Example 6.15. Diagnostic Tests:

$$\begin{aligned} P\left(T_P\right) &= P\left(T_P \cap D\right) + P\left(T_P \cap D^c\right) \\ &= P\left(T_P \,|D\right) P\left(D\right) + P\left(T_P \,|D^c\right) P\left(D^c\right). \\ &= (1 - p_{TE})p_D + p_{TE}(1 - p_D). \end{aligned}$$

6.16. Bayes' Theorem:

(a) Form 1:

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

(b) Form 2: If a (finite or infinitely) countable collection of events $\{B_1, B_2, \ldots\}$ is a partition of Ω, then

$$P(B_k|A) = P(A|B_k)\frac{P(B_k)}{P(A)} = \frac{P(A|B_k)P(B_k)}{\sum_i P(A|B_i)P(B_i)}.$$

- Very simple to derive.
- Extremely useful for making inferences about phenomena that cannot be observed directly.
- Sometimes, these inferences are described as "reasoning about causes when we observe effects".

Example 6.17. Decease testing:

$$\begin{aligned} P\left(D\,|T_P\right) &= \frac{P\left(D \cap T_P\right)}{P\left(T_P\right)} = \frac{P\left(T_P\,|D\right)P\left(D\right)}{P\left(T_P\right)} \\ &= \frac{(1-p_{TE})p_D}{(1-p_{TE})p_D + p_{TE}(1-p_D)} \end{aligned}$$

6.18. Probability of compound events

(a) $P(A \cap B) = P(A)P(B|A)$

(b) $P(A \cap B \cap C) = P(A \cap B) \times P(C|A \cap B)$

(c) $P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B)$

When we have many sets intersected in the conditioned part, we often use "," instead of "$\cap$".

Example 6.19. Most people reason as follows to find the probability of getting two aces when two cards are selected at random from an ordinary deck of cards:

(a) The probability of getting an ace on the first card is 4/52.

(b) Given that one ace is gone from the deck, the probability of getting an ace on the second card is 3/51.

(c) The desired probability is therefore

$$\frac{4}{52} \times \frac{3}{51}.$$

[22, p 243]

Example 6.20. In the early 1990s, a leading Swedish tabloid tried to create an uproar with the headline "Your ticket is thrown away!". This was in reference to the popular Swedish TV show "Bingolotto" where people bought lottery tickets and mailed them to the show. The host then, in live broadcast, drew one ticket from a large mailbag and announced a winner. Some observant reporter noticed that the bag contained only a small fraction of the hundreds of thousands tickets that were mailed. Thus the conclusion: Your ticket has most likely been thrown away!

Let us solve this quickly. Just to have some numbers, let us say that there are a total of $N = 100,000$ tickets and that $n = 1,000$ of them are chosen at random to be in the final drawing. If the drawing was from all tickets, your chance to win would be $1/N = 1/100,000$. The way it is actually done, you need to both survive the first drawing to get your ticket into the bag and

then get your ticket drawn from the bag. The probability to get your entry into the bag is $n/N = 1,000/100,000$. The conditional probability to be drawn from the bag, given that your entry is in it, is $1/n = 1/1,000$. Multiply to get $1/N = 1/100,000$ once more. There were no riots in the streets. [17, p 22]

6.21. Chain rule of conditional probability [9, p 58]:

$$P(A \cap B|B) = P(B|C)P(A|B \cap C).$$

Example 6.22. Your teacher tells the class there will be a surprise exam next week. On one day, Monday-Friday, you will be told in the morning that an exam is to be given on that day. You quickly realize that the exam will not be given on Friday; if it was, it would not be a surprise because it is the last possible day to get the exam. Thus, Friday is ruled out, which leaves Monday-Thursday. But then Thursday is impossible also, now having become the last possible day to get the exam. Thursday is ruled out, but then Wednesday becomes impossible, then Tuesday, then Monday, and you conclude: There is no such thing as a surprise exam! But the teacher decides to give the exam on Tuesday, and come Tuesday morning, you are surprised indeed.

This problem, which is often also formulated in terms of surprise fire drills or surprise executions, is known by many names, for example, the "hangman's paradox" or by serious philosophers as the "prediction paradox." To resolve it, let's treat it as a probability problem. Suppose that the day of the exam is chosen randomly among the five days of the week. Now start a new school week. What is the probability that you get the test on Monday? Obviously 1/5 because this is the probability that Monday is chosen. If the test was not given on Monday. what is the probability that it is given on Tuesday? The probability that Tuesday is chosen to start with is 1/5, but we are now asking for the conditional probability that the test is given on Tuesday, given that it was not given on Monday. As there are now four days left, this conditional probability is 1/4. Similarly, the conditional probabilities that the test is given on Wednesday, Thursday, and Friday conditioned on that it has not been given thus far are 1/3, 1/2, and 1, respectively.

We could define the "surprise index" each day as the probability that the test is not given. On Monday, the surprise index is therefore 0.8, on Tuesday it has gone down to 0.75, and it continues to go down as the week proceeds with no test given. On Friday, the surprise index is 0, indicating absolute certainty that the test will be given that day. Thus, it is possible to give a surprise test but not in a way so that you are equally surprised each day, and it is never possible to give it so that you are surprised on Friday. [17, p 23–24]

Example 6.23. Today Bayesian analysis is widely employed throughout science and industry. For instance, models employed to determine car insurance rates include a mathematical function describing, per unit of driving time, your personal probability of having zero, one, or more accidents. Consider, for our purposes, a simplified model that places everyone in one of two categories: high risk, which includes drivers who average at least one accident each year, and low risk, which includes drivers who average less than one.

If, when you apply for insurance, you have a driving record that stretches back twenty years without an accident or one that goes back twenty years with thirty-seven accidents, the insurance company can be pretty sure which category to place you in. But if you are a new driver, should you be classified as low risk (a kid who obeys the speed limit and volunteers to be the designated driver) or high risk (a kid who races down Main Street swigging from a half-empty $2 bottle of Boone's Farm apple wine)?

Since the company has no data on you, it might assign you an equal prior probability of being in either group, or it might use what it knows about the general population of new drivers and start you off by guessing that the chances you are a high risk are, say, 1 in 3. In that case the company would model you as a hybrid–one-third high risk and two-thirds low risk–and charge you one-third the price it charges high-risk drivers plus two-thirds the price it charges low-risk drivers.

Then, after a year of observation, the company can employ the new datum to reevaluate its model, adjust the one-third and two-

third proportions it previously assigned, and recalculate what it ought to charge. If you have had no accidents, the proportion of low risk and low price it assigns you will increase; if you have had two accidents, it will decrease. The precise size of the adjustment is given by Bayes's theory. In the same manner the insurance company can periodically adjust its assessments in later years to reflect the fact that you were accident-free or that you twice had an accident while driving the wrong way down a one-way street, holding a cell phone with your left hand and a doughnut with your right. That is why insurance companies can give out "good driver" discounts: the absence of accidents elevates the posterior probability that a driver belongs in a low-risk group. [13, p 111-112]

6.2 Event-based Independence

Plenty of random things happen in the world all the time, most of which have nothing to do with one another. If you toss a coin and I roll a die, the probability that you get heads is 1/2 regardless of the outcome of my die. Events that in this way are unrelated to each other are called *independent*.

6.24. Sometimes the definition for independence above does not agree with the everyday-language use of the word "independence". Hence, many authors use the term "statistically independence" for the definition above to distinguish it from other definitions.

Definition 6.25. Two events A, B are called (statistically) ***independent*** if

$$P(A \cap B) = P(A)\,P(B) \tag{12}$$

- Notation: $A \perp\!\!\!\perp B$
- Read "A and B are independent" or "A is independent of B"
- We call (12) the ***multiplication rule*** for probabilities.
- If two events are not independent, they are ***dependent***. If two events are dependent, the probability of one changes with the knowledge of whether the other has occurred.
- In classical probability, this is equivalent to

$$|A \cap B||\Omega| = |A||B|.$$

6.26. Intuition: Again, here is how you should think about independent events: "If one event has occurred, the probability of the other does not change."

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B). \tag{13}$$

In other words, "the unconditional and the conditional probabilities are the same". We can almost use (13) as the definitions for independence. However, we use (12) instead because it also works with events whose probabilities are zero. In fact, in 6.30, we show how (13) can be used to define independence with extra condition that deals with the case when zero probability is involved.

Example 6.27. [26, Ex. 5.4] Which of the following pairs of events are independent? (a) The card is a club, and the card is black. (b) The card is a king, and the card is black.

For part (a), we let C be the event that the card is a club and B be the event that it is black. Since there are 26 black cards in an ordinary deck of cards, 13 of which are clubs, the conditional probability $P(C|B)$ is 13/26 (given we are considering only black cards, we have 13 favorable outcomes for the card being a club). The probability of a club (event C), on the other hand, is $P(C) = 13/52$ (13 cards in a 52-card deck are clubs). In this case, $P(C|B) \neq P(C)$ so the events are not independent.

For part (b), we let K be the event that a king is drawn and event B be that it is black. In this case, the probability of a king given that the card is black is $P(K|B) = 2/26$ (two cards of the 26 black cards are kings). The probability of a king is simply $P(K) = 4/52$ (four kings in the 52-card deck). Hence, $P(K|B) = P(K)$. which shows that the events king and black are independent.

6.28. An event with probability 0 or 1 is independent of any event (including itself).

- In particular, $\emptyset$ and Ω are independent of any events.

6.29. An event A is independent of itself if and only if $P(A)$ is 0 or 1.

6.30. Two events A, B *with positive probabilities* are independent if and only if $P(B|A) = P(B)$, which is equivalent to $P(A|B) = P(A)$.

When A and/or B has zero probability, A and B are automatically independent.

6.31. When A and B have nonzero probabilities, the following statements are equivalent:

6.32. If A and B are independent events, then so are A and B^c, A^c and B, and A^c and B^c. By interchanging the roles of A and A^c and/or B and B^c, it follows that if any one of the four pairs is independent, then so are the other three. [9, p.31]

In fact, the following four statements are equivalent:

$$A \perp\!\!\!\perp B, \quad A \perp\!\!\!\perp B^c, \quad A^c \perp\!\!\!\perp B, \quad A^c \perp\!\!\!\perp B^c.$$

6.33. If $A \perp\!\!\!\perp B_i$ for all disjoint events $B_1, B_2, \ldots$, then $A \perp\!\!\!\perp \bigcup_i B_i$.

6.34. Keep in mind that **independent and disjoint** are ***not synonyms***. In some contexts these words can have similar meanings, but this is not the case in probability.

- If two events cannot occur at the same time (they are disjoint), are they independent? At first you might think so. After all, they have nothing to do with each other, right? Wrong! They have a lot to do with each other. If one has occurred, we know for certain that the other cannot occur. [17, p 12]

- The two statements $A \perp B$ and $A \perp\!\!\!\perp B$ can occur simultaneously only when $P(A) = 0$ and/or $P(B) = 0$.

    - Reverse is not true in general.

- Reminder: If events A and B are disjoint, you calculate the probability of the union $A \cup B$ by adding the probabilities of A and B. For independent events A and B you calculate the probability of the intersection $A \cap B$ by multiplying the probabilities of A and B.

Example 6.35. Experiment of flipping a fair coin twice. $\Omega = \{HH, HT, TH, TT\}$. Define event A to be the event that the first flip gives a H; that is $A = \{HH, HT\}$. Event B is the event that the second flip gives a H; that is $B = \{HH, TH\}$. $C = \{HH, TT\}$. Note also that even though the events A and B are not disjoint, they are independent.

Definition 6.36. Three events A_1, A_2, A_3 are independent if and only if

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)\,P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)\,P(A_3) \\ P(A_2 \cap A_3) &= P(A_2)\,P(A_3) \\ P(A_1 \cap A_2 \cap A_3) &= P(A_1)\,P(A_2)\,P(A_3) \end{aligned}$$

Remarks:

(a) When first three equations hold, we say that the three events are ***pairwise independent***.

We may use the term "mutually independence" to further emphasize that we have "independence" instead of "pairwise independence".

(b) It is tempting to think that the last equation is not needed. However, it is possible to construct events such that the first three equations hold (pairwise independence), but the last one does not as demonstrated in Example 6.37.

(c) It is tempting to think that the last equality alone is enough to guarantee the last three. This is not true. The last equality alone is not enough for independence.

- In fact, it is possible for the last equation to hold while the first three fail as shown in Example 6.38.

(d) The last condition can be replaced by $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \cap A_3)$; that is, $A_1 \perp\!\!\!\perp (A_2 \cap A_3)$.

Example 6.37. Having three pairwise independent events does ***not*** imply that the three events are jointly independent. In other

words,

$$A \perp\!\!\!\perp B, B \perp\!\!\!\perp C, A \perp\!\!\!\perp C \nRightarrow A \perp\!\!\!\perp B \perp\!\!\!\perp C.$$

In each of the following examples, each pair of events is independent (pairwise independent) but the three are not. (To show several events are independent, you have to check more than just that each pair is independent.)

(a) Toss three dice. Let N_i be the result of the ith dice.

(b) Let $\Omega = \{1,2,3,4\}$, $\mathcal{A} = 2^{\Omega}$, $p(i) = \frac{1}{4}$, $A_1 = \{1,2\}$, $A_2 = \{1,3\}$, $A_3 = \{2,3\}$. Then $P(A_i \cap A_j) = P(A_i) P(A_j)$ for all $i \neq j$ but$P(A_1 \cap A_2 \cap A_3) \neq P(A_1) P(A_2) P(A_3)$

(c) Let A = [Alice and Betty have the same birthday], B = [Betty and Carol have the same birthday], and C = [Carol and Alice have the same birthday].

Example 6.38. Let $\Omega = \{1,2,3,4,5,6\}$, $\mathcal{A} = 2^{\Omega}$, $p(i) = \frac{1}{6}$, $A_1 = \{1,2,3,4\}$, $A_2 = A_3 = \{4,5,6\}$. Then, $P(A_1 \cap A_2 \cap A_3) = P(A_1) P(A_2) P(A_3)$ but $P(A_i \cap A_j) \neq P(A_i) P(A_j)$ for all $i \neq j$

Example 6.39. Suppose A, B, and C are independent.

(a) Show that $A \perp\!\!\!\perp (B \cap C)$.

(b) Show that $A \perp\!\!\!\perp (B \cup C)$.

(c) Show that $A \perp\!\!\!\perp (B \cap C^c)$.

Definition 6.40. Independence between many events:

(a) Independence for finite collection $\{A_1, \ldots, A_n\}$ of sets:

$$\equiv P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j) \quad \forall J \subset [n] \text{ and } |J| \geq 2$$

- Note that the case when $j = 1$ automatically holds. The case when $j = 0$ can be regard as the $\emptyset$ event case, which is also trivially true.
- There are $\sum_{j=2}^{n} \binom{n}{j} = 2^n - 1 - n$ constraints.

$$\equiv P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1) P(B_2) \cdots P(B_n) \text{ where } B_i = A_i \text{ or } B_i = \Omega$$

(b) Independence for collection $\{A_\alpha : \alpha \in I\}$ of sets:

$$\equiv \forall \text{ finite } J \subset I,\ P\left(\bigcap_{\alpha \in J} A_\alpha\right) = \prod_{\alpha \in J} P(A)$$

$\equiv$ Each of the finite subcollection is independent.

Example 6.41. ***Prosecutor's fallacy***: In 1999, a British jury convicted Sally Clark of murdering two of her children who had died suddenly at the ages of 11 and 8 weeks, respectively. A pediatrician called in as an expert witness claimed that the chance of having two cases of infant sudden death syndrome, or "cot deaths," in the same family was 1 in 73 million. There was no physical or other evidence of murder, nor was there a motive. Most likely, the jury was so impressed with the seemingly astronomical odds against the incidents that they convicted. But where did the number come from? Data suggested that a baby born into a family similar to the Clarks faced a 1 in 8,500 chance of dying a cot death. Two cot deaths in the same family, it was argued, therefore had a probability of $(1/8,500)^2$ which is roughly equal to 1/73,000.000.

Did you spot the error? I hope you did. The computation assumes that successive cot deaths in the same family are *independent* events. This assumption is clearly questionable, and even a person without any medical expertise might suspect that genetic factors play a role. Indeed, it has been estimated that if there is one cot death, the next child faces a much larger risk, perhaps around 1/100. To find the probability of having two cot deaths in the same family, we should thus use conditional probabilities and arrive at the computation $1/8,500 \times 1/100$, which equals l/850,000. Now, this is still a small number and might not have made the jurors judge differently. But what does the probability 1/850,000 have to do with Sallys guilt? Nothing! When her first child died, it was certified to have been from natural causes and there was no suspicion of foul play. The probability that it would happen again without foul play was 1/100, and if that number had been presented to the jury, Sally would not have had to spend three years in jail before the verdict was finally overturned and the expert witness (certainly no expert in probability) found guilty of "serious professional misconduct."

You may still ask the question what the probability 1/100 has to do with Sallys guilt. Is this the probability that she is innocent? Not at all. That would mean that 99% of all mothers who experience two cot deaths are murderers! The number 1/100 is

simply the probability of a second cot death, which only means that among all families who experience one cot death, about 1% will suffer through another. If probability arguments are used in court cases, it is very important that all involved parties understand some basic probability. In Sallys case, nobody did.

References: [13, 118–119] and [17, 22–23].

6.3 Bernoulli Trials

Definition 6.42. A chance experiment is called a ***compound experiment*** [22, p 29 and 231] or ***combined experiment*** if it consists of several **subexperiments** (aka. **elementary experiments**). [19, Ch 3]

- The question arises as to how, in general, we define a probability space for a compound experiment.

Example 6.43. We are given two experiments:

(a) The first experiment is the rolling of a fair dice

$$\Omega_1 = \{1, 2, 3, 4, 5, 6\}, \quad P(\{i\}) = \frac{1}{6}, \forall i \in \Omega_1.$$

(b) The second experiment is the tossing of a fair coin

$$\Omega_2 = \{H, T\}, \quad P(\{H\}) = P(\{T\}) = \frac{1}{2}.$$

We perform both experiments and we want to find the probability that we get "2" on the dice and "H" on the coin.

If we make the reasonable assumption that the outcomes of the first experiment are (physically[16]) independent of the outcomes of the second, we conclude that the unknown probability equals

$$\frac{1}{6} \times \frac{1}{2} = \frac{1}{12}.$$

The above conclusion is reasonable; however, the notion of independence used in its derivation does not agree with the Definition

[16] By physically independent, we mean that the outcomes from any one of the subexperiments have no influence on the functioning or outcomes of any of the other subexperiments.

6.25 given earlier. In that definition, the events A and B were subsets of the same space.

In order to fit the above conclusion into our theory, we must, therefore, construct a new (and larger) sample space Ω having as subsets the events "2" and "H".

This is done as follows:

The two experiments are viewed as (combined into) a single experiment whose outcomes are pairs $\left(\omega^{(1)}, \omega^{(2)}\right)$ where $\omega^{(1)} \in \Omega_1$ and $\omega^{(2)} \in \Omega_2$. The resulting space consists of $6 \times 2 = 12$ elements. The new (and larger) sample space of all the possible pairs $\left(\omega^{(1)}, \omega^{(2)}\right)$ is then given by

$$\Omega = \Omega_1 \times \Omega_2$$

where "$\times$" denotes ***Cartesian product***.

Example 6.44. When $\Omega_1 = \Omega_2 = \{H, T\}$, we have

$$\Omega = \Omega_1 \times \Omega_2 = \{(H, H), (H, T), (T, H), (T, T)\}.$$

6.45. We now generalize Example 6.43 to arbitrary pair of experiments.

The ***Cartesian product of two experiments*** $\mathcal{E}_1$ and $\mathcal{E}_2$ whose sample spaces are Ω_1 and Ω_2 is a new experiment $\mathcal{E}$ with sample space

$$\Omega = \Omega_1 \times \Omega_2$$

whose events are

- all Cartesian products of the form $A_1 \times A_2$ where A_1 is an event of Ω_1 and A_2 is an event of Ω_2,
- and their unions and intersections.

Example 6.46. For countable Ω_1 and Ω_2, it is clear that if we take $A_1 = \left\{\omega^{(1)}\right\}$ and $A_2 = \left\{\omega^{(2)}\right\}$, then $A_1 \times A_2 = \left\{\left(\omega^{(1)}, \omega^{(2)}\right)\right\}$, which is a singleton subset of $\Omega = \Omega_1 \times \Omega_2$. In this way, by changing the values of $\omega^{(1)}$ and $\omega^{(2)}$, we get all the singleton subsets of Ω. Then, by (countable) union of these singleton subsets of Ω, we can get any subsets of Ω.

6.47. Continue from 6.45. There, we have a new sample space for our combined experiment. The next step is to define the probability measure. First, we need to make sure that the event that focuses on only one particular subexperiment still has the same probability as in the original experiment.

For example, in Example 6.43, if we only want to know the probability that we get "2" on the dice, the probability should still be 1/6. Now, to distinguish the probability measures in experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, we will use P_1 and P_2 to denote these probabilities, respectively.

Suppose we want to focus on the event A_1 of experiment $\mathcal{E}_1$ alone. Then, in the new sample space Ω, the equivalent event is $A_1 \times \Omega_2$. Note that the event $A_1 \times \Omega_2$ of the experiment $\mathcal{E}$ occurs if the event A_1 of the experiment $\mathcal{E}_1$ occurs no matter what the outcome of experiment $\mathcal{E}_2$ is. Therefore, the new probability measure should be defined such that

$$P(A_1 \times \Omega_2) = P_1(A_1).$$

Similarly, we must also have

$$P(\Omega_1 \times A_2) = P_2(A_2).$$

Again, $P_i(A_i)$ is the probability of the event A_i in the experiment Ω_i.

Not all events in $\mathcal{E}$ are of the two forms above. If we want to completely specify the probability measure for $\mathcal{E}$, we need more information about the dependency between the two subexperiments. One easy case which we will discuss next is the case when the two subexperiments are independent.

Definition 6.48. ***Independent trials/experiments*** = experiments consisting of ***independent*** repetitions of a ***subexperiment***.

6.49. Continue from refCartesianExp2. Let's now further assume that the two experiment are independent. Note that if we want to consider the combined event that event A_1 occurs in $\mathcal{E}_1$ and event

A_2 occurs in $\mathcal{E}_2$, the corresponding event in $\mathcal{E}$ is $A_1 \times A_2$. Note that

$$A_1 \times A_2 = (A_1 \times \Omega_2) \cap (\Omega_1 \times A_2).$$

We now have a more way to characterize the statement that the two subexperiments are independent. The independence simply requires

$$P(A_1 \times A_2) = P_1(A_1) \times P_2(A_2)$$

for any event A_i of the subexperiment $\mathcal{E}_i$. Equivalently, we may require that the events $A_1 \times \Omega_1$ and $\Omega_1 \times A_2$ of the combined experiment are independent. In which case, we have

$$\begin{aligned} P(A_1 \times A_2) &= P((A_1 \times \Omega_2) \cap (\Omega_1 \times A_2)) \\ &= P(A_1 \times \Omega_2) \times P(\Omega_1 \times A_2) \\ &= P_1(A_1) \times P_2(A_2) \end{aligned}$$

6.50. Given n subexperiments $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ with sample spaces

$$\Omega_1, \Omega_2, \ldots, \Omega_n,$$

their Cartesian product

$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$$

is defined as the sample space of the combined experiment.

Events are then

- all sets of the form $A_1 \times A_2 \times \cdots \times A_n$ where $A_i \subset \Omega_i$, and
- their unions and intersections.

If the subexperiments are independent and $P_i(A_i)$ is the probability of the event A_i in the subexperiment $\mathcal{E}_i$, then

$$P(A_1 \times A_2 \times \cdots \times A_n) = P_1(A_1) \times P_2(A_2) \times \cdots \times P_n(A_n).$$

6.51. In fact, when the Ω_i are all countable, the natural choice for the probability measure P on Ω arises by assigning the probability $p(\omega)$ to each element $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ by using the product rule

$$p(\omega) = p_1(\omega_1) \times p_2(\omega_2) \times \cdots \times p_n(\omega_n).$$

It turns out that this is the only choice that is valid for independent trials.

Definition 6.52. A ***Bernoulli trial*** involves performing an experiment once and noting whether a particular event A occurs.

The outcome of the Bernoulli trial is said to be

(a) a "success" if A occurs and

(b) a "failure" otherwise.

We may view the outcome of a single Bernoulli trial as the outcome of a toss of an unfair coin for which the probability of heads (success) is $p = P(A)$ and the probability of tails (failure) is $1 - p$.

- The labeling ("success" and "failure") is not meant to be literal and sometimes has nothing to do with the everyday meaning of the words.

Example 6.53. Examples of Bernoulli trials: Flipping a coin, deciding to vote for candidate A or candidate B, giving birth to a boy or girl, buying or not buying a product, being cured or not being cured, even dying or living are examples of Bernoulli trials.

- Actions that have multiple outcomes can also be modeled as Bernoulli trials if the question you are asking can be phrased in a way that has a yes or no answer, such as "Did the dice land on the number 4?" or "Is there any ice left on the North Pole?"

Definition 6.54. (Independent) ***Bernoulli Trials*** = independent trials whose subexperiment is a Bernoulli trial.

An outcome of the complete experiment is a sequence of successes and failures which can be denoted by a ***sequence of ones and zeroes***.

Example 6.55. If we toss ***unfair coin*** n times, we obtain the space $\Omega = \{H, T\}^n$ consisting of 2^n elements of the form $(\omega_1, \omega_2, \ldots, \omega_n)$ where $\omega_i =$ H or T.

Example 6.56. What is the probability of two failures and three successes in five Bernoulli trials with success probability p.

We observe that the outcomes with three successes in five trials are 11100, 11010, 11001, 10110, 10101, 10011,01110,01101,01011,

and 00111. We note that the probability of each outcome is a product of five probabilities, each related to one subexperiment. In outcomes with three successes, three of the probabilities are p and the other two are $1-p$. Therefore, each outcome with three successes has probability $(1-p)^2p^3$. There are 10 of them. Hence, the total probability is $10(1-p)^2p^3$

6.57. The probability of exactly n_1 success in $n = n_0+n_1$ bernoulli trials is

$$\binom{n}{n_1}(1-p)^{n-n_1}p^{n_1} = \binom{n}{n_0}(1-p)^{n_0}p^{n-n_0}.$$

Example 6.58. At least one occurrence of a 1-in-n-chance event in n repeated trials:

Example 6.59. ***Digital communication over unreliable channels***: Consider a communication system below

Here, we consider a simple channel called **binary symmetric channel**:

This channel can be described as a channel that introduces random bit errors with probability p.

A crude digital communication system would put binary information into the channel directly; the receiver then takes whatever value that shows up at the channel output as what the sender transmitted. Such communication system would directly suffer bit error probability of p.

In situation where this error rate is not acceptable, error control techniques are introduced to reduce the error rate in the delivered information.

One method of reducing the error rate is to use error-correcting codes:

A simple error-correcting code is the ***repetition code*** described below:

(a) At the transmitter, the "encoder" box performs the following task:

(i) To send a 1, it will send 11111 through the channel.

(ii) To send a 0, it will send 00000 through the channel.

(b) When the five bits pass through the channel, it may be corrupted. Assume that the channel is binary symmetric and that it acts on each of the bit independently.

(c) At the receiver, we (or more specifically, the decoder box) get 5 bits, but some of the bits may be changed by the channel.

To determine what was sent from the transmitter, the receiver apply the ***majority rule***: Among the 5 received bits,

(i) if $\#1 > \#0$, then it claims that "1" we transmitted,

(ii) if $\#0 > \#1$, then it claims that "0" we transmitted.

Example 6.60. The paradox of "almost sure" events: Consider two random events with probabilities of 99% and 99.99%, respectively. One could say that the two probabilities are nearly the same, both events are very likely. Nevertheless the difference may become significant in certain cases. Consider, for instance, independent events which may occur on any day of the year with probability $p = 99\%$; then the probability P that it will occur every day of the year is less than 3%, while if $p = 99.99\%$ then $P = 97\%$. (Check that $P = p^{365}$.)

ECS315 2011/1 Part III.1 Dr.Prapun

7 Random variables

In performing a chance experiment, one is often not interested in the particular outcome that occurs but in a specific numerical value associated with that outcome. In fact, for most applications, measurements and observations are expressed as numerical quantities.

7.1. The advantage of working with numerical quantities is that we can perform mathematical operations on them.

In order to exploit the axioms and properties of probability that we studied earlier, we technically define random variables as functions on an underlying sample space.

Fortunately, once some basic results are derived, we can think of random variables in the traditional manner, and not worry about, or even mention the underlying sample space.

Any function that assigns a real number to each outcome in the sample space of the experiment is called a random variable. Intuitively, a random variable is a variable that takes on its values by chance.

Definition 7.2. A real-valued function $X(\omega)$ defined for points ω in a sample space Ω is called a ***random variable***.

- Random variables are important because they provide a compact way of referring to events via their numerical attributes.
- The abbreviation r.v. will be used for "(real-valued) random variables" [10, p. 1].
- Technically, a random variable must be *measurable*.
- The convention is to use capital letters such as X, Y, Z to denote random variables.

Example 7.3. Take this course and observe your grades.

Remark: Unlike probability models defined on arbitrary sample space, random variables allow us to compute averages.

In the mathematics of probability, averages are called expectations or expected values.

Example 7.4. One Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Probability mass function: $p(\omega_i) = \frac{1}{6}$.
Define function

Example 7.5. If X is the sum of the dots when rolling twice one fair die, the random variable X assigns the numerical value $i + j$ to the outcome (i, j) of the chance experiment.

Example 7.6. Counting the number of heads in a sequence of three coin tosses.

7.7. Technically, if you look at the definition carefully, a random variable is a deterministic function; that is, it is not random and it is not a variable. [Toby Berger]

- As a function, it is simply a rule that maps points/outcomes ω in Ω to real numbers.
- It is also a deterministic function; nothing is random about the mapping/assignment. The randomness in the observed values is due to the underlying randomness of the argument of the function X, namely the experiment outcomes ω.
- In other words, the randomness in the observed value of X is induced by the underlying random experiment, and hence we should be able to compute the probabilities of the observed values in terms of the probabilities of the underlying outcomes.

Example 7.8. Continue from Example 7.4,

(a) What is the probability that $X = 4$?

(b) What is the probability that $Y = 4$?

7.9. At a certain point in most probability courses, the sample space is rarely mentioned anymore and we work directly with random variables. The sample space often "disappears" along with the "(ω)" of $X(\omega)$ but they are really there in the background.

Definition 7.10. Shorthand Notation:

- $[X \in B] = \{\omega \in \Omega : X(\omega) \in B\}$
- $[a \leq X < b] = [X \in [a, b)] = \{\omega \in \Omega : a \leq X(\omega) < b\}$
- $[X > a] = \{\omega \in \Omega : X(\omega) > a\}$
- $[X = x] = \{\omega \in \Omega : X(\omega) = x\}$
    - We usually use the corresponding lowercase letter to denote a possible alue (realization) of the random variable.

All of the above items are sets of outcomes. They are all events!

To avoid double use of brackets (round brackets over square brackets), we write $P\,[X \in B]$ when we means $P\,([X \in B])$. Hence,

$$P\,[X \in B] = P\,([X \in B]) = P\,(\{\omega \in \Omega : X(\omega) \in B\})\,.$$

Similarly,

$$P[X < x] = P\,([X < x]) = P\,(\{\omega \in \Omega : X(\omega) < x\})\,.$$

Example 7.11. Continue from Examples 7.4 and 7.8,

(a) $[X = 4] = \{\omega : X(\omega) = 4\}$

(b) $[Y = 4] = \{\omega : Y(\omega) = 4\} = \left\{\omega : (\omega - 3)^2 = 4\right\}$

Example 7.12. In Example 7.6, if the coins is fair, then
$P[N < 2] =$

Definition 7.13. A set S is called a ***support*** of a random variable X if $P[X \in S] = 1$.

- To emphasize that S is a support of a particular variable X, we denote a support of X by S_X.
- Recall that a support of a probability measure P is any set $A \subset \Omega$ such that $P(A) = 1$.

7.14. There are three types of of random variables. The first type, which will be discussed in Section 8, is called *discrete random variable*.

8 Discrete Random Variables

Definition 8.1. A random variable X is said to be a ***discrete random variable*** if there exists countable distinct real numbers x_k such that

$$\sum_k P[X = x_k] = 1. \tag{14}$$

$\equiv$ X has a countable support $\{x_1, x_2, \ldots\}$

$\equiv$ $\exists$ a countable set $S = \{x_1, x_2, \ldots\}$ such that $P[X \in S] = 1$

Definition 8.2. An ***integer-valued random variable*** is a discrete random variable whose x_k in (14) above are all integers.

Theorem 8.3. A random variable can have at most countably many point x such that$P[X = x] > 0$.

Definition 8.4. When X is a discrete random variable satisfying (14), we define its ***probability mass function*** (pmf) by

$$p_X(x) = P[X = x].$$

- Sometimes, when we only deal with one random variable or when it is clear which random variable the pmf is associated with, we write $p(x)$ or p_x instead of $p_X(x)$.
- The argument of a pmf ranges over all real numbers. Hence, it is defined for x that is not among the x_k in (14). In such case, the pmf is simply 0.
- Many references (including `MATLAB`) use $f_X(x)$ for pmf instead of $p_X(x)$. We will *NOT* use $f_X(x)$ for pmf. Later, we will define $f_X(x)$ as a probability density function which will be used primarily for another type of random variable (continuous r.v.)

Note that for any subset B of $\mathbb{R}$, we can find

$$P[X \in B] = \sum_{x_k \in B} P[X = x_k] = \sum_{x_k \in B} p_X(x_k).$$

In particular, for integer-valued random variables,

$$P[X \in B] = \sum_{k \in B} P[X = k] = \sum_{k \in B} p_X(k).$$

8.5. We can use ***stem plot*** to visualize p_X. To do this, we graph a pmf by marking on the horizontal axis each value with nonzero probability and drawing a vertical bar with length proportional to the probability.

Example 8.6. Let X be the number of heads in a sequence of three coin tosses.

8.7. Any pmf $p(\cdot)$ satisfies two properties:

(a) $p(\cdot) \geq 0$

(b) there exists numbers $x_1, x_2, x_3, \ldots$ such that $\sum_k p(x_k) = 1$ and $p(x) = 0$ for other x.

Example 8.8. Suppose a random variable X has pmf

$$p_X(x) = \begin{cases} c/x, & x = 1, 2, 3, \\ 0, & \text{otherwise.} \end{cases}$$

(a) The value of the constant c is

(b) Sketch of pmf

(c) $P[X = 1]$

(d) $P[X \geq 2]$

(e) $P[X > 3]$

8.9. Any function $p(\cdot)$ on $\mathbb{R}$ which satisfies

(a) $p(\cdot) \geq 0$, and

(b) there exists numbers $x_1, x_2, x_3, \ldots$ such that $\sum_k p(x_k) = 1$ and $p(x) = 0$ for other x

is a pmf of some discrete random variable.

8.1 CDF: Cumulative Distribution Function

Definition 8.10. The (***cumulative***) ***distribution function*** (***cdf***) of a random variable X is the function $F_X(x)$ defined $\forall x \in \mathbb{R}$ by

$$F_X(x) = P[X \leq x].$$

- From its definition, we know that $0 \leq F_X \leq 1$.
- Think of it as a function that collects the "probability mass" from $-\infty$ up to the point x.

Example 8.11. Continue from Example 8.11 where X is defined as the number of heads in a sequence of three coin tosses. We have

$$p_X(0) = p_X(3) = \frac{1}{8} \text{ and } p_X(1) = p_X(2) = \frac{3}{8}.$$

(a) $F_X(0)$

(b) $F_X(1.5)$

(c) Sketch of cdf

8.12. Characterizing properties of cdf:

CDF1 F_X is non-decreasing (monotone increasing)

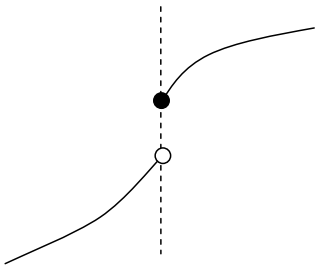

Figure 3: Right-continuous function at jump point

CDF2 F_X is right continuous (continuous from the right):

CDF3 $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.

8.13. If a jump happens at $x = c$, then $P[X = c]$ is the same as the amount of jump at c.

8.14. Properties of the cdf of a discrete random variable: Suppose X is a discrete r.v.

(a) F_X is a right-continuous, staircase function of x with jumps at a countable set of points x_k.

(b) F_X can be written as

$$F_X(x) = \sum_{x_k} p_X(x_k)u(x - x_k),$$

where $u(x) = 1_{[0,\infty)}(x)$ is the unit step function.

ECS315 2011/1 Part III.2 Dr.Prapun

8.2 Families of Discrete Random Variables

Definition 8.15. X is ***uniformly distributed*** on a finite set S if

$$p_X(x) = P[X = x] = \frac{1}{|S|}, \quad \forall\, x \in S.$$

- We write $X \sim \mathcal{U}(S)$ or $X \sim \text{Uniform}(S)$.
- Read "X is uniform on S".
- The pmf is usually referred to as the uniform discrete distribution.
- Simulation: When the support S contains only consecutive integers, it can be generated by the command `randi` in `MATLAB` (R2008b).

Example 8.16. X is uniformly distributed on $1, 2, \ldots, n$ if

Example 8.17. Uniform pmf is used when the random variable can take finite number of "equally likely" or "totally random" values.

- Classical game of chance / classical probability drawing at random
- Fair gaming devices (well-balanced coins and dice, well shuffled decks of cards)

Example 8.18. Toss a fair dice. Let X be the outcome.

Definition 8.19. X is a ***Bernoulli*** random variable if

$$p_X(x) = \begin{cases} 1-p, & x = 0, \\ p, & x = 1, \\ 0, & \text{otherwise}, \end{cases} \qquad p \in (0,1)$$

- Write $X \sim \mathcal{B}(1,p)$ or $X \sim \text{Bernoulli}(p)$
- Some references denote $1-p$ by q for brevity.
- $p_0 = q = 1-p$, $p_1 = p$
- Bernoulli random variable is usually denoted by I. (Think about indicator function; it also has only two possible values, 0 and 1.)

Definition 8.20. X is a ***binary*** random variable if

$$p_X(x) = \begin{cases} 1-p, & x = a, \\ p, & x = b, \\ 0, & \text{otherwise}, \end{cases} \qquad p \in (0,1), \quad b > a.$$

- X takes only two values: a and b

Definition 8.21. X is a ***Binomial*** random variable with size $n \in \mathbb{N}$ and parameter $p \in (0,1)$ if

$$p_X(x) = \begin{cases} \binom{n}{k} p^x (1-p)^{n-x}, & x \in \{0,1,2,\ldots,n\} \\ 0, & \text{otherwise} \end{cases}$$

- Write $X \sim \mathcal{B}(n,p)$.
  - Observe that $\mathcal{B}(n,p)$ is Bernoulli with parameter p.
- Use `binopdf(x,n,p)` in MATLAB.
- Interpretation: X is the number of successes in n independent Bernoulli trials.

Example 8.22. Daily Airlines flies from Amsterdam to London every day. The price of a ticket for this extremely popular flight route is \$75. The aircraft has a passenger capacity of 150. The airline management has made it a policy to sell 160 tickets for this flight in order to protect themselves against no-show passengers. Experience has shown that the probability of a passenger being a no-show is equal to 0.1. The booked passengers act independently of each other. Given this overbooking strategy, what is the probability that some passengers will have to be bumped from the flight?

Solution: This problem can be treated as 160 independent trials of a Bernoulli experiment with a success rate of $p = 9/10$, where a passenger who shows up for the flight is counted as a success. Use the random variable X to denote number of passengers that show up for a given flight. The random variable X is binomial distributed with the parameters $n = 160$ and $p = 9/10$. The probability in question is given by

$$P[X > 150] = 1 - P[X \leq 150] = 1 - F_X(150).$$

In `MATLAB`, we can enter `1-binocdf(150,160,9/10)` to get 0.0359. Thus, the probability that some passengers will be bumped from any given flight is roughly 3.6%. [22, Ex 4.1]

Definition 8.23. A geometric random variable X is defined by the fact that for some $\beta \in (0,1)$, $p_X(k+1) = \beta \times p_X(k)$ for all $k \in S$ where S can be either $\mathbb{N}$ or $\mathbb{N} \cup \{0\}$.

- When its support is $\mathbb{N}$,

$$p_X(x) = \begin{cases} (1-\beta)\,\beta^{x-1}, & x \in \mathbb{N} \\ 0, & \text{otherwise.} \end{cases}$$

  - Write $X \sim \mathcal{G}_1(\beta)$ or geometric$_1(\beta)$.
  - In `MATLAB`, use `geopdf(k-1,1-`β`)`.
  - Interpretation: X is the number of trials required in a Bernoulli trials to achieve the first success.

- When its support is $\mathbb{N} \cup \{0\}$,

$$p_X(k) = (1-\beta)\,\beta^k, \quad \forall\, k \in \mathbb{N} \cup \{0\}$$

  - Write $X \sim \mathcal{G}_0(\beta)$ or geometric$_0(\beta)$.
  - In `MATLAB`, use `geopdf(k,1-`β`)`.
  - Interpretation: X is the number of failures in a Bernoulli trials before the first success occurs.

8.24. In 1837, the famous French mathematician Poisson introduced a probability distribution that would later come to be known

as the Poisson distribution, and this would develop into one of the most important distributions in probability theory. As is often remarked, Poisson did not recognize the huge practical importance of the distribution that would later be named after him. In his book, he dedicates just one page to this distribution. It was Bortkiewicz in 1898, who first discerned and explained the importance of the Poisson distribution in his book Das Gesetz der Kleinen Zahlen (*The Law of Small Numbers*).

Definition 8.25. X is a ***Poisson*** random variable with ***parameter*** $\alpha > 0$ if

$$p_X(k) = \begin{cases} e^{-\alpha}\frac{\alpha^k}{k!}, & k \in \{0, 1, 2, \ldots\} \\ 0, & \text{otherwise} \end{cases}$$

- In `MATLAB`, use `poisspdf(k,alpha)`.
- Write $X \sim \mathcal{P}(\alpha)$ or Poisson(α).
- We will see later in Example 8.54 that α is the "average" or expected value of X.
- Instead of X, Poisson random variable is usually denoted by Λ. The parameter α is often replaced by $\lambda\tau$ where λ is referred to as the ***intensity/rate parameter*** of the distribution

8.26. Summary:

$X \sim$	Support set $\mathcal{X}$	$p_X(k)$	$\varphi_X(u)$
Uniform $\mathcal{U}_n$	$\{1, 2, \ldots, n\}$	$\frac{1}{n}$	
$\mathcal{U}_{\{0,1,\ldots,n-1\}}$	$\{0, 1, \ldots, n-1\}$	$\frac{1}{n}$	$\frac{1-e^{iun}}{n(1-e^{iu})}$
Bernoulli $\mathcal{B}(1,p)$	$\{0, 1\}$	$\begin{cases} 1-p, & k=0 \\ p, & k=1 \end{cases}$	
Binomial $\mathcal{B}(n,p)$	$\{0, 1, \ldots, n\}$	$\binom{n}{k}p^k(1-p)^{n-k}$	$(1-p+pe^{ju})^n$
Geometric $\mathcal{G}_0(\beta)$	$\mathbb{N} \cup \{0\}$	$(1-\beta)\beta^k$	$\frac{1-\beta}{1-\beta e^{iu}}$
Geometric $\mathcal{G}_1(\beta)$	$\mathbb{N}$	$(1-\beta)\beta^{k-1}$	
Poisson $\mathcal{P}(\alpha)$	$\mathbb{N} \cup \{0\}$	$e^{-\alpha}\frac{\alpha^k}{k!}$	$e^{\alpha\left(e^{iu}-1\right)}$

Table 3: Examples of probability mass functions. Here, $p, \beta \in (0,1)$. $\alpha > 0$. $n \in \mathbb{N}$

Example 8.27. The first use of the Poisson model is said to have been by a Prussian (German) physician, Bortkiewicz, who found that the annual number of late-19th-century Prussian (German) soldiers kicked to death by horses fitted a Poisson distribution [6, p 150],[3, Ex 2.23][17].

Example 8.28. The number of hits to a popular website during a 1-minute interval is given by $N \sim \mathcal{P}(\alpha)$ where $\alpha = 2$.

(a) Find the probability that there is at least one hit between 3:00AM and 3:01AM.

(b) Find the probability that there are at least 2 hits during the time interval above.

8.2.1 A Preview of Poisson Process

8.29. One of the reasons why Poisson distribution is important is because many natural phenomenons can be modeled by ***Poisson processes***.

Definition 8.30. A ***Poisson process*** (PP) is a random arrangement of "marks" (denoted by "$\times$" below) on the time line.

The "marks" may indicate the arrival times or occurrences of event/phenomenon of interest.

[17] I. J. Good and others have argued that the Poisson distribution should be called the Bortkiewicz distribution, but then it would be very difficult to say or write.

Example 8.31. Examples of processes that can be modeled by ***Poisson process*** include

(a) the sequence of times at which lightning strikes occur or mail carriers get bitten within some region

(b) the emission of particles from a radioactive source

(c) the arrival of

- telephone calls at a switchboard or at an automatic phone-switching system
- urgent calls to an emergency center
- (filed) claims at an insurance company
- incoming spikes (action potential) to a neuron in human brain

(d) the occurrence of

- serious earthquakes
- traffic accidents
- power outages

in a certain area.

(e) page view requests to a website

8.32. It is convenient to consider the Poisson process in terms of customers arriving at a facility.

We shall focus on a type of Poisson process that is called *homogeneous Poisson process*.

Definition 8.33. For ***homogeneous Poisson process***, there is only one parameter that describes the whole process. This number is call the ***rate*** and usually denoted by λ.

Example 8.34. If you think about modeling customer arrival as a Poisson process with rate $\lambda = 5$ customers/hour, then it means that during any fixed time interval of duration 1 hour (say, from

noon to 1PM), you expect to have about 5 customers arriving in that interval. If you consider a time interval of duration two hours (say, from 1PM to 3PM), you expect to have about $2 \times 5 = 10$ customers arriving in that time interval.

8.35. More generally, For a homogeneous Poisson process of rate λ, during a time interval of length τ, the average number of arrivals will be $\lambda \times \tau$.

One important fact which we will show later is that, when we consider a fixed time interval, the ***number of arrivals*** for a Poisson process is a Poisson random variable. So, now we know that the "average" or expected value of this random variable must be λT.

Summary: **For a homogeneous Poisson process, the number of arrivals during a time interval of duration T is a Poisson random variable with parameter $\alpha = \lambda T$.**

Example 8.36. Examples of Poisson *random variables*:

- #photons emitted by a light source of intensity λ [photons/second] in time τ
- #atoms of radioactive material undergoing decay in time τ
- #clicks in a Geiger counter in τ seconds when the average number of click in 1 second is λ.
- #dopant atoms deposited to make a small device such as an FET
- #customers arriving in a queue or workstations requesting service from a file server in time τ
- Counts of demands for telephone connections in time τ
- Counts of defects in a semiconductor chip.

Example 8.37. Thongchai produces a new hit song every 7 months on average. Assume that songs are produced according to a Poisson process. Find the probability that Thongchai produces more than two hit songs in 1 year.

8.2.2 Poisson Approximation

8.38. ***Poisson approximation*** of Binomial distribution: When p is small and n is large, $\mathcal{B}(n,p)$ can be approximated by $\mathcal{P}(np)$

(a) In a large number of independent repetitions of a Bernoulli trial having a small probability of success, the total number of successes is approximately Poisson distributed with parameter $\alpha = np$, where n = the number of trials and p = the probability of success. [22, p 109]

(b) More specifically, suppose $X_n \sim \mathcal{B}(n,p_n)$. If $p_n \to 0$ and $np_n \to \alpha$ as $n \to \infty$, then

$$P[X_n = k] \to e^{-\alpha}\frac{\alpha^k}{k!}.$$

To see this, note that

$$\binom{n}{k}p_n^k(1-p_n)^{n-k} = \underbrace{\binom{n}{k}\frac{1}{n^k}}_{\to\frac{1}{k!}}\underbrace{(np_n)^k}_{\to\alpha^i}\underbrace{(1-p_n)^n}_{=(1-\frac{np_n}{n})^n\to e^{-\alpha}}\underbrace{(1-p_n)^{-k}}_{\to 1}.$$

Example 8.39. Recall that Bortkiewicz applied the Poisson model to the number of Prussian cavalry deaths attributed to fatal horse kicks. Here, indeed, one encounters a very large number of trials (the Prussian cavalrymen), each with a very small probability of "success" (fatal horse kick).

8.40. Poisson approximation for weakly dependent trials [22, Section 4.2.3]:

The Poisson distribution is derived for the situation of many independent trials each having a small probability of success. In case the independence assumption is not satisfied, but there is a "weak" dependence between the trial outcomes, the Poisson model may still be useful as an approximation. In surprisingly many probability problems, the Poisson approximation method enables us to obtain quick estimates for probabilities that are otherwise difficult to calculate.

This approach requires that the problem is reformulated in the framework of a series of (weakly dependent) trials. The idea of the method is first illustrated by the birthday problem.

Example 8.41. [22, p 114] The birthday problem revisited: Recall that, in a (randomly formed) group of r people, the probability that at least two of them will have birthdays on the same day of the year is given by

$$1 - \frac{(n)_r}{n^r} \approx 1 - e^{-\frac{r(r-1)}{2n}},$$

where $n = 365$. It is interesting to write the approximated quantity above as

$$1 - e^{-\binom{r}{2} \times \frac{1}{n}}.$$

To place the birthday problem in the context of a series of trials, some creativity is called for. The idea is to consider all of the possible combinations of two people and to trace whether, in any of those combinations, the two persons have birthdays on the same day. Only when such a combination exists can it be said that two or more people out of the whole group have birthdays on the same day. What you are doing, in fact, is conducting $m = \binom{r}{2}$ trials. Every trial has the same probability of success $p = 1/n = 1/365$ in showing the probability that two given people will have birthdays on the same day (this probability is the same as the probability that a person chosen at random matches your birthday).

Assume that the random variable X indicates the number of trials where both people have birthdays on the same day. The probability that, in a group of r people, two or more people will

have birthdays on the same day is then equal to $P[X \geq 1]$. Although the outcomes of the trials are dependent on one another, this dependence is considered to be weak because of the vast number (n = 365) of possible birth dates. It is therefore reasonable to approximate X as a Poisson random variable with parameter

$$\alpha = mp = \binom{r}{2} \times \frac{1}{n}.$$

In particular,

$$P[X \geq 1] = 1 - P[X = 0] \approx 1 - e^{-\alpha} = 1 - e^{-\binom{r}{2} \times \frac{1}{n}}.$$

This results in an approximate value of $1 - e^{-0.69315} = 0.5000$ for the probability that, in a group of 23 people, two or more people will have their birthdays on the same day. This is an excellent approximation for the exact value 0.5073 of this probability. The approximation approach with $\binom{23}{2} = 253$ trials and a success probability of 1/365 on each trial explains why a relatively small group of 23 people is sufficient to give approximately a 50% probability of encountering two people with birthdays on the same day. The exact solution for the birthday problem does not provide this insight.

Example 8.42. [22, Section 3.1.4 and p 115] The "almost" birthday problem:

In the almost-birthday problem, we undertake the task of determining the probability of two or more people in a (randomly assembled) group of r people having their birthdays within d days of each other. This probability is given by

$$1 - \frac{(n - 1 - rd)!}{n^{r-1}\,(n - (d+1)\,r)!},$$

where, again, $n = 365$. The proof of this formula is rather tricky and can be found in J.I. Nauss, "An Extension of the Birthday Problem," The American Statistician 22 (1968): 2729.

Although the derivation of an exact formula above is far from simple, a Poisson approximation is particularly simple to give.

Let's try the case where $d = 1$: what is the probability that, within a randomly formed group of r people, two or more people will have birthdays within one day of each other?

You must reconsider all the possible combinations of two people, that is, you must run $m = \binom{r}{2}$ trials. The probability of success in a given trial is now equal to $p = 3/365$ (the probability that two given people will have birthdays within one day of each other). The number of successful trials is approximately Poisson distributed with parameter $\alpha = mp$. In particular, the probability that two or more people will have birthdays within one day of each other is approximately equal to

$$P[X \geq 1] = 1 - P[X = 0] \approx 1 - e^{-\alpha} = 1 - e^{-\binom{r}{2} \times \frac{3}{365}}.$$

For $r = 14$, the approximate value is $1 - e^{-0.74795} = 0.5267$ (the exact value of the probability is 0.5375).

For $d \neq 1$, we simply replace the number 3 in the formula above by $2d + 1$.

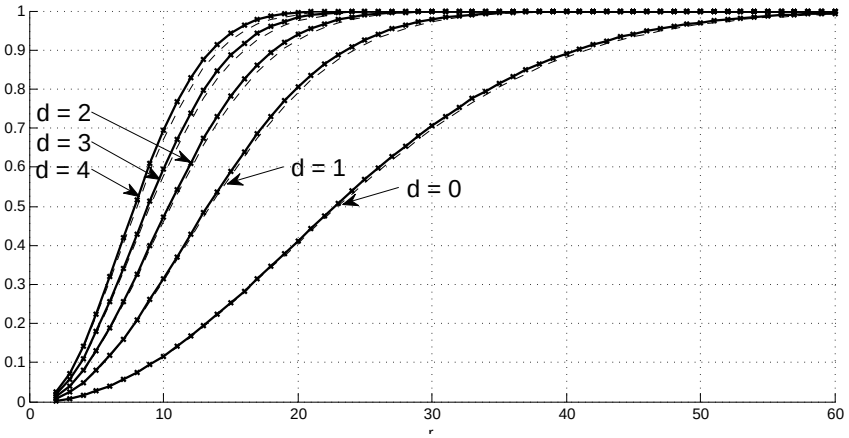

Figure 4: The probability that, within a randomly formed group of r people, two or more people will have birthdays within d day of each other.

8.3 Some Remarks

8.43. To emphasize the the importance of the parameters n, p, α, β, some references include them in the pmf notation.

For example, the pmf of $\mathcal{B}(n,p)$ is expressed as $p_X(x;n,p)$ to emphasize that fact that it depends on the parameters n and p.

8.44. Sometimes, it is useful to define and think of pmf as a vector $\underline{p}$ of probabilities.

When you use `MATLAB`, it is also useful to keep track of the values of x corresponding to the probabilities in p. This can be done via defining a vector $\underline{x}$.

Example 8.45. For $\mathcal{B}\left(3, \frac{1}{3}\right)$, we may define

$$\underline{x} = [0, 1, 2, 3]$$

and

$$\begin{aligned}\underline{p} &= \left[\binom{3}{0}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^3, \binom{3}{1}\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^2, \binom{3}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^1, \binom{3}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^0\right] \\ &= \left[\frac{8}{27}, \frac{4}{9}, \frac{2}{9}, \frac{1}{27}\right]\end{aligned}$$

8.46. At this point, we have a couple of ways to define probabilities that are associated with a random variable X

(a) We can define $P[X \in B]$ for all possible set B.

(b) For discrete random variable, we only need to define its pmf $p_X(x)$ which is defined as $P[X = x] = P[X \in \{x\}]$.

(c) We can also define the cdf $F_X(x)$.

Definition 8.47. If $p_X(c) = 1$, that is $P[X = c] = 1$, for some constant c, then X is called a ***degenerated*** random variable.

8.4 Expectation of Discrete Random Variable

The most important characteristic of a random variable is its expectation. Synonyms for expectation are expected value, mean, and first moment.

The definition of expectation is motivated by the conventional idea of numerical average. Recall that the numerical average of n numbers, say $a_1, a_2, \ldots, a_n$ is

$$\frac{1}{n}\sum_{k=1}^{n} a_k.$$

We use the average to summarize or characterize the entire collection of numbers $a_1, \ldots, a_n$ with a single value.

Example 8.48. Consider 10 numbers: 5, 2, 3, 2, 5, -2, 3, 2, 5, 2.

The average is

$$\frac{5+2+3+2+5+(-2)+3+2+5+2}{10} = \frac{27}{10} = 2.7.$$

We can rewrite the above calculation as

$$-2 \times \frac{1}{10} + 2 \times \frac{4}{10} + 3 \times \frac{2}{10} + 5 \times \frac{3}{10}$$

Definition 8.49. Suppose X is a discrete random variable, we define the **expectation** (or ***mean*** or ***expected value***) of X by

$$\mathbb{E}X = \sum_x x \times P[X = x] = \sum_x x \times p_X(x). \tag{15}$$

In other words, The expected value of a discrete random variable is a weighted mean of the values the random variable can take on where the weights come from the pmf of the random variable.

- Some references use m_X or μ_X to represent $\mathbb{E}X$.

- For conciseness, we simply write x under the summation symbol in (15); this means that the sum runs over all x values in the support of X. (Of course, for x outside of the support, $p_X(x)$ is 0 anyway.)

8.50. In mechanics, think of point masses on a line with a mass of $p_X(x)$ kg. at a distance x meters from the origin.

In this model, $\mathbb{E}X$ is the center of mass.

This is why $p_X(x)$ is called probability mass function.

Example 8.51. When $X \sim \text{Bernoulli}(p)$ with $p \in (0,1)$,

Note that, since X takes only the values 0 and 1, its expected value p is "never seen".

8.52. Interpretation: The expected value is in general not a typical value that the random variable can take on. It is often helpful to interpret the expected value of a random variable as the ***long-run average value*** of the variable over many independent repetitions of an experiment

Example 8.53. $p_X(x) = \begin{cases} 1/4, & x = 0 \\ 3/4, & x = 2 \\ 0, & \text{otherwise} \end{cases}$

Example 8.54. For $X \sim \mathcal{P}(\alpha)$,

$$\mathbb{E}X = \sum_{i=0}^{\infty} i e^{-\alpha} \frac{(\alpha)^i}{i!} = \sum_{i=1}^{\infty} e^{-\alpha} \frac{(\alpha)^i}{i!} i + 0 = e^{-\alpha}(\alpha) \sum_{i=1}^{\infty} \frac{(\alpha)^{i-1}}{(i-1)!}$$
$$= e^{-\alpha} \alpha \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} = e^{-\alpha} \alpha e^{\alpha} = \alpha.$$

Example 8.55. For $X \sim \mathcal{B}(n,p)$,

$$\begin{aligned}\mathbb{E}X &= \sum_{i=0}^{n} i\binom{n}{i}p^i(1-p)^{n-i} = \sum_{i=1}^{n} i\frac{n!}{i!\,(n-i)!}p^i(1-p)^{n-i} \\ &= n\sum_{i=1}^{n}\frac{(n-1)!}{(i-1)!\,(n-i)!}p^i(1-p)^{n-i} = n\sum_{i=1}^{n}\binom{n-1}{i-1}p^i(1-p)^{n-i}\end{aligned}$$

Let $k = i-1$. Then,

$$\mathbb{E}X = n\sum_{k=0}^{n-1}\binom{n-1}{k}p^{k+1}(1-p)^{n-(k+1)} = np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{n-1-k}$$

We now have the expression in the form that we can apply the binomial theorem which finally gives

$$\mathbb{E}X = np(p+(1-p))^{n-1} = np.$$

We shall revisit this example again using another approach in Example 9.35.

Example 8.56. ***Pascal's wager***: Suppose you concede that you don't know whether or not God exists and therefore assign a 50 percent chance to either proposition. How should you weigh these odds when deciding whether to lead a pious life? If you act piously and God exists, Pascal argued, your gain–eternal happiness–is infinite. If, on the other hand, God does not exist, your loss, or negative return, is small–the sacrifices of piety. To weigh these possible gains and losses, Pascal proposed, you multiply the probability of each possible outcome by its payoff and add them all up, forming a kind of average or expected payoff. In other words, the mathematical expectation of your return on piety is one-half infinity (your gain if God exists) minus one-half a small number (your loss if he does not exist). Pascal knew enough about infinity to know that the answer to this calculation is infinite, and thus the expected return on piety is infinitely positive. Every reasonable person, Pascal concluded, should therefore follow the laws of God. [13, p 76]

- Pascals wager is often considered the founding of the mathematical discipline of game theory, the quantitative study of optimal decision strategies in games.

Example 8.57. A sweepstakes sent through the mail offered a grand prize of $5 million. All you had to do to win was mail in your entry. There was no limit on how many times you could enter, but each entry had to be mailed in separately. The sponsors were apparently expecting about 200 million entries, because the fine print said that the chances of winning were 1 in 200 million. Does it pay to enter this kind of "free sweepstakes offer"?

Multiplying the probability of winning times the payoff, we find that each entry was worth 1/40 of $1, or $0.025 far less than the cost of mailing it in. In fact, the big winner in this contest was the post office, which, if the projections were correct, made nearly $80 million in postage revenue on all the submissions. [13, p 77]

8.58. Technical issue: Definition (15) is only meaningful if the sum is well defined.

The sum of infinitely many nonnegative terms is always well-defined, with $+\infty$ as a possible value for the sum.

- ***Infinite Expectation***: Consider a random variable X whose pmf is defined by

$$p_X(x) = \begin{cases} \frac{1}{cx^2}, & x = 1, 2, 3, \ldots \\ 0, & \text{otherwise} \end{cases}$$

  Then, $c = \sum_{n=1}^{\infty} \frac{1}{n^2}$ which is a finite positive number $(\pi^2/6)$. However,

$$\mathbb{E}X = \sum_{k=1}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k \frac{1}{c} \frac{1}{k^2} = \frac{1}{c} \sum_{k=1}^{\infty} \frac{1}{k} = +\infty.$$

Some care is necessary when computing expectations of signed random variables that take infinitely many values.

- The sum over countably infinite many terms is not always well defined when both positive and negative terms are involved.
- For example, the infinite series $1-1+1-1+\ldots$ has the sum 0 when you sum the terms according to $(1-1)+(1-1)+\cdots$, whereas you get the sum 1 when you sum the terms according to $1+(-1+1)+(-1+1)+(-1+1)+\cdots$.

- Such abnormalities cannot happen when all terms in the infinite summation are nonnegative.

It is the convention in probability theory that $\mathbb{E}X$ should be evaluated as

$$\mathbb{E}X = \sum_{x \geq 0} x p_X(x) - \sum_{x<0} (-x) p_X(x),$$

- If at least one of these sums is finite, then it is clear what value should be assigned as $\mathbb{E}X$.
- If both sums are $+\infty$, then no value is assigned to $\mathbb{E}X$, and we say that $\mathbb{E}X$ is **undefined**.

Example 8.59. Undefined Expectation: Let

$$p_X(x) = \begin{cases} \frac{1}{2cx^2}, & x = \pm 1, \pm 2, \pm 3, \ldots \\ 0, & \text{otherwise} \end{cases}$$

Then,

$$\mathbb{E}X = \sum_{k=1}^{\infty} k p_X(k) - \sum_{k=-\infty}^{-1} (-k) p_X(k).$$

The first sum gives

$$\sum_{k=1}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k \frac{1}{2ck^2} = \frac{1}{2c} \sum_{k=1}^{\infty} \frac{1}{k} = \frac{\infty}{2c}.$$

The second sum gives

$$\sum_{k=-\infty}^{-1} (-k) p_X(k) = \sum_{k=1}^{\infty} k p_X(-k) = \sum_{k=1}^{\infty} k \frac{1}{2ck^2} = \frac{1}{2c} \sum_{k=1}^{\infty} \frac{1}{k} = \frac{\infty}{2c}.$$

Because both sums are infinite, we conclude that $\mathbb{E}X$ is undefined.

8.60. More rigorously, to define $\mathbb{E}X$, we let $X^+ = \max\{X, 0\}$ and $X^- = -\min\{X, 0\}$. Then observe that $X = X^+ - X^-$ and that both X^+ and X^- are nonnegative r.v.'s. We say that a random variable X ***admits an expectation*** if $\mathbb{E}X^+$ and $\mathbb{E}X^-$ are not both equal to $+\infty$. In which case, $\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$.

ECS315 2011/1 Part III.3 Dr.Prapun

8.5 Function of a Discrete Random Variable

Given a random variable X, we will often have occasion to define a new random variable by $Z \equiv g(X)$, where $g(x)$ is a real-valued function of the real-valued variable x. More precisely, recall that a random variable X is actually a function taking points of the sample space, $\omega \in \Omega$, into real numbers $X(\omega)$. Hence, we have the following definition

Definition 8.61. The notation $Y = g(X)$ is actually shorthand for $Y(\omega) := g(X(\omega))$.

- The random variable $Y = g(X)$ is sometimes called **derived** random variable.

Example 8.62. Let

$$p_X(x) = \begin{cases} \frac{1}{c}x^2, & x = \pm 1, \pm 2 \\ 0, & \text{otherwise} \end{cases}$$

and

$$Y = X^4.$$

Find $p_Y(y)$ and then calculate $\mathbb{E}Y$.

8.63. For discrete random variable X, the pmf of a derived random variable $Y = g(X)$ is given by

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

Example 8.64. A "binary" random variable X takes only two values a and b with

$$P[X = b] = 1 - P[X = a] = p.$$

X can be expressed as $X = (b - a)I + a$, where I is a Bernoulli random variable with parameter p.

Example 8.65. Suppose X is $\mathcal{G}_0(\beta)$. Then, $Y = X + 1$ is $\mathcal{G}_1(\beta)$.

8.6 Expectation of a Function of a Discrete Random Variable

Recall that for discrete random variable X, the pmf of a derived random variable $Y = g(X)$ is given by

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

If we want to compute $\mathbb{E}Y$, it might seem that we first have to find the pmf of Y. Typically, this requires a detailed analysis of g which can be complicated, and it is avoided by the following result.

8.66. Suppose X is a discrete random variable.

$$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x).$$

This is referred to as the **law/rule of the lazy/unconcious statistician** (LOTUS) [23, Thm 3.6 p 48],[9, p. 149],[8, p. 50] because it is so much easier to use the above formula than to first find the pmf of Y. It is also called **substitution rule** [22, p 271].

Example 8.67. Back to Example 8.62. Recall that

$$p_X(x) = \begin{cases} \frac{1}{c}x^2, & x = \pm 1, \pm 2 \\ 0, & \text{otherwise} \end{cases}$$

(a) When $Y = X^4$, $\mathbb{E}Y =$

(b) $\mathbb{E}[2X - 1]$

8.68. Caution: A frequently made *mistake* of beginning students is to set $\mathbb{E}[g(X)]$ equal to $g(\mathbb{E}X)$. In general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}X)$.

(a) In particular, $\mathbb{E}\left[\frac{1}{X}\right]$ is not the same as $\frac{1}{\mathbb{E}X}$.

(b) An exception is the case of a linear function $g(x) = ax + b$. See also (8.71) and (**??**).

Example 8.69. For $X \sim \text{Bernoulli}(p)$,

(a) $\mathbb{E}X = p$

(b) $\mathbb{E}\left[X^2\right] = 0^2 \times (1-p) + 1^2 \times p = p \neq (\mathbb{E}X)^2$.

Example 8.70. Continue from Example 8.54. Suppose $X \sim \mathcal{P}(\alpha)$.

$$\mathbb{E}\left[X^2\right] = \sum_{i=0}^{\infty} i^2 e^{-\alpha} \frac{\alpha^i}{i!} = e^{-\alpha}\alpha \sum_{i=0}^{\infty} i \frac{\alpha^{i-1}}{(i-1)!} \tag{16}$$

We can evaluate the infinite sum in (16) by rewriting i as $i-1+1$:

$$\begin{aligned}\sum_{i=1}^{\infty} i\frac{\alpha^{i-1}}{(i-1)!} &= \sum_{i=1}^{\infty} (i-1+1)\frac{\alpha^{i-1}}{(i-1)!} = \sum_{i=1}^{\infty} (i-1)\frac{\alpha^{i-1}}{(i-1)!} + \sum_{i=1}^{\infty} \frac{\alpha^{i-1}}{(i-1)!} \\ &= \alpha \sum_{i=2}^{\infty} \frac{\alpha^{i-2}}{(i-2)!} + \sum_{i=1}^{\infty} \frac{\alpha^{i-1}}{(i-1)!} = \alpha e^{\alpha} + e^{\alpha} = e^{\alpha}(\alpha+1).\end{aligned}$$

Plugging this back into (16), we get

$$\mathbb{E}\left[X^2\right] = \alpha\left(\alpha + 1\right) = \alpha^2 + \alpha.$$

8.71. Some Basic Properties of Expectations

(a) For $c \in \mathbb{R}$, $\mathbb{E}\left[c\right] = c$

(b) For $c \in \mathbb{R}$, $\mathbb{E}\left[X + c\right] = \mathbb{E}X + c$ and $\mathbb{E}\left[cX\right] = c\mathbb{E}X$

(c) For constants a, b, we have $\mathbb{E}\left[aX + b\right] = a\mathbb{E}X + b$.

(d) $\mathbb{E}\left[X - \mathbb{E}X\right] = 0$.

Definition 8.72. Some definitions involving expectation of a function of a random variable:

(a) ***Absolute moment***: $\mathbb{E}\left[\left|X\right|^k\right]$, where we define $\mathbb{E}\left[\left|X\right|^0\right] = 1$

(b) ***Moment***: $m_k = \mathbb{E}\left[X^k\right]$ = the k^{th} moment of X, $\quad k \in \mathbb{N}$.

- The first moment of X is its expectation $\mathbb{E}X$.
- The second moment of X is $\mathbb{E}\left[X^2\right]$.

8.7 Variance and Standard Deviation

An average (expectation) can be regarded as one number that summarizes an entire probability model. After finding an average, someone who wants to look further into the probability model might ask, "How typical is the average?" or, "What are the chances of observing an event far from the average?" A measure of **dispersion/deviation/spread** is an answer to these questions wrapped up in a single number. (The opposite of this measure is the **peakedness**.) If this measure is small, observations are likely to be near the average. A high measure of dispersion suggests that it is not unusual to observe events that are far from the average.

Example 8.73. Consider your score on the midterm exam. After you find out your score is 7 points above average, you are likely to ask, "How good is that? Is it near the top of the class or somewhere near the middle?".

Example 8.74. In the case that the random variable X is the random payoff in a game that can be repeated many times under identical conditions, the expected value of X is an informative measure on the grounds of the law of large numbers. However, the information provided by $\mathbb{E}X$ is usually not sufficient when X is the random payoff in a nonrepeatable game.

Suppose your investment has yielded a profit of \$3,000 and you must choose between the following two options:

- the first option is to take the sure profit of \$3,000 and
- the second option is to reinvest the profit of \$3,000 under the scenario that this profit increases to \$4,000 with probability 0.8 and is lost with probability 0.2.

The expected profit of the second option is

$$0.8 \times \$4,000 + 0.2 \times \$0 = \$3,200$$

and is larger than the \$3,000 from the first option. Nevertheless, most people would prefer the first option. The downside ***risk*** is too big for them. A measure that takes into account the aspect of risk is the *variance* of a random variable. [22, p 35]

Definition 8.75. The most important ***measures of dispersion*** are the standard deviation and its close relative, the variance.

(a) ***Variance***:

$$\operatorname{Var} X = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]. \tag{17}$$

- Read "the variance of X"
- *Notation*: D_X, or $\sigma^2(X)$, or σ_X^2, or $\mathbb{V}X$ [23, p. 51]
- In some references, to avoid confusion from the two expectation symbols, they first define $m = \mathbb{E}X$ and then define the variance of X by

$$\operatorname{Var} X = \mathbb{E}\left[(X - m)^2\right].$$

- We can also calculate the variance via another identity:

$$\operatorname{Var} X = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2$$

- The units of the variance are squares of the units of the random variable.
- $\operatorname{Var} X \geq 0$.
- $\operatorname{Var} X \leq \mathbb{E}\left[X^2\right]$.
- $\operatorname{Var}[cX] = c^2 \operatorname{Var} X$.
- $\operatorname{Var}[X + c] = \operatorname{Var} X$.
- $\operatorname{Var}[aX + b] = a^2 \operatorname{Var} X$.

(b) ***Standard Deviation***: $\sigma_X = \sqrt{\operatorname{Var}[X]}$.

- One uses the standard deviation, which is defined as the square root of the variance to measure of the *spread* of the possible values of X.
- It is useful to work with the standard deviation since it has the same units as $\mathbb{E}X$.
- $\sigma_{aX+b} = |a|\, \sigma_X$.
- Informally we think of outcomes within $\pm\sigma_X$ of $\mathbb{E}X$ as being in the center of the distribution. Some references would informally interpret sample values within $\pm\sigma_X$ of

the expected value, $x \in [\mathbb{E}X - \sigma_X, \mathbb{E}X + \sigma_X]$, as "typical" values of X and other values as "unusual".

8.76. In finance, standard deviation is a key concept and is used to measure the ***volatility*** (risk) of investment returns and stock returns.

It is common wisdom in finance that diversification of a portfolio of stocks generally reduces the total risk exposure of the investment. We shall return to this point in Example 9.52.

Example 8.77. Continue from Example 8.73. If the standard deviation of exam scores is 12 points, the student with a score of +7 with respect to the mean can think of herself in the middle of the class. If the standard deviation is 3 points, she is likely to be near the top.

Example 8.78. Suppose $X \sim \text{Bernoulli}(p)$.

(a) $\mathbb{E}\left[X^2\right] = 0^2 \times (1-p) + 1^2 \times p = p.$

(b) $\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2 = p(1-p).$
Alternatively, if we directly use (17), we have

$$\begin{aligned}\operatorname{Var} X = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] &= (0-p)^2 \times (1-p) + (1-p)^2 \times p \\ &= p(1-p)(p + (1-p)) = p(1-p).\end{aligned}$$

Example 8.79. Continue from Example 8.54 and Example 8.70. Suppose $X \sim \mathcal{P}(\alpha)$. We have

$$\operatorname{Var} X = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2 = \alpha^2 + \alpha - \alpha^2 = \alpha.$$

Therefore, for Poisson random variable, the expected value is the same as the variance.

Example 8.80. For X uniform on `[N:M]` (the set of integers from N to M), we have

$$\mathbb{E}X = \frac{M+N}{2}$$

and

$$\operatorname{Var} X = \frac{1}{12}(M-N)(M-N-2) = \frac{1}{12}(n^2-1),$$

where $n = M - N + 1$. [22, p 280]

- For X uniform on `[-M:M]`, we have $\mathbb{E}X = 0$ and $\operatorname{Var} X = \frac{M(M+1)}{3}$.

Example 8.81. Continue from Example 8.64. Suppose X is a binary random variable with

$$P[X = b] = 1 - P[X = a] = p.$$

(a) $\operatorname{Var} X = (b-a)^2 \operatorname{Var} I = (b-a)^2 p(1-p)$.

(b) Suppose $a = -b$. Then, $X = -2aI + a = 2bI - b$. In which case, $\operatorname{Var} X = 2b^2 p(1-p)$.

Example 8.82. Consider the two pmfs shown in Figure 5. The random variable X with pmf at the left has a smaller variance than the random variable Y with pmf at the right because more probability mass is concentrated near zero (their mean) in the graph at the left than in the graph at the right. [9, p. 85]

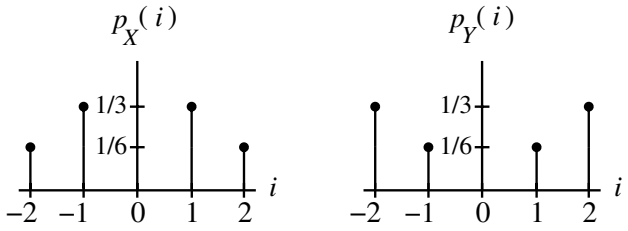

Figure 5: Example 8.82 shows that a random variable whose probability mass is concentrated near the mean has smaller variance. [9, Fig. 2.9]

8.83. We have already talked about variance and standard deviation as a number that indicates spread/dispersion of the pmf. More specifically, let's imagine a pmf that shapes like a bell curve. As the value of σ_X gets smaller, the spread of the pmf will be smaller and hence the pmf would "look sharper". Therefore, the probability that the random variable X would take a value that is far from the mean would be smaller.

The next property involves the use of σ_X to bound "the tail probability" of a random variable.

8.84. *Chebyshev's Inequality*:

$$P\left[|X - \mathbb{E}X| \geq \alpha\right] \leq \frac{\sigma_X^2}{\alpha^2}$$

or equivalently

$$P\left[|X - \mathbb{E}X| \geq n\sigma_X\right] \leq \frac{1}{n^2}$$

- Useful only when $\alpha > \sigma_X$

Definition 8.85. More definitions involving expectation of a function of a random variable:

(a) ***Coefficient of Variation***: $CV_X = \frac{\sigma_X}{\mathbb{E}X}$.

- It is the standard deviation of the "normalized" random variable $\frac{X}{\mathbb{E}X}$.
- 1 for exponential.

(b) ***Fano Factor*** (index of dispersion): $\frac{\operatorname{Var} X}{\mathbb{E}X}$.

- 1 for Poisson.

(c) ***Central Moments***: A generalization of the variance is the nth central moment which is defined to be $\mu_n = \mathbb{E}\left[(X - \mathbb{E}X)^n\right]$.

(i) $\mu_1 = \mathbb{E}\left[X - \mathbb{E}X\right] = 0$.

(ii) $\mu_2 = \sigma_X^2 = \operatorname{Var} X$: the second central moment is the variance.

Example 8.86. If X has mean m and variance σ^2, it is sometimes convenient to introduce the normalized random variable

$$Y = \frac{X - m}{\sigma}.$$

ECS315 2011/1 Part III.4 Dr.Prapun

9 Multiple Random Variables

One is often interested not only in individual random variables, but also in relationships between two or more random variables. Furthermore, one often wishes to make inferences about one random variable on the basis of observations of other random variables.

Example 9.1. If the experiment is the testing of a new medicine, the researcher might be interested in cholesterol level, blood pressure, and the glucose level of a test person.

9.1 A Pair of Random Variables

Definition 9.2. If X and Y are random variables, we use the shorthand

$$\begin{aligned}[X \in B, Y \in C] &= [X \in B \text{ and } Y \in C] \\ &= \{\omega \in \Omega : X(\omega) \in B \text{ and } Y(\omega) \in C\} \\ &= \{\omega \in \Omega : X(\omega) \in B\} \cap \{\omega \in \Omega : Y(\omega) \in C\} \\ &= [X \in B] \cap [Y \in C]\,.\end{aligned}$$

- Observe that the "," in $[X \in B, Y \in B]$ means "and".

Consequently,

$$\begin{aligned}P[X \in B, Y \in C] &= P[X \in B \text{ and } Y \in C] \\ &= P([X \in B] \cap [Y \in C])\,.\end{aligned}$$

Similarly, the concept of conditional probability can be straightforwardly applied to random variables via

$$P\left[X \in B \middle| Y \in C\right] = P\left([X \in B] \middle| [Y \in C]\right) = \frac{P\left([X \in B] \cap [Y \in C]\right)}{P\left([Y \in C]\right)}$$
$$= \frac{P\left[X \in B, Y \in C\right]}{P\left[Y \in C\right]}.$$

Example 9.3. We also have

$$P\left[X = x, Y = y\right] = P\left[X = x \text{ and } Y = y\right],$$
$$P\left[X = x \middle| Y = y\right] = \frac{P\left[X = x \text{ and } Y = y\right]}{P\left[Y = y\right]},$$

and

$$P\left[3 \leq X < 4, Y < 1\right] = P\left[3 \leq X < 4 \text{ and } Y < 1\right]$$
$$= P\left[X \in [3, 4) \text{ and } Y \in (-\infty, 1)\right].$$
$$P\left[3 \leq X < 4 \middle| Y < 1\right] = \frac{P\left[3 \leq X < 4 \text{ and } Y < 1\right]}{P\left[Y < 1\right]}$$

Definition 9.4. ***Joint pmf***: If X and Y are two discrete random variables (defined on a same sample space with probability measure P), the probability mass function $p_{X,Y}(x, y)$ defined by

$$p_{X,Y}(x, y) = P\left[X = x, Y = y\right]$$

is called the ***joint probability mass function*** of X and Y. We can then evaluate $P\left[(X, Y) \in R\right]$ by $\sum_{(x,y):(x,y)\in R} p_{X,Y}(x, y)$.

Definition 9.5. The ***joint cdf*** of X and Y is defined by

$$F_{X,Y}(x, y) = P\left[X \leq x, Y \leq y\right].$$

Definition 9.6. The ***conditional pmf*** of X given Y is defined as

$$p_{X|Y}(x|y) = P\left[X = x | Y = y\right]$$

which gives

$$p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x).$$

Example 9.7. Toss-and-Roll Game:

Step 1 Toss a fair coin. Define X by

$$X = \begin{cases} 1, & \text{if result} = \text{H}, \\ 0, & \text{if result} = \text{T}. \end{cases}$$

Step 2 You have two dice, Dice 1 and Dice 2. Dice 1 is fair. Dice 2 is unfair with $p(1) = p(2) = p(3) = \frac{2}{9}$ and $p(4) = p(5) = p(6) = \frac{1}{9}$.

(i) If $X = 0$, roll Dice 1.

(ii) If $X = 1$, roll Dice 2.

Record the result as Y.

Definition 9.8. When X and Y take finitely many values (have finite supports), say $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, respectively, we can arrange the probabilities $p_{X,Y}(x_i, y_j)$ in the $m \times n$ matrix

$$\begin{bmatrix} p_{X,Y}(x_1, y_1) & p_{X,Y}(x_1, y_2) & \ldots & p_{X,Y}(x_1, y_n) \\ p_{X,Y}(x_2, y_1) & p_{X,Y}(x_2, y_2) & \ldots & p_{X,Y}(x_2, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ p_{X,Y}(x_m, y_1) & p_{X,Y}(x_m, y_2) & \ldots & p_{X,Y}(x_m, y_n) \end{bmatrix}.$$

- The sum of all the entries in the matrix is one.
- The sum of the entries in the ith row is $p_X(x_i)$, and the sum of the entries in the jth column is $p_Y(y_j)$:

$$p_X(x_i) = \sum_{j=1}^{n} p_{X,Y}(x_i, y_j) \tag{18}$$

$$p_Y(y_j) = \sum_{i=1}^{m} p_{X,Y}(x_i, y_j) \tag{19}$$

To show (18), we consider $A = [X = x_i]$ and a collection defined by $B_j = [Y = y_j]$ and $B_0 = [Y \notin \{y_1, \ldots, y_n\}]$. Note that the collection $B_0, B_1, \ldots, B_n$ partitions Ω. So, $P(A) = \sum_{j=0}^{n} P(A \cap B_j)$. Of course, because the support of Y is $\{y_1, \ldots, y_n\}$, we have $P(A \cap B_0) = 0$. Hence, the sum can start at $j = 1$ instead of $j = 0$.

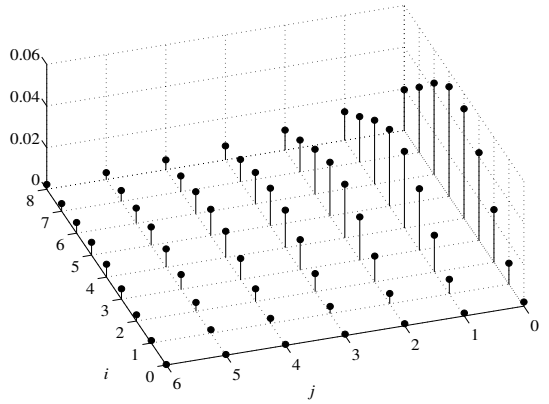

Figure 6: Example of the plot of a joint pmf. [9, Fig. 2.8]

9.9. From the joint pmf, we can find $p_X(x)$ and $p_Y(y)$ by

$$p_X(x) = \sum_y p_{X,Y}(x, y) \tag{20}$$

$$p_Y(y) = \sum_x p_{X,Y}(x, y) \tag{21}$$

In this setting, $p_X(x)$ and $p_Y(y)$ are call the ***marginal pmfs*** (to distinguish them from the joint one).

In `MATLAB`, if we define the joint pmf matrix as `P_XY`, then the marginal pmf (row) vectors `p_X` and `p_Y` can be found by

```
p_X = (sum(P_XY,2))'
p_Y = (sum(P_XY,1))
```

Example 9.10. Consider the following joint pmf matrix

Definition 9.11. Two random variables X and Y are said to be ***identically distributed*** if, for every B, $P[X \in B] = P[Y \in B]$.

9.12. The following statements are equivalent:

(a) Random variables X and Y are ***identically distributed***.

(b) For every B, $P[X \in B] = P[Y \in B]$

(c) $p_X(c) = p_Y(c)$ for all c

(d) $F_X(c) = F_Y(c)$ for all c

Definition 9.13. Two random variables X and Y are said to be ***independent*** if the events $[X \in B]$ and $[Y \in C]$ are independent for all sets B and C.

9.14. The following statements are equivalent:

(a) Random variables X and Y are ***independent***.

(b) $[X \in B] \perp\!\!\!\perp [Y \in C]$ for all B, C.

(c) $P[X \in B, Y \in C] = P[X \in B] \times P[Y \in C]$ for all B, C.

(d) $p_{X,Y}(x, y) = p_X(x) \times p_Y(y)$ for all x, y.

(e) $F_{X,Y}(x, y) = F_X(x) \times F_Y(y)$ for all x, y.

Definition 9.15. Two random variables X and Y are said to be ***independent and identically distributed (i.i.d.)*** if X and Y are both independent and identically distributed.

9.2 Extending the Definitions to Multiple RVs

Definition 9.16. Joint pmf:

$$p_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = P\left[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right].$$

Joint cdf:

$$F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = P\left[X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n\right].$$

Definition 9.17. ***Identically distributed*** random variables: The following statements are equivalent.

(a) Random variables $X_1, X_2, \ldots$ are ***identically distributed***

(b) For every B, $P\left[X_j \in B\right]$ does not depend on j.

(c) $p_{X_i}(x) = p_{X_j}(x)$ for all x, i, j.

(d) $F_{X_i}(x) = F_{X_j}(x)$ for all x, i, j.

Definition 9.18. ***Independence*** among finite number of random variables: The following statements are equivalent.

(a) $X_1, X_2, \ldots, X_n$ are ***independent***

(b) $[X_1 \in B_1], [X_2 \in B_2], \ldots, [X_n \in B_n]$ are independent, for all $B_1, B_2, \ldots, B_n$.

(c) $P\left[X_i \in B_i, \forall i\right] = \prod_{i=1}^n P\left[X_i \in B_i\right]$, for all $B_1, B_2, \ldots, B_n$.

(d) $p_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = \prod_{i=1}^n p_{X_i}(x_i)$ for all $x_1, x_2, \ldots, x_n$.

(e) $F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = \prod_{i=1}^n F_{X_i}(x_i)$ for all $x_1, x_2, \ldots, x_n$.

Example 9.19. Toss a coin n times. For the ith toss, let

$$X_i = \begin{cases} 1, & \text{if H happens on the } i\text{th toss,} \\ 0, & \text{if T happens on the } i\text{th toss.} \end{cases}$$

We then have a collection of i.i.d. random variables $X_1, X_2, X_3, \ldots, X_n$.

Example 9.20. Roll a dice n times. Let N_i be the result of the ith roll. We then have another collection of i.i.d. random variables $N_1, N_2, N_3, \ldots, N_n$.

Example 9.21. Let X_1 be the result of tossing a coin. Set $X_2 = X_3 = \cdots = X_n = X_1$.

9.22. If $X_1, X_2, \ldots, X_n$ are independent, then so is any subset of them.

9.23. For i.i.d. $X_i \sim \text{Bernoulli}(p)$, $Y = X_1 + X_2 + \cdots + X_n$ is $\mathcal{B}(n, p)$.

Definition 9.24. A ***pairwise independent*** collection of random variables is a set of random variables any two of which are independent.

(a) Any collection of (mutually) independent random variables is pairwise independent

(b) Some pairwise independent collections are not independent. See Example (9.26).

Definition 9.25. A family of random variables $\{X_i : i \in I\}$ is ***independent*** if $\forall$ finite $J \subset I$, the family of random variables $\{X_i : i \in J\}$ is independent. In words, "an infinite collection of random elements is by definition independent if each finite subcollection is." Hence, we only need to know how to test independence for finite collection.

Example 9.26. Let suppose X, Y, and Z have the following joint probability distribution: $p_{X,Y,Z}(x, y, z) = \frac{1}{4}$ for $(x, y, z) \in \{(0, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 0)\}$. This, for example, can be constructed by starting with independent X and Y that are Bernoulli-$\frac{1}{2}$. Then set $Z = X \oplus Y = X + Y \mod 2$.

(a) X, Y, Z are pairwise independent.

(b) X, Y, Z are not independent.

ECS315 2011/1 Part III.5 Dr.Prapun

Example 9.27. Suppose the pmf of a random variable X is given by

$$p_X(x) = \begin{cases} 1/4, & x = 3, \\ \alpha, & x = 4, \\ 0, & \text{otherwise.} \end{cases}$$

Let Y be another random variable. Assume that X and Y are i.i.d.

Find

(a) α,

(b) the pmf of Y, and

(c) the joint pmf of X and Y.

Example 9.28. Consider a pair of random variables X and Y whose joint pmf is given by

$$p_{X,Y}(x,y) = \begin{cases} 1/15, & x=3, y=1, \\ 2/15, & x=4, y=1, \\ 4/15, & x=3, y=3, \\ \beta, & x=4, y=3, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Are X and Y identically distributed?

(b) Are X and Y independent?

9.3 Function of Discrete Random Variables

9.29. For discrete random variable X, the pmf of a derived random variable $Y = g(X)$ is given by

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

Similarly, for discrete random variables X and Y, the pmf of a derived random variable $Z = g(X, Y)$ is given by

$$p_Z(z) = \sum_{(x,y):g(x,y)=z} p_{X,Y}(x, y).$$

Example 9.30. Suppose the joint pmf of X and Y is given by

$$p_{X,Y}(x, y) = \begin{cases} 1/15, & x = 0, y = 0, \\ 2/15, & x = 1, y = 0, \\ 4/15, & x = 0, y = 1, \\ 8/15, & x = 1, y = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Let $Z = X + Y$. Find the pmf of Z.

9.31. In general, when $Z = X + Y$,

$$\begin{aligned} p_Z(z) &= \sum_{(x,y):x+y=z} p_{X,Y}(x,y) \\ &= \sum_y p_{X,Y}(z-y,y) = \sum_x p_{X,Y}(x,z-x). \end{aligned}$$

Furthermore, if X and Y are independent,

$$\begin{aligned} p_Z(z) &= \sum_{(x,y):x+y=z} p_X(x)\, p_Y(y) \\ &= \sum_y p_X(z-y)\, p_Y(y) = \sum_x p_X(x)\, p_Y(z-x). \end{aligned}$$

Example 9.32. Suppose $\Lambda_1 \sim \mathcal{P}(\lambda_1)$ and $\Lambda_2 \sim \mathcal{P}(\lambda_2)$ are independent. Find the pmf of $\Lambda = \Lambda_1 + \Lambda_2$.

First, note that $p_\Lambda(x)$ would be positive only on nonnegative integers because a sum of nonnegative integers is still a nonnegative integer. So, the support of Λ is the same as the support for Λ_1 and Λ_2. Now, we know that

$$P[\Lambda = k] = P[\Lambda_1 + \Lambda_2 = k] = \sum_i P[\Lambda_1 = i]\, P[\Lambda_2 = k - i]$$

Of course, we are interested in k that is a nonnegative integer. The summation runs over $i = 0, 1, 2, \ldots$. Other values of i would make $P[\Lambda_1 = i] = 0$. Note also that if $i > k$, then $k - i < 0$ and $P[\Lambda_2 = k - i] = 0$. Hence, we conclude that the index i can only be integers from 0 to k:

$$\begin{aligned} P[\Lambda = k] &= \sum_{i=0}^{k} e^{-\lambda_1} \frac{\lambda_1^i}{i!} e^{-\lambda_2} \frac{\lambda_2^{k-i}}{(k-i)!} \\ &= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{i=0}^{k} \frac{k!}{i!\,(k-i)!} \lambda_1^i \lambda_2^{k-i} \\ &= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{i=0}^{k} \lambda_1^i \lambda_2^{k-i} = e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^k}{k!}, \end{aligned}$$

where the last equality is from the binomial theorem. Hence, the sum of two independent Poisson random variables is still Poisson!

There are a couple of interesting results that are related to the above example:

- **Finite additivity**: Suppose we have independent $\Lambda_i \sim \mathcal{P}(\lambda_i)$, then $\sum_{i=1}^n \Lambda_i \sim \mathcal{P}(\sum_{i=1}^n \lambda_i)$.
- **Raikov's theorem**: Independent random variables can have their sum Poisson-distributed only if every component of the sum is Poisson-distributed.

9.4 Expectation of function of discrete random variables

9.33. Suppose X is a discrete random variable.

$$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x).$$

Similarly,

$$\mathbb{E}[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y).$$

These are called the **law/rule of the lazy statistician** (LOTUS) [23, Thm 3.6 p 48],[9, p. 149] because it is so much easier to use the above formula than to first find the pmf of Y. It is also called **substitution rule** [22, p 271].

9.34. $\mathbb{E}[\cdot]$ is a **linear** operator: $\mathbb{E}[aX + bY] = a\mathbb{E}X + b\mathbb{E}Y$.

(a) Homogeneous: $\mathbb{E}[cX] = c\mathbb{E}X$

(b) Additive: $\mathbb{E}[X + Y] = \mathbb{E}X + \mathbb{E}Y$

(c) Extension: $\mathbb{E}[\sum_{i=1}^n c_i X_i] = \sum_{i=1}^n c_i \mathbb{E}X_i$.

Example 9.35. Recall from 9.23 that when i.i.d. $X_i \sim \text{Bernoulli}(p)$, $Y = X_1 + X_2 + \cdots X_n$ is $\mathcal{B}(n,p)$. Also, from Example 8.51, we have $\mathbb{E}X_i = p$. Hence,

$$\mathbb{E}Y = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n p = np.$$

Therefore, the expectation of a binomial random variable with parameters n and p is np.

	Discrete
$P[X \in B]$	$\sum_{x \in B} p_X(x)$
$P[(X,Y) \in R]$	$\sum_{(x,y):(x,y)\in R} p_{X,Y}(x,y)$
Joint to Marginal: (Law of Total Prob.)	$p_X(x) = \sum_y p_{X,Y}(x,y)$ $p_Y(y) = \sum_x p_{X,Y}(x,y)$
$P[X > Y]$	$\sum_x \sum_{y:\, y<x} p_{X,Y}(x,y)$ $= \sum_y \sum_{x:\, x>y} p_{X,Y}(x,y)$
$P[X = Y]$	$\sum_x p_{X,Y}(x,x)$
$X \perp\!\!\!\perp Y$	$p_{X,Y}(x,y) = p_X(x)p_Y(y)$
Conditional	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$
$\mathbb{E}[g(X,Y)]$	$\sum_x \sum_y g(x,y) p_{X,Y}(x,y)$
$P[g(X,Y) \in B]$	$\sum_{(x,y):\, g(x,y)\in B} p_{X,Y}(x,y)$
$Z = X + Y$	$p_Z(z) = \sum_x p_{X,Y}(x, z-x)$

Table 4: Joint pmf: A Summary

Example 9.36. A binary communication link has bit-error probability p. What is the expected number of bit errors in a transmission of n bits?

Theorem 9.37 (Expectation and Independence)**.** Two random variables X and Y are independent if and only if

$$\mathbb{E}[h(X)g(Y)] = \mathbb{E}[h(X)]\,\mathbb{E}[g(Y)]$$

for all functions h and g.

- In other words, X and Y are independent if and only if for every pair of functions h and g, the expectation of the product $h(X)g(Y)$ is equal to the product of the individual expectations.
- One special case is that

$$X \perp\!\!\!\perp Y \quad \text{implies} \quad \mathbb{E}[XY] = \mathbb{E}X \times \mathbb{E}Y. \tag{22}$$

However, independence means more than this property. In other words, having $\mathbb{E}[XY] = (\mathbb{E}X)(\mathbb{E}Y)$ does not necessarily imply $X \perp\!\!\!\perp Y$. See Example 9.46.

9.38. It is useful to incorporate what we have just learned about independence into the definition that we already have.

The following statements are equivalent:

(a) Random variables X and Y are ***independent***.

(b) $[X \in B] \perp\!\!\!\perp [Y \in C]$ for all B, C.

(c) $P[X \in B, Y \in C] = P[X \in B] \times P[Y \in C]$ for all B, C.

(d) $p_{X,Y}(x, y) = p_X(x) \times p_Y(y)$ for all x, y.

(e) $F_{X,Y}(x, y) = F_X(x) \times F_Y(y)$ for all x, y.

(f)

9.39. To quantify the amount of dependence between two random variables, we may calculate their ***mutual information***. This quantity is crucial in the study of digital communications and information theory. However, in introductory probability class (and introductory communication class), it is traditionally omitted.

9.5 Linear Dependence

Definition 9.40. Given two random variables X and Y, we may calculate the following quantities:

(a) **Correlation**: $\mathbb{E}[XY]$.

(b) **Covariance**: $\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$.

(c) **Correlation coefficient**: $\rho_{X,Y} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$

9.41. $\operatorname{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y$

- Note that $\operatorname{Var} X = \operatorname{Cov}[X, X]$.

9.42. $\operatorname{Var}[X + Y] = \operatorname{Var} X + \operatorname{Var} Y + 2\operatorname{Cov}[X, Y]$

Definition 9.43. X and Y are said to be ***uncorrelated*** if and only if $\operatorname{Cov}[X, Y] = 0$.

9.44. The following statements are equivalent:

(a) X and Y are ***uncorrelated***.

(b) $\operatorname{Cov}[X, Y] = 0$.

(c) $\mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y$.

9.45. If $X \perp\!\!\!\perp Y$, then $\operatorname{Cov}[X, Y] = 0$. The converse is not true. Being uncorrelated does not imply independence.

Example 9.46. Let X be uniform on $\{\pm 1, \pm 2\}$ and $Y = |X|$.

Example 9.47. Suppose two fair dice are tossed. Denote by the random variable V_1 the number appearing on the first die and by the random variable V_2 the number appearing on the second die. Let $X = V_1 + V_2$ and $Y = V_1 - V_2$.

(a) X and Y are not independent.

(b) $\mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y$.

Definition 9.48. X and Y are said to be ***orthogonal*** if $\mathbb{E}[XY] = 0$.

9.49. When $\mathbb{E}X = 0$ or $\mathbb{E}Y = 0$, orthogonality is equivalent to uncorrelatedness.

Definition 9.50. ***Correlation coefficient***:

$$\begin{aligned}\rho_{X,Y} &= \frac{\mathrm{Cov}[X,Y]}{\sigma_X\sigma_Y} \\ &= \mathbb{E}\left[\left(\frac{X-\mathbb{E}X}{\sigma_X}\right)\left(\frac{Y-\mathbb{E}Y}{\sigma_Y}\right)\right] = \frac{\mathbb{E}[XY]-\mathbb{E}X\mathbb{E}Y}{\sigma_X\sigma_Y}.\end{aligned}$$

- $\rho_{X,Y}$ is dimensionless
- $\rho_{X,X} = 1$
- $\rho_{X,Y} = 0$ if and only if X and Y are uncorrelated.

9.51. Linear Dependence and Caychy-Schwartz Inequality

(a) If $Y = aX + b$, then $\rho_{X,Y} = \mathrm{sign}(a) = \begin{cases} 1, & a > 0 \\ -1, & a < 0. \end{cases}$

- To be rigorous, we should also require that $\sigma_X > 0$ and $a \neq 0$.

(b) Cauchy-Schwartz Inequality:

$$(\operatorname{Cov}[X,Y])^2 \leq \sigma_X^2 \sigma_Y^2$$

(c) This implies $|\rho_{X,Y}| \leq 1$. In other words, $\rho_{XY} \in [-1,1]$.

(d) When $\sigma_Y, \sigma_X > 0$, equality occurs if and only if the following conditions holds

$$\begin{aligned}
&\equiv \exists a \neq 0 \text{ such that } (X - \mathbb{E}X) = a(Y - \mathbb{E}Y) \\
&\equiv \exists a \neq 0 \text{ and } b \in \mathbb{R} \text{ such that } X = aY + b \\
&\equiv \exists c \neq 0 \text{ and } d \in \mathbb{R} \text{ such that } Y = cX + d \\
&\equiv |\rho_{XY}| = 1
\end{aligned}$$

In which case, $|a| = \frac{\sigma_X}{\sigma_Y}$ and $\rho_{XY} = \frac{a}{|a|} = \operatorname{sgn} a$. Hence, ρ_{XY} is used to quantify **linear dependence** between X and Y. The closer $|\rho_{XY}|$ to 1, the higher degree of linear dependence between X and Y.

Example 9.52. [22, Section 5.2.3] Consider an important fact that *investment experience* supports: spreading investments over a variety of funds (diversification) diminishes risk. To illustrate, imagine that the random variable X is the return on every invested dollar in a local fund, and random variable Y is the return on every invested dollar in a foreign fund. Assume that random variables X and Y are i.i.d. with expected value 0.15 and standard deviation 0.12.

If you invest all of your money, say c, in either the local or the foreign fund, your return R would be cX or cY, The expected return is $\mathbb{E}R = c\mathbb{E}X = c\mathbb{E}Y = c \times 0.15$.

Now imagine that your money is equally distributed over the two funds. Then, the return R is $\frac{1}{2}cX + \frac{1}{2}cY$. The expected return is $\mathbb{E}R = \frac{1}{2}c\mathbb{E}X + \frac{1}{2}c\mathbb{E}Y = 0.15 \times c$. Hence, the expected return remains at 15%. However,

$$\operatorname{Var}\left[\frac{1}{2}(X+Y)\right] = \frac{1}{4}\operatorname{Var}X + \frac{1}{4}\operatorname{Var}Y = \frac{1}{2} \times 0.12.$$

So, the standard deviation is $\frac{0.12}{\sqrt{2}} \approx 0.0849$.

In comparison with the distributions of X and Y, the pmf of $\frac{1}{2}(X+Y)$ is concentrated more around the expected value. The centralization of the distribution as random variables are averaged together is a manifestation of the central limit theorem which we will soon discuss.

9.53. [22, Section 5.2.3] Example 9.52 is based on the assumption that return rates X and Y are independent from each other. In the world of investment, however, risks are more commonly reduced by combining negatively correlated funds (two funds are negatively correlated when one tends to go up as the other falls).

This becomes clear when one considers the following hypothetical situation. Suppose that two stock market outcomes ω_1 and ω_2 are possible, and that each outcome will occur with a probability of $\frac{1}{2}$ Assume that domestic and foreign fund returns X and Y are determined by $X(\omega_1) = Y(\omega_2) = 0.25$ and $X(\omega_2) = Y(\omega_1) = -0.10$. Each of the two funds then has an expected return of 7.5%, with equal probability for actual returns of 25% and .10%. The random variable $Z = \frac{1}{2}(X+Y)$ satisfies $Z(\omega_1) = Z(\omega_2) = 0.075$. In other words, Z is equal to 0.075 with certainty. This means that an investment that is equally divided between the domestic and foreign funds has a guaranteed return of 7.5%.

Example 9.54. The input X and output Y of a system subject to random perturbations are described probabilistically by the following joint pmf matrix:

x \ y	2	4	5
1	0.02	0.10	0.08
3	0.08	0.32	0.40

(a) Create a `MATLAB` m-file to evaluate the following quantities.

(i) $\mathbb{E}X$

(ii) $P[X = Y]$

(iii) $P[XY < 6]$

(iv) $\mathbb{E}\left[(X-3)(Y-2)\right]$

(v) $\mathbb{E}\left[X(Y^3 - 11Y^2 + 38Y)\right]$

(vi) $\mathrm{Cov}\left[X, Y\right]$

(vii) $\rho_{X,Y}$

(b) Calculate the following quantities using what you got from part (a).

(i) $\mathrm{Cov}\left[3X + 4, 6Y - 7\right]$

(ii) $\rho_{3X+4,6Y-7}$

(iii) $\mathrm{Cov}\left[X, 6X - 7\right]$

(iv) $\rho_{X,6X-7}$

Solution:

(a) The MATLAB code is provided in a separate file.

(i) $\mathbb{E}X = 2.6$

(ii) $P[X = Y] = 0$

(iii) $P[XY < 6] = 0.2$

(iv) $\mathbb{E}\left[(X-3)(Y-2)\right] = -0.88$

(v) $\mathbb{E}\left[X(Y^3 - 11Y^2 + 38Y)\right] = 104$

(vi) $\mathrm{Cov}\left[X, Y\right] = 0.032$

(vii) $\rho_{X,Y} = 0.0447$

(b)

(i) Note that

$$\begin{aligned}
\mathrm{Cov}\left[aX+b, cY+d\right] &= \mathbb{E}\left[\left((aX+b) - \mathbb{E}\left[aX+b\right]\right)\left((cY+d) - \mathbb{E}\left[cY+d\right]\right)\right] \\
&= \mathbb{E}\left[\left((aX+b) - (a\mathbb{E}X+b)\right)\left((cY+d) - (c\mathbb{E}Y+d)\right)\right] \\
&= \mathbb{E}\left[(aX - a\mathbb{E}X)(cY - c\mathbb{E}Y)\right] \\
&= ac\mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right] \\
&= ac\mathrm{Cov}\left[X, Y\right].
\end{aligned}$$

Hence, $\mathrm{Cov}\left[3X+4, 6Y-7\right] = 3 \times 6 \times \mathrm{Cov}\left[X, Y\right] \approx 3 \times 6 \times 0.032 \approx \boxed{0.576}$.

(ii) Note that

$$\begin{aligned}
\rho_{aX+b,cY+d} &= \frac{\mathrm{Cov}\left[aX+b, cY+d\right]}{\sigma_{aX+b}\sigma_{cY+d}} \\
&= \frac{ac\mathrm{Cov}\left[X, Y\right]}{|a|\sigma_X|c|\sigma_Y} = \frac{ac}{|ac|}\rho_{X,Y} = \mathrm{sign}(ac) \times \rho_{X,Y}.
\end{aligned}$$

Hence, $\rho_{3X+4,6Y-7} = \mathrm{sign}(3 \times 4)\rho_{X,Y} = \rho_{X,Y} = \boxed{0.0447}$.

(iii) $\mathrm{Cov}\left[X, 6X-7\right] = 1 \times 6 \times \mathrm{Cov}\left[X, X\right] = 6 \times \mathrm{Var}[X] \approx \boxed{3.84}$.

(iv) $\rho_{X,6X-7} = \mathrm{sign}(1 \times 6) \times \rho_{X,X} = \boxed{1}$.

ECS315 2011/1 Part IV.1 Dr.Prapun

10 Continuous Random Variables

10.1 From Discrete to Continuous Random Variables

In many practical applications of probability, physical situations are better described by random variables that can take on a *continuum* of possible values rather than a *discrete* number of values. The interesting fact is that, for this type of random variable, any individual value has probability zero:

$$P[X = x] = 0 \quad \text{for all } x. \tag{23}$$

These random variables are called **continuous random variables**.

10.1. We can already see from (23) that the pmf is going to be useless for this type of random variable. It turns out that the cdf F_X is still useful and we shall introduce another useful function called pdf to replace the role of pmf. However, integral calculus[18] is required to formulate this continuous analog of a pmf.

Example 10.2. If you can measure the heights of people with infinite precision, the height of a randomly chosen person is a continuous random variable. In reality, heights cannot be measured with infinite precision, but the mathematical analysis of the distribution of heights of people is greatly simplified when using a mathematical model in which the height of a randomly chosen person is modeled as a continuous random variable. [22, p 284]

[18] This is always a difficult concept for the beginning student.

Example 10.3. Continuous random variables are important models for

(a) voltages in communication receivers

(b) file download times on the Internet

(c) velocity and position of an airliner on radar

(d) lifetime of a battery

(e) decay time of a radioactive particle

(f) time until the occurrence of the next earthquake in a certain region

Example 10.4. The most simple example of a continuous random variable is the "random choice" of a number from the interval $(0, 1)$.

- In `MATLAB`, this can be generated by the command `rand`.
- The generation is "unbiased" in the sense that "any number in the range is as likely to occur as another number."
- Again, the probability that the randomly chosen number will take on a pre-specified value is zero.
    - So the above statement is true but not useful because it is true for all continuous random variables anyway. For any continuous random variable, the probability of a particular value x is 0 for any x; so they all have the same probability.
- We can speak of the probability of the randomly chosen number falling in a given subinterval of (0,1).
    - For numbers generated from the command `rand`, this probability is equal to the length of that subinterval.

- For example, if a dart is thrown at random to the interval $(0, 1)$,
  - the probability of the dart hitting exactly the point 0.25 is zero,
  - but the probability of the dart landing somewhere in the interval between 0.2 and 0.3 is 0.1 (assuming that the dart has an infinitely thin point).

  - No matter how small Δx is, any subinterval of the length Δx has probability Δx of containing the point at which the dart will land.

Definition 10.5. We say that X is a **continuous random variable**[19] if there exists a (real-valued) function f such that, for any event B, $P[X \in B]$ has the form

$$P[X \in B] = \int_B f(x)dx.$$

- The function f is called the ***probability density function*** (pdf) or simply ***density***.
- When we want to emphasize that the function f is a density of a particular random variable X, we write f_X instead of f.
- Recall that when X is a discrete random variable,

$$P[X \in B] = \sum_{x \in B} p_X(x).$$

Definition 10.6. Indicator function:

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

[19] To be more rigorous, this is the definition for *absolutely* continuous random variable. At this level, we will not distinguish between the continuous random variable and absolutely continuous random variable. When the distinction between them is considered, a random variable X is said to be continuous (not necessarily absolutely continuous) when condition (23) is satisfied. Alternatively, condition (23) is equivalent to requiring the cdf F_X to be continuous. Another fact worth mentioning is that if a random variable is absolutely continuous, then it is continuous. So, absolute continuity is a stronger condition.

For example, $1_{[a,b)}(x)$ is shown in Figure 7a, and $1_{(a,b]}(x)$ is shown in Figure 7b. Some references use $I_A(x)$ instead of $1_A(x)$.

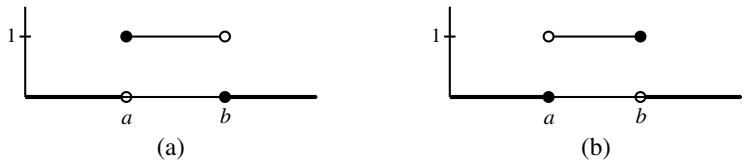

Figure 7: (a) Indicator function $I_{[a,b)}(x)$. (b) Indicator function $I_{(a,b]}(x)$. [9, Fig. 2.10]

Example 10.7. For the random variable generated by the `rand` command in `MATLAB`, $f(x) = 1_{(0,1)}(x)$.

10.8. The `rand` command in `MATLAB` is an approximation for two reasons:

(a) It produces pseudorandom numbers; the numbers seem random but are actually the output of a deterministic algorithm.

(b) It produces a double precision floating point number, represented in the computer by 64 bits. Thus MATLAB distinguishes no more than 2^{64} unique double precision floating point numbers. By comparison, there are uncountably infinite real numbers in the interval from 0 to 1.

10.2 Properties of PDF and CDF for Continuous Random Variables

10.9. f_X is determined only almost everywhere[20]. That is, if we construct a function g by changing the function f at a countable number of points[21], then g can also serve as a density for X.

Example 10.10. For the random variable generated by the `rand` command in `MATLAB`,

10.11. Recall that the cdf of a random variable X is defined as

$$F_X(x) = P\left[X \leq x\right].$$

10.12. Note that even through there are more than one valid pdf's for any given random variable, the cdf is unique. There is only one cdf for each random variable.

10.13. From the pdf $f_X(x)$, we can find the cdf of X by

$$F_X(x) = P\left[X \leq x\right] = P\left[X \in (-\infty, x]\right] = \int_{-\infty}^{x} f_X(t)dt.$$

[20] Lebesgue-a.e, to be exact

[21] More specifically, if $g = f$ Lebesgue-a.e., then g is also a pdf for X.

- If F_X is differentiable at x,

$$\frac{d}{dx}F_X(x) = f_X(x).$$

- In general F_X need not differentiate to f_X everywhere.

Example 10.14. For the random variable generated by the `rand` command in `MATLAB`,

Example 10.15. Suppose that the lifetime X of a device has the cdf

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}x^2, & 0 \le x \le 2 \\ 1, & x > 2 \end{cases}$$

Observe that it is differentiable at each point x except for the two points $x = 0$ and $x = 2$. The probability density function is obtained by differentiation of the cdf which gives

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

In each of the finite number of points x at which F_X has no derivative, it does not matter what value we give f_X. These values do not affect $\int_B f_X(x)dx$. Usually, we give $f_X(x)$ the value 0 at any of these exceptional points.

10.16. Unlike the cdf of a discrete random variable, the cdf of a continuous random variable has no jumps and is continuous everywhere.

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

In other words, the **area under the graph** of $f_X(x)$ between the points a and b gives the probability $P[a < X \leq b]$.

10.18. $p_X(x) = P[X = x] = P[x \leq X \leq x] = \int_x^x f_X(t)dt = 0$.

Again, it makes no sense to speak of the probability that X will take on a pre-specified value. This probability is always zero.

10.19. $P[X = a] = P[X = b] = 0$. Hence,

$$P[X \in [a, b]] = P[X \in [a, b)] = P[X \in (a, b]] = P[X \in (a, b)]$$

- The corresponding integrals over an interval are not affected by whether or not the endpoints are included or excluded.
- When we work with continuous random variables, it is usually not necessary to be precise about specifying whether or not a range of numbers includes the endpoints. This is quite different from the situation we encounter with discrete random variables where it is critical to carefully examine the type of inequality.

10.20. f_X is nonnegative a.e. [9, p. 138] and $\int_{\mathbb{R}} f(x)dx = 1$.

Example 10.21. Random variable X has pdf

$$f_X(x) = \begin{cases} ce^{-2x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

Find the constant c and sketch the pdf.

Theorem 10.22. Any nonnegative[22] function that integrates to one is a ***probability density function*** (pdf) [9, p. 139].

10.23. Intuition/Interpretation:

The use of the word "density" originated with the analogy to the distribution of matter in space. In physics, any finite volume, no matter how small, has a positive mass, but there is no mass at a single point. A similar description applies to continuous random variables.

Approximately, for a small Δx,

$$P\left[X \in [x, x+\Delta x]\right] = \int_x^{x+\Delta x} f_X(t)dt \approx f_X(x)\Delta x.$$

This is why we call f_X the density function.

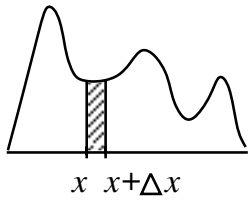

Figure 8: $P\left[x \leq X \leq x+\Delta x\right]$ is the area of the shaded vertical strip.

In other words, the probability of random variable X taking on a value in a *small* interval around point c is approximately equal to $f(c)\Delta c$ when Δc is the length of the interval.

- In fact, $f_X(x) = \lim_{\Delta x \to 0} \frac{P[x<X\leq x+\Delta x]}{\Delta x}$
- The number $f_X(x)$ itself is ***not a probability***. In particular, it does not have to be between 0 and 1.
- $f_X(c)$ is a relative measure for the likelihood that random variable X will take on a value in the immediate neighborhood of point c.

  Stated differently, the pdf $f_X(x)$ expresses how densely the probability mass of random variable X is smeared out in the neighborhood of point x. Hence, the name of density function.

[22] or nonnegative a.e.

10.24. Histogram and pdf [22, p 143 and 145]:

(a) A (probability) **histogram** is a bar chart that divides the range of values covered by the samples/measurements into intervals of the same width, and shows the proportion (relative frequency) of the samples in each interval.

- To make a histogram, you break up the range of values covered by the samples into a number of disjoint adjacent intervals each having the same width, say width Δ. The height of the bar on each interval $[j\Delta, (j+1)\Delta)$ is taken such that the area of the bar is equal to the proportion of the measurements falling in that interval (the proportion of measurements within the interval is divided by the width of the interval to obtain the height of the bar).
- The total area under the histogram is thus standardized/normalized to one.

(b) If you take sufficiently many independent samples from a continuous random variable and make the width Δ of the base intervals of the probability histogram smaller and smaller, the graph of the histogram will begin to look more and more like the pdf.

(c) Conclusion: A probability density function can be seen as a "smoothed out" version of a probability histogram

Example 10.25. Another popular `MATLAB` command: `randn`

10.26. In many situations when you are asked to find pdf, it may be easier to find cdf first and then differentiate it to get pdf.

Example 10.27. A point is "picked at random" in the inside of a circular disk with radius r. Let the random variable X denote the distance from the center of the disk to this point. Find $f_X(x)$.

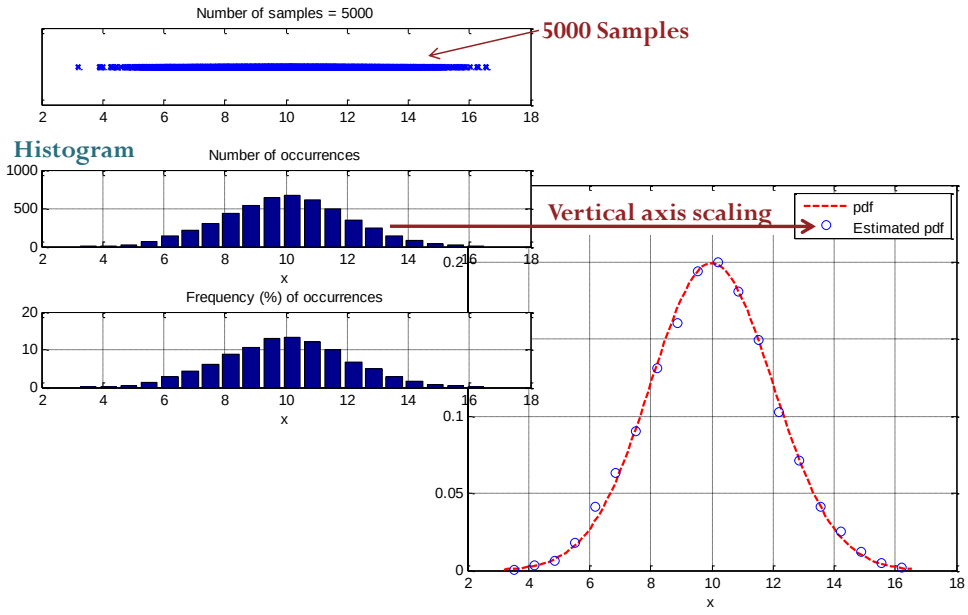

Figure 9: From histogram to pdf.

$$F_X(x) = P\left[X \leq x\right] = \begin{cases} \frac{\pi x^2}{\pi r^2}, & 0 \leq x \leq r \\ 0, & x < 0 \\ 1, & x > r \end{cases}$$

$$f_X(x) = \begin{cases} \frac{2x}{r^2}, & 0 \leq x < r \\ 0, & \text{otherwise} \end{cases}$$

10.3 Expectation and Variance

10.28. ***Expectation***: Suppose X is a continuous random variable with probability density function $f_X(x)$.

$$\mathbb{E}X = \int_{\mathbb{R}} x f_X(x) dx \tag{24}$$

$$\mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}} g(x) f_X(x) dx \tag{25}$$

In particular,

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f_X(x)dx$$
$$\operatorname{Var} X = \int_{-\infty}^{\infty} (x - \mathbb{E}X)^2 f_X(x)dx.$$

Example 10.29. Consider the random variable X from Example 10.27. The expected value of the distance X equals

$$\mathbb{E}X = \int_0^r x\frac{2x}{r^2} = \frac{2}{3}\frac{x^3}{r^2}\bigg|_0^r = \frac{2}{3}r.$$

10.30. If we compare other characteristics of discrete and continuous random variables, we find that with discrete random variables, many facts are expressed as sums. With continuous random variables, the corresponding facts are expressed as integrals.

10.31. Intuition/interpretation: As $n \to \infty$, the average of n independent samples of X will approach $\mathbb{E}X$.

- This observation is known as the "Law of Large Numbers".

10.32. All of the properties for the expectation and variance of discrete random variables also work for continuous random variables as well:

(a) For $c \in \mathbb{R}$, $\mathbb{E}[c] = c$

(b) For $c \in \mathbb{R}$, $\mathbb{E}[X + c] = \mathbb{E}X + c$ and $\mathbb{E}[cX] = c\mathbb{E}X$

(c) For constants a, b, we have $\mathbb{E}[aX + b] = a\mathbb{E}X + b$.

(d) $\mathbb{E}[\cdot]$ is a **linear** operator: $\mathbb{E}[aX + bY] = a\mathbb{E}X + b\mathbb{E}Y$.

  (i) Homogeneous: $\mathbb{E}[cX] = c\mathbb{E}X$

  (ii) Additive: $\mathbb{E}[X + Y] = \mathbb{E}X + \mathbb{E}Y$

  (iii) Extension: $\mathbb{E}[\sum_{i=1}^n c_i X_i] = \sum_{i=1}^n c_i \mathbb{E}X_i$.

(e) $\operatorname{Var} X = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2$

(f) $\operatorname{Var} X \geq 0$.

(g) $\operatorname{Var} X \leq \mathbb{E}\left[X^2\right]$.

(h) $\operatorname{Var}[aX + b] = a^2 \operatorname{Var} X$.

10.33. ***Chebyshev's Inequality***:

$$P\left[|X - \mathbb{E}X| \geq \alpha\right] \leq \frac{\sigma_X^2}{\alpha^2}$$

or equivalently

$$P\left[|X - \mathbb{E}X| \geq n\sigma_X\right] \leq \frac{1}{n^2}$$

- Useful only when $\alpha > \sigma_X$

Example 10.34. A circuit is designed to handle a current of 20 mA plus or minus a deviation of less than 5 mA. If the applied current has mean 20 mA and variance 4 mA^2, use the Chebyshev inequality to bound the probability that the applied current violates the design parameters.

Let X denote the applied current. Then X is within the design parameters if and only if $|X - 20| < 5$. To bound the probability that this does not happen, write

$$P\left[|X - 20| < 5\right] \leq \frac{\operatorname{Var} X}{5^2} = \frac{4}{25} = 0.16.$$

Hence, the probability of violating the design parameters is at most 16%.

10.35. Interesting applications of expectation:

(a) $f_X(x) = \mathbb{E}\left[\delta(X - x)\right]$

(b) $P\left[X \in B\right] = \mathbb{E}\left[1_B(X)\right]$

10.4 Families of Continuous Random Variables

Theorem 10.22 states that any nonnegative function $f(x)$ whose integral over the interval $(-\infty, +\infty)$ equals 1 can be regarded as a probability density function of a random variable. In real-world applications, however, special mathematical forms naturally show up. In this section, we introduce several families of continuous random variables that frequently appear in practical applications. The probability densities of the members of each family all have the same mathematical form but differ only in one or more parameters.

10.4.1 Uniform Distribution

Definition 10.36. For a uniform random variable on an interval $[a, b]$, we denote its family by uniform$([a, b])$ or $\mathcal{U}([a, b])$. Expressions that are synonymous with X is a uniform random variable are X is uniformly distributed and X has a uniform distribution.

This family is characterized by

(a) $f(x) = \frac{1}{b-a} U(x-a)\, U(b-x) = \begin{cases} 0 & x < a, x > b \\ \frac{1}{b-a} & a \le x \le b \end{cases}$

- The random variable X is just as likely to be near any value in $[a, b]$ as any other value.

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

10.37. When $X \sim \mathcal{U}(a, b)$, F_X is not differentiable at a nor b.

Example 10.38. In `MATLAB`, use `a+(b-a).*rand`.

10.39. The uniform distribution provides a probability model for selecting a point at random from the interval $[a, b]$.

- Use with caution to model a quantity that is known to vary randomly between a and b but about which little else is known.
  - Represent ignorance about a parameter taking value in $[a, b]$.

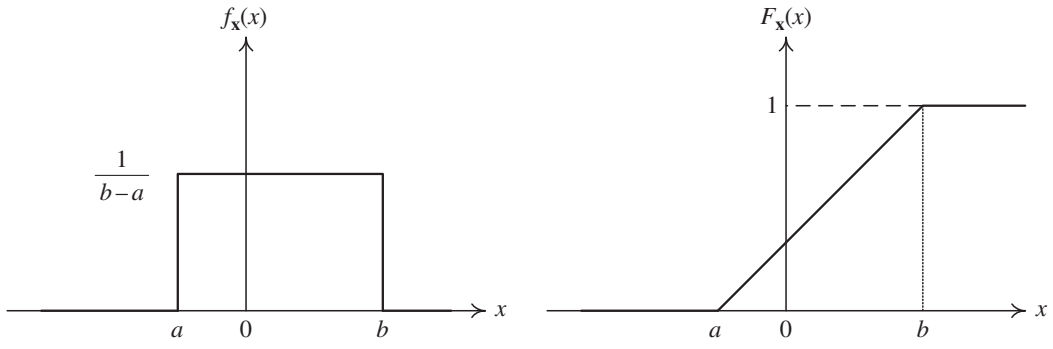

Figure 10: The pdf and cdf for the uniform random variable. [16, Fig. 3.5]

Example 10.40. In coherent radio communications, the phase difference between the transmitter and the receiver, denoted by Θ, is modeled as having a uniform density on $[-\pi, \pi]$.

(a) $P[\Theta \leq 0] = \frac{1}{2}$

(b) $P\left[\Theta \leq \frac{\pi}{2}\right] = \frac{3}{4}$

10.41. $\mathbb{E}X = \frac{a+b}{2}$, $\operatorname{Var} X = \frac{(b-a)^2}{12}$, $\mathbb{E}\left[X^2\right] = \frac{1}{3}\left(b^2 + ab + a^2\right)$.

10.4.2 Gaussian Distribution

Definition 10.42. ***Gaussian*** random variables:

(a) Often called ***normal*** random variables because they occur so frequently in practice

(b) Denoted by $\mathcal{N}\left(m, \sigma^2\right)$.

(c) $\mathcal{N}(0, 1)$ is the ***standard*** Gaussian (normal) distribution.

(d) $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} =$ `normpdf(x,m,sigma)`.

- Figure 13 displays the famous ***bell-shaped*** graph of the Gaussian pdf. This curve is also called the *normal* curve.
- Advanced calculus is required to prove that the area under the graph is indeed 1.

(e) $F_X(x) =$ `normcdf(x,m,sigma)`.

- The standard normal cdf is sometimes denoted by $\Phi(x)$. It inherits all properties of cdf. Moreover, note that $\Phi(-x) = 1 - \Phi(x)$.

- If X is a $\mathcal{N}(m, \sigma^2)$ random variable, the CDF of X is

$$\boxed{F_X(x) = \Phi\left(\frac{x-m}{\sigma}\right).}$$

- It is impossible to express the integral of a Gaussian PDF between non-infinite limits as a function that appears on most scientific calculators.
  - An old but still popular technique to find integrals of the Gaussian PDF is to refer to tables that have been obtained by numerical integration.
    - One such table is the table that lists $\Phi(z)$ for many values of positive z.

(f) An arbitrary Gaussian random variable with mean m and variance σ^2 can be represented as $\sigma S + m$, where $S \sim \mathcal{N}(0, 1)$.

10.43. $\mathbb{E}X = m$ and $\operatorname{Var} X = \sigma^2$.

10.44. $P[|X - \mu| < \sigma] = 0.6827$; $P[|X - \mu| > \sigma] = 0.3173$
$P[|X - \mu| > 2\sigma] = 0.0455$; $P[|X - \mu| < 2\sigma] = 0.9545$

10.45. Unquestionably the Gaussian pdf model is the one most frequently encountered in nature. This is because most random phenomena are due to the action of many different factors. By the *central limit theorem*, the resultant random variable tends to be Gaussian regardless of the underlying probabilities of the individual actors.

10.46. Moments and central moments:

n	0	1	2	3	4
$\mathbb{E}X^n$	1	μ	$\mu^2 + \sigma^2$	$\mu\left(\mu^2 + 3\sigma^2\right)$	$\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$
$\mathbb{E}\left[(X-\mu)^n\right]$	1	0	σ^2	0	$3\sigma^4$

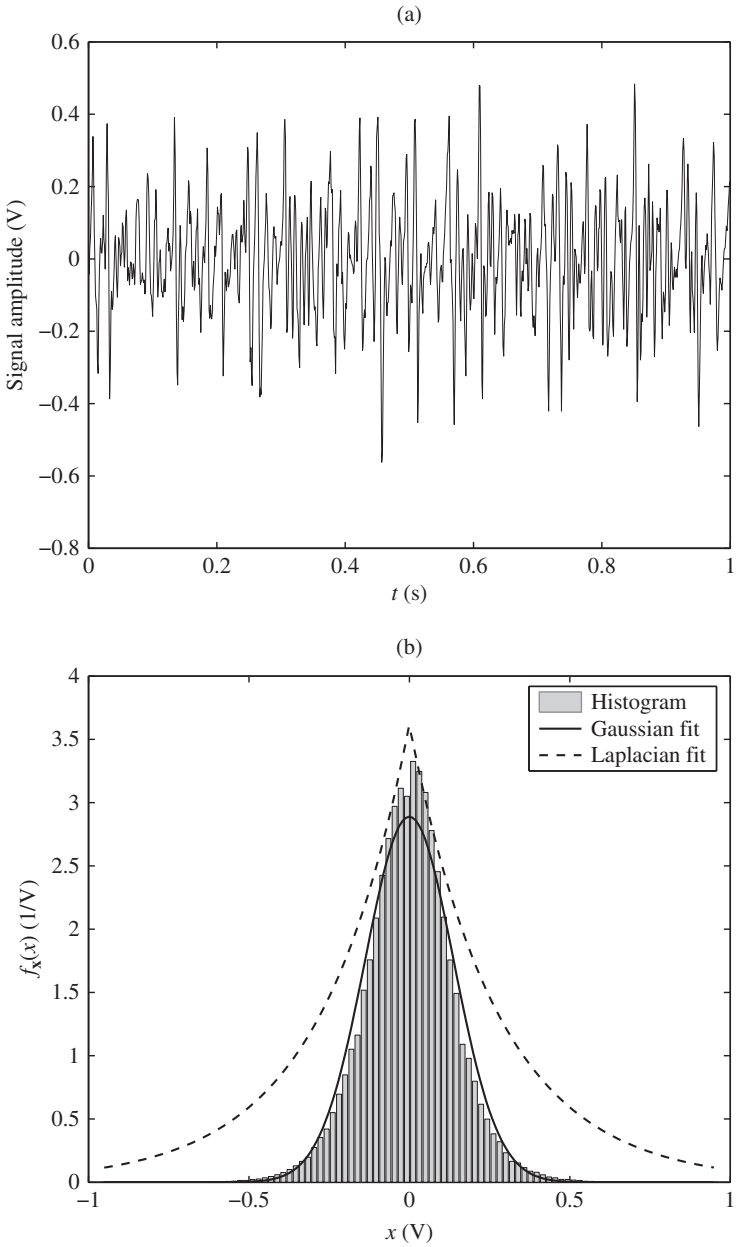

Figure 11: Electrical activity of a skeletal muscle: (a) A sample skeletal muscle (emg) signal, and (b) its histogram and pdf fits. [16, Fig. 3.14]

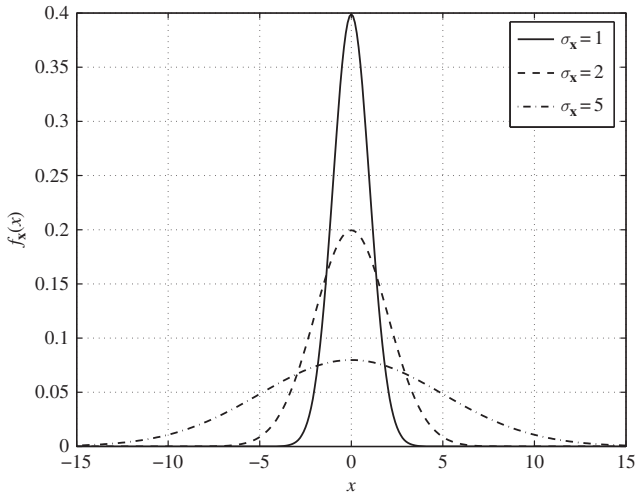

Figure 12: Plots of the zero-mean Gaussian pdf for different values of standard deviation, σ_X. [16, Fig. 3.15]

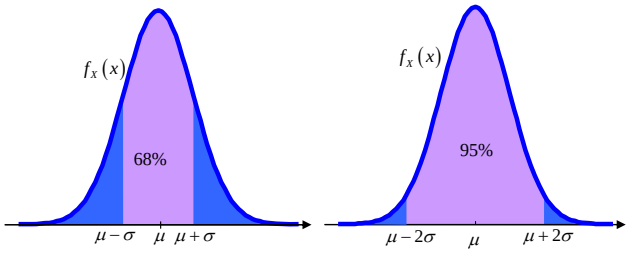

Figure 13: Probability density function of $X \sim \mathcal{N}(\mu, \sigma^2)$.

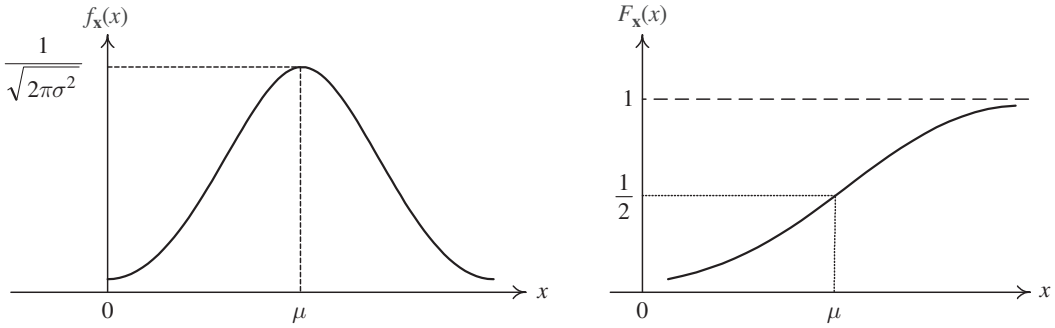

Figure 14: The pdf and cdf of $\mathcal{N}(\mu, \sigma^2)$. [16, Fig. 3.6]

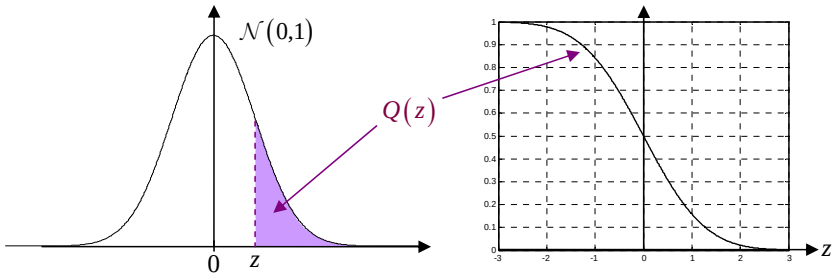

Figure 15: Q-function

10.47. *Q-function*: $Q(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ corresponds to $P[X > z]$ where $X \sim \mathcal{N}(0,1)$; that is $Q(z)$ is the probability of the "tail" of $\mathcal{N}(0,1)$. The Q function is then a complementary cdf (ccdf).

(a) Q is a decreasing function with $Q(0) = \frac{1}{2}$.

(b) $Q(-z) = 1 - Q(z) = \Phi(z)$

(c) $Q^{-1}(1 - Q(z)) = -z$

10.48. *Error function* (MATLAB): $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx = 1 - 2Q\left(\sqrt{2}z\right)$

(a) It is an odd function of z.

(b) For $z \geq 0$, it corresponds to $P[|X| < z]$ where $X \sim \mathcal{N}\left(0, \frac{1}{2}\right)$.

(c) $\lim_{z\to\infty} \operatorname{erf}(z) = 1$

(d) $\operatorname{erf}(-z) = -\operatorname{erf}(z)$

(e) $Q(z) = \frac{1}{2}\operatorname{erfc}\left(\frac{z}{\sqrt{2}}\right) = \frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right)$

(f) $\Phi(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{(2)}}\right)\right) = \frac{1}{2}\operatorname{erfc}\left(-\frac{x}{\sqrt{2}}\right)$

(g) $Q^{-1}(q) = \sqrt{2}\operatorname{erfc}^{-1}(2q)$

(h) The complementary error function:
$\operatorname{erfc}(z) = 1 - \operatorname{erf}(z) = 2Q\left(\sqrt{2}z\right) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-x^2} dx$

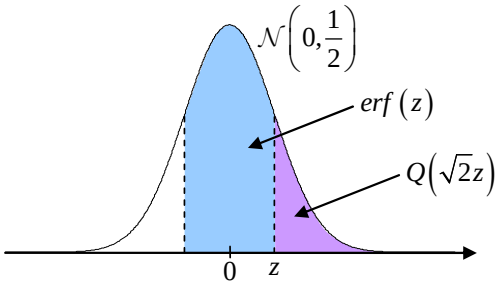

Figure 16: erf-function and Q-function

10.49. $P[X > x] = P[X \geq x] = Q\left(\frac{x-m}{\sigma}\right) = 1 - \Phi\left(\frac{x-m}{\sigma}\right) = \Phi\left(-\frac{x-m}{\sigma}\right)$.
$P[X < x] = P[X \leq x] = 1 - Q\left(\frac{x-m}{\sigma}\right) = Q\left(-\frac{x-m}{\sigma}\right) = \Phi\left(\frac{x-m}{\sigma}\right)$.

10.4.3 Exponential Distribution

Definition 10.50. The exponential distribution is denoted by $\mathcal{E}(\lambda)$.

(a) $\lambda > 0$ is a parameter of the distribution, often called the **rate parameter**.

(b) Characterized by

- $f_X(x) = \lambda e^{-\lambda x} U(x)$;
- $F_X(x) = \left(1 - e^{-\lambda x}\right) U(x)$;

(c) `MATLAB`:

- $X =$ `exprnd(1/lambda)`
- $f_X(x) =$ `exppdf(x,1/lambda)`
- $F_X(x) =$ `expcdf(x,1/lambda)`

10.51. Coefficient of variation: $\mathrm{CV} = \frac{\sigma_X}{\mathbb{E}X} = 1$

10.52. It is a continuous version of geometric distribution. In fact, $\lfloor X \rfloor \sim \mathcal{G}_0(e^{-\lambda})$ and $\lceil X \rceil \sim \mathcal{G}_1(e^{-\lambda})$

10.53. Can be generated by $X = -\frac{1}{\lambda} \ln U$ where $U \sim \mathcal{U}(0,1)$.

Example 10.54. The exponential distribution is intimately related to the Poisson process. It is often used as a probability model for the (waiting) time until a "rare" event occurs.

- time elapsed until the next earthquake in a certain region
- decay time of a radioactive particle
- time between independent events such as arrivals at a service facility or arrivals of customers in a shop.
- duration of a cell-phone call
- time it takes a computer network to transmit a message from one node to another.

Example 10.55. Phone Company A charges \$0.15 per minute for telephone calls. For any fraction of a minute at the end of a call, they charge for a full minute. Phone Company B also charges \$0.15 per minute. However, Phone Company B calculates its charge based on the exact duration of a call. If T, the duration of a call in minutes, is exponential with parameter $\lambda = 1/3$, what are the expected revenues per call $\mathbb{E}[R_A]$ and $\mathbb{E}[R_B]$ for companies A and B?

Solution: First, note that $\mathbb{E}T = \frac{1}{\lambda} = 3$. Hence,

$$\mathbb{E}[R_B] = \mathbb{E}[0.15 \times T] = 0.15\mathbb{E}T = \$0.45.$$

and

$$\mathbb{E}[R_A] = \mathbb{E}[0.15 \times \lceil T \rceil] = 0.15\mathbb{E}\lceil T \rceil.$$

Now, recall that $\lceil T \rceil \sim \mathcal{G}_1\left(e^{-\lambda}\right)$. Hence, $\mathbb{E}\lceil T \rceil = \frac{1}{1-e^{-\lambda}} \approx 3.53$. Therefore,

$$\mathbb{E}[R_A] = 0.15\mathbb{E}\lceil T \rceil \approx 0.5292.$$

10.56. ***Memoryless property***: The exponential r.v. is the only continuous r.v. on $[0, \infty)$ that satisfies the memoryless property:

$$P[X > s + x \,|X > s] = P[X > x]$$

for all $x > 0$ and all $s > 0$ [18, p. 157–159]. In words, the future is independent of the past. The fact that it hasn't happened yet, tells us nothing about how much longer it will take before it does happen.

- Imagining that the exponentially distributed random variable X represents the lifetime of an item, the residual life of an item has the same exponential distribution as the original lifetime, regardless of how long the item has been already in use.
- In particular, suppose we define the set $B+x$ to be $\{x+b : b \in B\}$. For any $x > 0$ and set $B \subset [0, \infty)$, we have

$$P[X \in B + x | X > x] = P[X \in B]$$

because

$$\frac{P\left[X \in B + x\right]}{P\left[X > x\right]} = \frac{\int_{B+x} \lambda e^{-\lambda t} dt}{e^{-\lambda x}} \overset{\tau = t-x}{=} \frac{\int_{B} \lambda e^{-\lambda(\tau+x)} d\tau}{e^{-\lambda x}}.$$

10.57. If X and Y are independent Gaussian random variables, then $U^2 + V^2$ is exponential.

ECS315 2011/1 Part IV.2 Dr.Prapun

10.5 Function of Continuous Random Variables: SISO

Reconsider the derived random variable $Y = g(X)$.

Recall that we can find $\mathbb{E}Y$ easily by (25):

$$\mathbb{E}Y = \mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}} g(x) f_X(x) dx.$$

However, there are cases when we have to deal directly with $P\left[Y \in B\right]$ or find $f_Y(y)$ directly.

Remark: For discrete random variables, it is easy to find $p_Y(y)$ by adding all $p_X(x)$ such that $g(x) = y$:

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

For continuous random variables, it turns out that we can't simply integrate the pdf.

Definition 10.58. Image and Inverse image: Consider a function $g : \mathbb{R} \to \mathbb{R}$.

(a) Given a set A, then the ***image*** of A under g, denoted by $g(A)$, is given by

$$g(A) = \{g(x) : x \in A\}$$

(b) Given a set B, then the ***inverse image*** of B under g, denoted by $g^{-1}(B)$, is given by

$$g^{-1}(B) = \{x : g(x) \in B\}.$$

In words, $g^{-1}(B)$ is the set of all the x whose $g(x)$ are in the set B.

Example 10.59. Suppose B is the singleton set, say $B = \{4\}$. If we consider $g(x) = x^2$, then

$$g^{-1}(B) = g^{-1}(\{4\}) = \left\{x : x^2 = 4\right\} = \{2, -2\}.$$

Observe that if g is a bijective function (one-to-one and onto), then the inverse image of a singleton $\{c\}$ would be another singleton containing the number $g^{-1}(c)$ where g^{-1} is now the usual inverse function that you have studied in calculus.

Example 10.60. Suppose B is the interval, say $B = [4, \infty)$. If we consider $g(x) = x^2$, then

$$\begin{aligned} g^{-1}(B) &= g^{-1}([4, \infty)) = \left\{x : x^2 \in [4, \infty)\right\} \\ &= \left\{x : x^2 \geq 4\right\} = (-\infty, -2] \cup [2, +\infty). \end{aligned}$$

10.61. Suppose $Y = g(X)$, then

$$\boxed{P[Y \in B] = P\left[X \in g^{-1}(B)\right].}$$

This properties is intuitive, easy to understand, and always true regardless of the type of random variables.

Example 10.62. For discrete random variables,

$$P[Y \in B] = P\left[X \in g^{-1}(B)\right] = \sum_{x \in g^{-1}(B)} p_X(x).$$

10.63. For $Y = g(X)$, if you want to find $f_Y(y)$, the following ***two-step procedure*** will always work and is very easy to remember:

(a) Find the cdf $F_Y(y) = P[Y \leq y]$.

(b) Compute the pdf from the cdf by "finding the derivative" $f_Y(y) = \frac{d}{dy} F_Y(y)$.

10.64. ***Linear Transformation***: Suppose $Y = aX + b$. Then, the cdf of Y is given by

$$F_Y(y) = P[Y \leq y] = P[aX + b \leq y] = \begin{cases} P\left[X \leq \frac{y-b}{a}\right], & a > 0, \\ P\left[X \geq \frac{y-b}{a}\right], & a < 0. \end{cases}$$

Now, by definition, we know that

$$P\left[X \leq \frac{y-b}{a}\right] = F_X\left(\frac{y-b}{a}\right),$$

and

$$\begin{aligned} P\left[X \geq \frac{y-b}{a}\right] &= P\left[X > \frac{y-b}{a}\right] + P\left[X = \frac{y-b}{a}\right] \\ &= 1 - F_X\left(\frac{y-b}{a}\right) + P\left[X = \frac{y-b}{a}\right]. \end{aligned}$$

For continuous random variable, $P\left[X = \frac{y-b}{a}\right] = 0$. Hence,

$$\boxed{F_Y(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right), & a > 0, \\ 1 - F_X\left(\frac{y-b}{a}\right), & a < 0. \end{cases}}$$

Finally, fundamental theorem of calculus and chain rule gives

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} \frac{1}{a} f_X\left(\frac{y-b}{a}\right), & a > 0, \\ -\frac{1}{a} f_X\left(\frac{y-b}{a}\right), & a < 0. \end{cases}$$

Note that we can further simplify the final formula by using the $|\cdot|$ function:

$$\boxed{f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right), \quad a \neq 0.}$$

Graphically, to get the plots of f_Y, we compress f_X horizontally by a factor of a, scale it vertically by a factor of $1/|a|$, and shift it to the right by b.

Of course, if $a = 0$, then we get the uninteresting degenerated random variable $Y \equiv b$.

10.65. Do we have to go through that much work to arrive at the above formula?

Example 10.66. Amplitude modulation in certain communication systems can be accomplished using various nonlinear devices such as a semiconductor diode. Suppose we model the nonlinear device by the function $Y = X^2$. If the input X is a continuous random variable, find the density of the output $Y = X^2$.

We will solve this problem using both techniques that we discussed so far:

(a) First technique: the two-step procedure discussed in 10.63.

(b) Second technique: the technique discussed in 10.65.

10.67. Generalization of 10.65: For "nice" function g, first, solve the equation $y = g(x)$. For a point y that has countable number of real-valued roots, denote the real roots by x_k. Then,

$$\boxed{f_Y(y) = \sum_k \frac{f_X(x_k)}{|g'(x_k)|}.} \tag{26}$$

Example 10.68. $Y = X^2$.

10.69. Proof of (26):

Consider Figure 17 where there is unique x such that $g(x) = y$. Then, For small dx and dy, $P[y < Y \leq y + dy] = P[x < X \leq x + dx]$ because the inverse image of the intervals $(y, y+dy]$ corresponds to the interval $(x, x + dx]$ This gives $f_Y(y)|dy| = f_X(x)|dx|$.

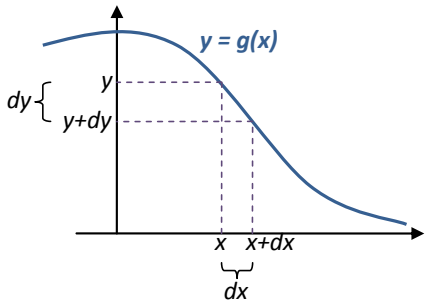

Figure 17: The mapping from $(x, x+dx]$ to $y+dy]$ for bijective relation $y = g(x)$

We can then extend this idea to the the case when there are multiple solutions for $y = g(x)$ as in Figure 18. By similar reasoning, the inverse image of $(y, y + dy]$ contains three disjoint intervals: $(x_1, x_1 + dx_1]$, $(x_2, x_2 + dx_2]$, and $(x_3, x_3 + dx_3]$. Hence, $P[y < Y \leq y + dy]$ can be well approximated by

$$P[x_1 < X \leq x_1 + dx_1] + P[x_2 < X \leq x_2 + dx_2] + P[x_3 < X \leq x_3 + dx_3].$$

Therefore,

$$f_Y(y)|dy| = f_X(x_1)|dx_1| + f_X(x_2)|dx_2| + f_X(x_3)|dx_3| = \sum_k f_X(x_k)|dx_k|$$

which implies

$$f_Y(y) = \sum_k f_X(x_k) \left| \frac{dx_k}{dy} \right| .$$

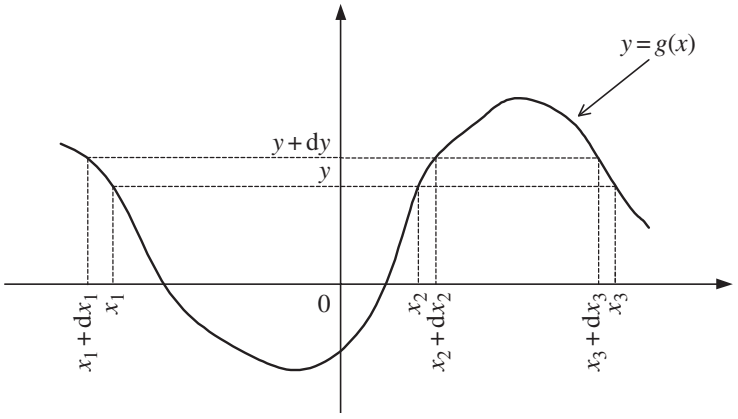

Figure 18: The mapping to $y + dy]$ when there are multiple solutions for the relation $y = g(x)$. [16, Fig. 3.29]

The next step is to find $\frac{dx_k}{dy}$. By recalling the property of derivative, we know that $\Delta y \approx g'(x_k)\Delta x_k$. Hence, $\frac{dx_k}{dy} = \frac{1}{g'(x_k)}$ and we get (26).

10.70. Special Case: When g bijective, we can apply the by inversion mapping theorem which says $\frac{1}{|g'(x)|} = \left| \frac{d}{dy} g^{-1}(y) \right|$ where $y = g(x)$. Consequently,

$$\boxed{f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X\left(g^{-1}(y)\right).}$$

	Discrete	Continuous
$P[X \in B] =$	$\sum_{x \in B} p_X(x)$	$\int_B f_X(x)dx$
$P[X = x] =$	$p_X(x) = F(x) - F(x^-)$	0
Interval prob.	$P^X((a,b]) = F(b) - F(a)$ $P^X([a,b]) = F(b) - F(a^-)$ $P^X([a,b)) = F(b^-) - F(a^-)$ $P^X((a,b)) = F(b^-) - F(a)$	$P^X((a,b]) = P^X([a,b])$ $= P^X([a,b)) = P^X((a,b))$ $= \int_a^b f_X(x)dx = F(b) - F(a)$
$\mathbb{E}X =$	$\sum_x x p_X(x)$	$\int_{-\infty}^{+\infty} x f_X(x)dx$
For $Y = g(X)$,	$p_Y(y) = \sum_{x:\, g(x)=y} p_X(x)$	$f_Y(y) = \frac{d}{dy} P[g(X) \le y]$. Alternatively, $f_Y(y) = \sum_k \frac{f_X(x_k)}{\lvert g'(x_k)\rvert}$, x_k are the real-valued roots of the equation $y = g(x)$.
For $Y = g(X)$, $P[Y \in B] =$	$\sum_{x:g(x)\in B} p_X(x)$	$\int_{\{x:g(x)\in B\}} f_X(x)dx$
$\mathbb{E}[g(X)] =$	$\sum_x g(x) p_X(x)$	$\int_{-\infty}^{+\infty} g(x) f_X(x)dx$
$\mathbb{E}[X^2] =$	$\sum_x x^2 p_X(x)$	$\int_{-\infty}^{+\infty} x^2 f_X(x)dx$
$\operatorname{Var} X =$	$\sum_x (x - \mathbb{E}X)^2 p_X(x)$	$\int_{-\infty}^{+\infty} (x - \mathbb{E}X)^2 f_X(x)dx$

Table 5: Important Formulas for Discrete and Continuous Random Variables

10.6 Pairs of Continuous Random Variables

In this section, we start to look at more than one continuous random variables. You should find that many of the concepts and formulas are similar if not the same as the ones for pairs of *discrete* random variables which we have already studied. For discrete random variables, we use summations. Here, for continuous random variables, we use integrations.

Let's consider one example where a pair of continuous random variables comes up. This example can be solved with a geometric probability approach (ratio of areas). The solution can also be easily checked with a MATLAB simulation.

Example 10.71. Will Lil and Bill Meet at the Malt Shop? [15, p 38–39]

Lil and Bill agree to meet at the malt shop sometime between 3:30 PM and 4 PM later that afternoon. They're pretty casual about details, however, because each knows that the other, while he or she will show up during that half-hour, is as likely to do so at any time during that half-hour as at any other time.

(a) If Lil arrives first, she'll wait five minutes for Bill, and then leave if he hasn't appeared by then.

(b) If Bill arrives first, however, he'll wait seven minutes for Lil before leaving if she hasn't appeared by then.

(c) Neither will wait past 4 PM.

One may ask:

- What's the probability that Lil and Bill meet?
- What's the probability of their meeting if Bill reduces his waiting time to match Lil's (i.e., if both waiting times are five minutes)?
- What's the probability of their meeting if Lil increases her waiting time to match Bill's (i.e., if both waiting times are seven minutes)?

Solution: Let L and B denote the arrival times of Lil and Bill. Suppose $B = 3:45\text{PM}$ and $L = 3:50\text{PM}$. Will they meet?

In Example 10.71, we see an example of a pair of random variables *uniformly distributed* over a rectangle in the Cartesian plane. We know that the model will need to involve some "uniform" probability model because it is given that each person is as likely to show up at any time during that half-hour as at any other time. In other words, the pdf for both L and B are uniform over the interval from 3:30PM to 4PM. Suppose the example is changed so that Bill tends to come at early time. That is, he is more likely to arrive at time that is closer to 3:30PM than 4PM. Then, his pdf will not be uniform anymore. Furthermore, hidden in Example 10.71 is the condition that L and B are independent. This means, for example, if Lil arrives first, she will not make a call to Bill so that he would run faster to the shop.

Recall that for a pair of discrete random variables, the joint pmf $p_{X,Y}(x, y)$ completely characterizes the probability model of two random variables X and Y. In particular, it does not only capture the probability of X and probability of Y individually, it also capture the relationship between them. For continuous random variable, we replace the joint pmf by joint pdf.

Definition 10.72. We say that two random variables X and Y are ***jointly continuous*** with ***joint pdf*** $f_{X,Y}(x, y)$ if for any **region**

R on the (x, y) plane

$$P\left[(X,Y) \in R\right] = \iint\limits_{\{(x,y):(x,y)\in R\}} f_{X,Y}(x,y)dxdy \tag{27}$$

To understand where Definition 10.72 comes from, it is helpful to take a careful look at Table 6.

	Discrete	Continuous
$P\left[X \in B\right]$	$\sum\limits_{x\in B} p_X(x)$	$\int\limits_B f_X(x)dx$
$P\left[(X,Y) \in R\right]$	$\sum\limits_{(x,y):(x,y)\in R} p_{X,Y}(x,y)$	$\iint\limits_{\{(x,y):(x,y)\in R\}} f_{X,Y}(x,y)dxdy$

Table 6: pmf vs. pdf

Remark: If you want to check that a function $f(x, y)$ is the joint pdf of a pair of random variables (X, Y) by using the above definition, you will need to check that (27) is true for any region R. This is not an easy task. Hence, we do not usually use this definition for such kind of test. There are some mathematical facts that can be derived from this definition. Such facts produce easier condition(s) than (27) but we will not talk about them here.

For us, Definition 10.72 is *useful* because if you know that a function $f(x, y)$ is a joint pdf of a pair of random variables, then you can *calculate* countless possibilities of probabilities involving these two random variables via (27). (See, e.g. Example 10.75.) However, the actual calculation of probability from (27) can be difficult if you have non-rectangular region R or if you have a complicated joint pdf. The formula itself is straight-forward and simple, but to carry it out may require that you review some multi-variable integration technique from your calculus class.

Note also that the joint pdf's definition extends the interpretation/approximation that we previously discussed for one random variable. Recall that for a single random variable, the pdf is a measure of ***probability per unit length***. In particular, if you want to find the probability that the value of a random variable X falls inside some small interval, say the interval $[1.5, 1.6]$, then

this probability can be approximated by

$$P\left[1.5 \leq X \leq 1.6\right] \approx f_X(1.5) \times 0.1.$$

More generally, for small value of interval length d, the probability that the value of X falls within a small interval $[x, x+d]$ can be approximated by

$$P\left[x \leq X \leq x+d\right] \approx f_X(x) \times d. \tag{28}$$

Usually, instead of using d, we use Δx and hence

$$P\left[x \leq X \leq x+\Delta x\right] \approx f_X(x) \times \Delta x. \tag{29}$$

10.73. Intuition/Approximation: For two random variables X and Y, the joint pdf $f_{X,Y}(x,y)$ measures ***probability per unit area***:

$$P\left[x \leq X \leq x+\Delta x\,,\; y \leq Y \leq y+\Delta y\right] \approx f_{X,Y}(x,y) \times \Delta x \times \Delta y. \tag{30}$$

Do not forget that the comma signifies the "and" (intersection) operation.

10.74. There are two important characterizing properties of joint pdf:

(a) $f_{X,Y} \geq 0$ a.e.

(b) $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$

Example 10.75. Consider a probability model of a pair of random variables uniformly distributed over a rectangle in the X-Y plane:

$$f_{X,Y}(x,y) = \begin{cases} c, & 0 \le x \le 5, 0 \le y \le 3 \\ 0, & \text{otherwise.} \end{cases}$$

Find the constant c, $P[2 \le X \le 3, 1 \le Y \le 3]$, and $P[Y > X]$

10.76. Other important properties and definitions for a pair of continuous random variables are summarized in Table 7 along with their "discrete counterparts".

Definition 10.77. The ***joint cumulative distribution function*** (***joint cdf***) of random variables X and Y (of any type(s)) is defined as

$$F_{X,Y}(x,y) = P[X \le x, Y \le y].$$

- Although its definition is simple, we rarely use the joint cdf to study probability models. It is easier to work with a probability mass function when the random variables are discrete, or a probability density function if they are continuous.

10.78. The joint cdf for a pair of random variables (of any type(s)) has the following properties[23].:

[23]Note that when we write $F_{X,Y}(x,\infty)$, we mean $\lim_{y\to\infty} F_{X,Y}(x,y)$. Similar limiting definition applies to $F_{X,Y}(\infty,\infty)$, $F_{X,Y}(-\infty,y)$, $F_{X,Y}(x,-\infty)$, and $F_{X,Y}(\infty,y)$

	Discrete	Continuous
$P[(X,Y) \in R]$	$\sum_{(x,y):(x,y)\in R} p_{X,Y}(x,y)$	$\iint_{\{(x,y):(x,y)\in R\}} f_{X,Y}(x,y)dxdy$
Joint to Marginal:	$p_X(x) = \sum_y p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy$
(Law of Total Prob.)	$p_Y(y) = \sum_x p_{X,Y}(x,y)$	$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx$
$P[X > Y]$	$\sum_x \sum_{y:\, y<x} p_{X,Y}(x,y)$ $= \sum_y \sum_{x:\, x>y} p_{X,Y}(x,y)$	$\int_{-\infty}^{+\infty} \int_{-\infty}^{x} f_{X,Y}(x,y)dydx$ $= \int_{-\infty}^{+\infty} \int_{y}^{\infty} f_{X,Y}(x,y)dxdy$
$P[X = Y]$	$\sum_x p_{X,Y}(x,x)$	0
$X \perp\!\!\!\perp Y$	$p_{X,Y}(x,y) = p_X(x)p_Y(y)$	$f_{X,Y}(x,y) = f_X(x)f_Y(y)$
Conditional	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

Table 7: Important formulas for a pair of discrete RVs and a pair of Continuous RVs

(a) $0 \le F_{X,Y}(x,y) \le 1$

  (i) $F_{X,Y}(\infty,\infty) = 1$.

  (ii) $F_{X,Y}(-\infty,y) = F_{X,Y}(x,-\infty) = 0$.

(b) Joint to Marginal: $F_X(x) = F_{X,Y}(x,\infty)$ and $F_Y(y) = F_{X,Y}(\infty,y)$. In words, we obtain the marginal cdf F_X and F_Y directly from $F_{X,Y}$ by setting the unwanted variable to ∞.

(c) If $x_1 \le x_2$ and $y_1 \le y_2$, then $F_{X,Y}(x_1,y_1) \le F_{X,Y}(x_2,y_2)$

10.79. The joint cdf for a pair of *continuous* random variables also has the following properties:

(a) $F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dvdu$.

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

10.80. ***Independence***:

The following statements are equivalent:

(a) Random variables X and Y are ***independent***.

(b) $[X \in B] \perp\!\!\!\perp [Y \in C]$ for all B, C.

(c) $P[X \in B, Y \in C] = P[X \in B] \times P[Y \in C]$ for all B, C.

(d) $f_{X,Y}(x,y) = f_X(x) \times f_Y(y)$ for all x, y.

(e) $F_{X,Y}(x,y) = F_X(x) \times F_Y(y)$ for all x, y.

10.7 Function of a Pair of Continuous Random Variables: MISO

There are many situations in which we observe two random variables and use their values to compute a new random variable.

Example 10.81. Signal in ***additive noise***: When we says that a random signal X is transmitted over a channel subject to additive noise N, we mean that at the receiver, the received signal Y will be $X+N$. Usually, the noise is assumed to be zero-mean Gaussian noise; that is $N \sim \mathcal{N}(0, \sigma_N^2)$ for some noise power σ_N^2.

Example 10.82. In a ***wireless channel***, the transmitted signal X is corrupted by fading (multiplicative noise). More specifically, the received signal Y at the receiver's antenna is $Y = H \times X$.

Remark: In the actual situation, the signal is further corrupted by additive noise N and hence $Y = HX + N$. However, this expression for Y involves more than two random variables and hence we we will not consider it here.

10.83. Consider a new random variable Z defined by

$$Z = g(X, Y).$$

Table 8 summarizes the basic formulas involving this derived random variable.

10.84. When X and Y are continuous random variables, it may be of interest to find the pdf of the derived random variable $Z = g(X, Y)$. It is usually helpful to devide this task into two steps:

(a) Find the cdf $F_Z(z) = P[Z \leq z] = \iint_{g(x,y) \leq z} f_{X,Y}(x,y) dx dy$

(b) $f_W(w) = \frac{d}{dw} F_W(w)$.

	Discrete	Continuous
$\mathbb{E}[Z]$	$\sum_x \sum_y g(x,y) p_{X,Y}(x,y)$	$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y) dxdy$
$P[Z \in B]$	$\sum_{(x,y):\, g(x,y)\in B} p_{X,Y}(x,y)$	$\iint_{\{(x,y):\, g(x,y)\in B\}} f_{X,Y}(x,y) dxdy$
$Z = X + Y$	$p_Z(z) = \sum_x p_{X,Y}(x, z-x)$ $= \sum_y p_{X,Y}(z-y, y)$	$f_Z(z) = \int_{-\infty}^{+\infty} f_{X,Y}(x, z-x) dx$ $= \int_{-\infty}^{+\infty} f_{X,Y}(z-y, y) dy$
$X \perp\!\!\!\perp Y$	$p_{X+Y} = p_X * p_Y$	$f_{X+Y} = f_X * f_Y$

Table 8: Important formulas for function of a pair of RVs. Unless stated otherwise, the function is defined as $Z = g(X, Y)$

Example 10.85. Suppose X and Y are i.i.d. $\mathcal{E}(3)$. Find the pdf of $W = Y/X$.

10.86. Observe that finding the pdf of $Z = g(X, Y)$ is a time-consuming task. If you goal is to find $\mathbb{E}[Z]$ do not forget that it

can be calculated directly from

$$\mathbb{E}\left[g(X,Y)\right] = \int\int g(x,y) f_{X,Y}(x,y) dxdy.$$

	Discrete	Continuous
$P\left[X \in B\right]$	$\sum\limits_{x \in B} p_X(x)$	$\int\limits_B f_X(x)dx$
$P\left[(X,Y) \in R\right]$	$\sum\limits_{(x,y):(x,y)\in R} p_{X,Y}(x,y)$	$\iint\limits_{\{(x,y):(x,y)\in R\}} f_{X,Y}(x,y)dxdy$
Joint to Marginal:	$p_X(x) = \sum\limits_y p_{X,Y}(x,y)$	$f_X(x) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)dy$
(Law of Total Prob.)	$p_Y(y) = \sum\limits_x p_{X,Y}(x,y)$	$f_Y(y) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)dx$
$P\left[X > Y\right]$	$\sum\limits_x \sum\limits_{y:\, y<x} p_{X,Y}(x,y)$ $= \sum\limits_y \sum\limits_{x:\, x>y} p_{X,Y}(x,y)$	$\int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{x} f_{X,Y}(x,y)dydx$ $= \int\limits_{-\infty}^{+\infty} \int\limits_{y}^{\infty} f_{X,Y}(x,y)dxdy$
$P\left[X = Y\right]$	$\sum\limits_x p_{X,Y}(x,x)$	0
$X \perp\!\!\!\perp Y$	$p_{X,Y}(x,y) = p_X(x)p_Y(y)$	$f_{X,Y}(x,y) = f_X(x)f_Y(y)$
Conditional	$p_{X\mid Y}(x\mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\mid Y}(x\mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
$\mathbb{E}\left[g(X,Y)\right]$	$\sum\limits_x \sum\limits_y g(x,y) p_{X,Y}(x,y)$	$\int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y)dxdy$
$P\left[g(X,Y) \in B\right]$	$\sum\limits_{(x,y):\, g(x,y)\in B} p_{X,Y}(x,y)$	$\iint\limits_{\{(x,y):\, g(x,y)\in B\}} f_{X,Y}(x,y)dxdy$
$Z = X + Y$	$p_Z(z) = \sum\limits_x p_{X,Y}(x, z-x)$ $= \sum\limits_y p_{X,Y}(z-y, y)$	$f_Z(z) = \int_{-\infty}^{+\infty} f_{X,Y}(x, z-x)dx$ $= \int_{-\infty}^{+\infty} f_{X,Y}(z-y, y)$

Table 9: pmf vs. pdf

10.87. The following property is valid for any kind of random variables:

$$\mathbb{E}\left[\sum_i Z_i\right] = \sum_i \mathbb{E}\left[Z_i\right].$$

Furthermore,

$$\mathbb{E}\left[\sum_i g_i(X,Y)\right] = \sum_i \mathbb{E}\left[g_i(X,Y)\right].$$

10.88. ***Independence***: At this point, it is useful to summarize what we know about independence. The following statements are equivalent:

(a) Random variables X and Y are ***independent***.

(b) $[X \in B] \perp\!\!\!\perp [Y \in C]$ for all B, C.

(c) $P[X \in B, Y \in C] = P[X \in B] \times P[Y \in C]$ for all B, C.

(d) For discrete RVs, $p_{X,Y}(x,y) = p_X(x) \times p_Y(y)$ for all x, y.
For continuous RVs, $f_{X,Y}(x,y) = f_X(x) \times f_Y(y)$ for all x, y.

(e) $F_{X,Y}(x,y) = F_X(x) \times F_Y(y)$ for all x, y.

(f) $\mathbb{E}[h(X)g(Y)] = \mathbb{E}[h(X)]\,\mathbb{E}[g(Y)]$ for all functions h and g.

Definition 10.89. All of the definitions involving expectation of a function of two random variables are the same as in the discrete case:

- **Correlation** between X and Y: $\mathbb{E}[XY]$.
- **Covariance** between X and Y:
$$\operatorname{Cov}[X,Y] = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y.$$
- $\operatorname{Var} X = \operatorname{Cov}[X,X]$.
- X and Y are said to be ***uncorrelated*** if and only if $\operatorname{Cov}[X,Y] = 0$.
- X and Y are said to be ***orthogonal*** if $\mathbb{E}[XY] = 0$.
- ***Correlation coefficient***: $\rho_{XY} = \frac{\operatorname{Cov}[X,Y]}{\sigma_X \sigma_Y}$

Example 10.90. The ***bivariate Gaussian*** or ***bivariate normal density*** is a generalization of the univariate $\mathcal{N}(m, \sigma^2)$ density. For bivariate normal, $f_{X,Y}(x,y)$ is

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\frac{\left(\frac{x-\mathbb{E}X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mathbb{E}X}{\sigma_X}\right)\left(\frac{y-\mathbb{E}Y}{\sigma_Y}\right) + \left(\frac{y-\mathbb{E}Y}{\sigma_Y}\right)^2}{2(1-\rho^2)} \right\}, \tag{31}$$

where $\rho = \frac{\operatorname{Cov}[X,Y]}{\sigma_X\sigma_Y} \in (-1,1)$.

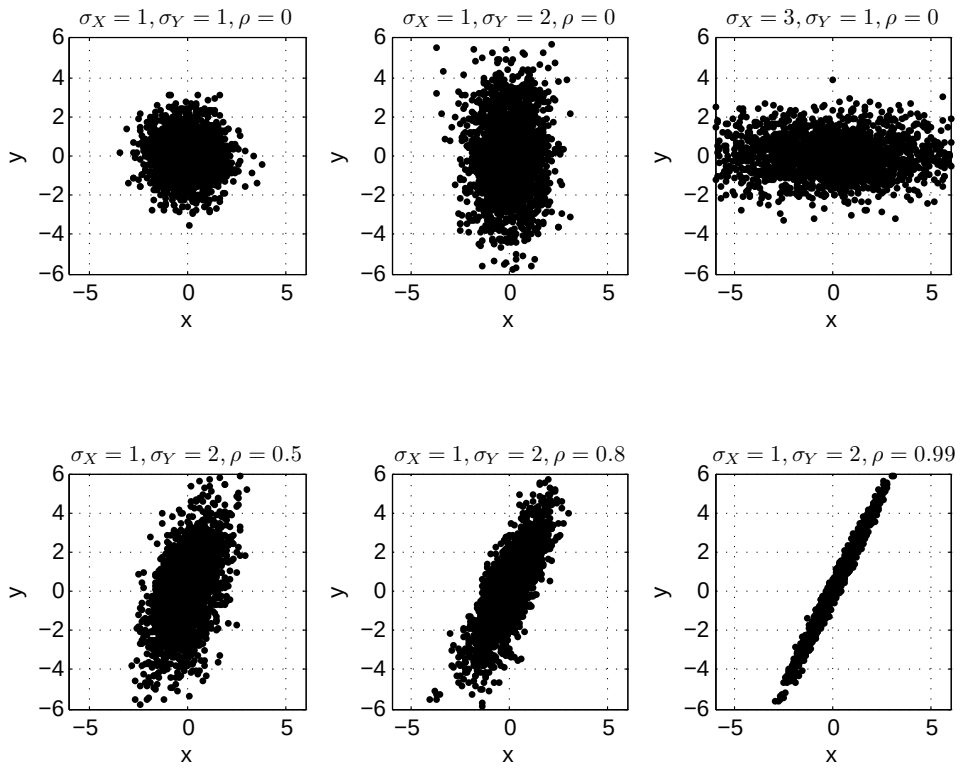

Figure 19: Samples from bivariate Gaussian distributions.

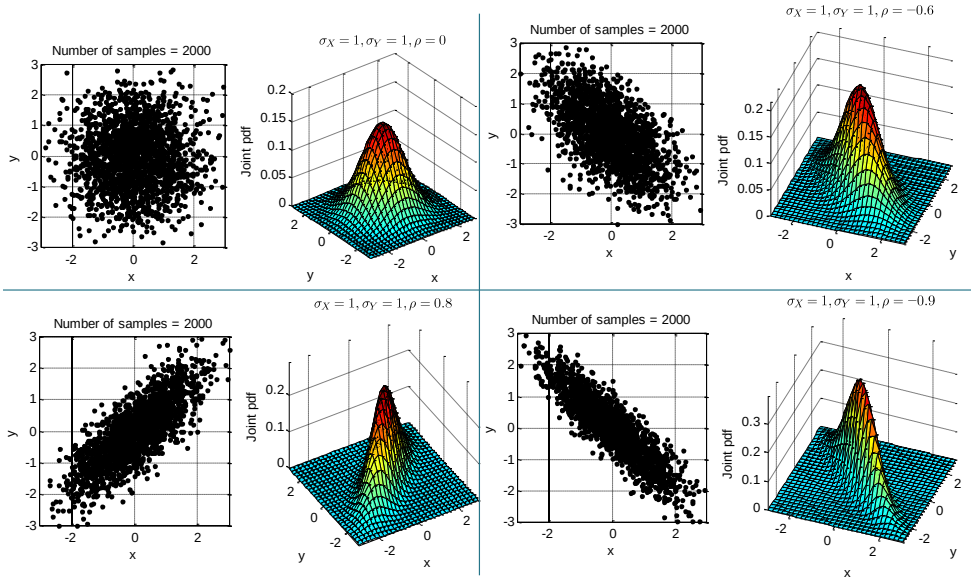

Figure 20: Effect of ρ on bivariate Gaussian distribution. Note that the marginal pdfs for both X and Y are all standard Gaussian.

ECS315 2011/1 Part V Dr.Prapun

11 Mixed Random Variables

Before we even begin to talk about mixed random variables, we will first present the mathematical prerequisite that is needed to enlarge the concept of probability density function (pdf). This will require the use of the delta function.

11.1. Recall that the (Dirac) ***delta function*** or (unit) impulse function is denoted by $\delta(t)$. It is usually depicted as a vertical arrow at the origin. Note that $\delta(t)$ is not a true function; it is undefined at $t = 0$. Make sure that you know the following properties for δ function.

(a) It is zero everywhere except at $t = 0$.

(b) The value is undefined at $t = 0$ but we usually think of $\delta(0)$ as having ∞ value.

(c) $\int_{-\infty}^{+\infty} \delta(t)dt = 1$. In fact, a more general version of this property is

$$\int_A \delta(t)dt = 1_A(0). \tag{32}$$

More specifically,

$$\int_a^b \delta(t)dt = \begin{cases} 1, & a \le 0 \le b, \\ 0, & \text{otherwise.} \end{cases}$$

(d) ***Sifting/sampling property***:

$$\boxed{\int_{-\infty}^{+\infty} g(t)\delta(t-t_0)dt = g(t_0)}. \tag{33}$$

for any function $g(t)$ which is continuous at t_0.

We will now use (32) to derive one important relationship which we shall frequently use throughout this section.

11.2. Consider an integration that resembles one in fundamental theorem of calculus (and also the integration that you perform when finding the cdf from pdf):

$$\int_{-\infty}^{x} \delta(t)dt.$$

We can evaluate the above integration via (32) which gives

$$\boxed{\int_{-\infty}^{x} \delta(t)dt = u(x)} \tag{34}$$

where $u(x)$ is the ***unit step function*** which can be represented as $1_{(-\infty,0]}(x)$. Note that, for this class, the unit step function is required to have $u(0) = 1$ (for right continuity).

Looking at (34), it seems reasonable to say

$$\boxed{\frac{d}{dx}u(x) = \delta(x).} \tag{35}$$

To put it as described in [25, p 127], (35) "embodies a certain kind of consistency in its inconsistency." We know that $\delta(x)$ does not really exist at $x = 0$. Similarly, the derivative of $u(x)$ does not really exist at $x = 0$. However, (35) allows us to use $\delta(x)$ to define a generalized pdf that applies to discrete random variables as well as to continuous random variables.

11.1 PDF for Discrete Random Variables

11.3. Recall that the cdf of a discrete random variable X can be expressed via its pmf as

$$F_X(x) = \sum_{x_i} p_X(x_i)u(x - x_i).$$

If we want to find its pdf, we may try to apply the formula $f_X(x) = F'_X(x)$. With the help of (35), we now have

$$\boxed{f_X(x) = \sum_{x_i} p_X(x_i)\delta(x - x_i)}. \tag{36}$$

So, miraculously, we now have a pdf for discrete random variable!

Although the delta function is not a well-defined function[24], this technique does allow easy manipulation of mixed distribution. The definition of quantities involving discrete random variables and the corresponding properties can then be derived from the pdf and hence there is no need to talk about pmf at all!

Example 11.4. Consider a discrete random variable X whose $p_X(1) = p_X(5) = p_X(7) = \frac{1}{3}$. Plot its pmf, cdf, and pdf.

Observe that $f_X(x)$ for a discrete random variable consists of a series of impulses. The value of $f_X(x)$ is either 0 or ∞. By contrast, the pdf of a continuous random variable has nonzero, finite values over intervals of x.

11.5. It may be helpful to review what we have done at this point.

(a) The unit step function $u(x)$ is always constant except at 0. So, the slope is zero everywhere except at 0.

(b) Suppose we want to define the slope at 0, we know that the value of the unit step function *increase* abruptly from 0 to 1. So, the slope should be *positive* and the amount should be unimaginably large. So, setting it at ∞ as given by the δ function makes sense.

[24] Rigorously, it is a unit measure at 0.

(c) If we look closely at the cdf of discrete random variables, we see that most of the time it stays constant. Of course, the derivative is 0 for those places. Now, there will be a number of places (at most countably many of them) at which the cdf jumps. For a jump, e.g. at c, the amount of jump is given by $p_X(c)$. Now if the increase from 0 to 1 in the unit step function is interpreted as having slope $1 \times \delta(0)$, it is reasonable to assign the slope at c of the cdf to be $\delta(0)$ scaled by a factor of $p_X(c)$. Of course, because the jump happens at c not at 0, we should have the term $p_X(c)\delta(x-c)$ in the slope expression.

11.6. From the pdf in (36), we can also find $\mathbb{E}[g(X)]$ from the pdf-version formula:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx = \int_{-\infty}^{+\infty} g(x) \sum_i p_X(x_i)\delta(x-x_i) dx.$$

By interchanging the integration and the sum, we have

$$\mathbb{E}[g(X)] = \sum_i p_X(x_i) \int_{-\infty}^{+\infty} g(x)\delta(x-x_i) dx.$$

Applying the sifting property (33) gives

$$\mathbb{E}[g(X)] = \sum_i p_X(x_i) g(x_i)$$

which is the same as what we have studied in earlier section(s).

11.2 Three Types of Random Variables

11.7. Review: You may recall[25] the following properties for cdf of discrete random variables. These properties hold for any kind of random variables.

(a) The cdf is defined as $F_X(x) = P[X \leq x]$. This is valid for any type of random variables.

(b) Moreover, the cdf for any kind of random variable must satisfies three properties which we have discussed earlier:

[25] If you don't know these properties by now, you should review them as soon as possible.

CDF1 F_X is non-decreasing

CDF2 F_X is right continuous

CDF3 $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.

(c) $P[X = x] = F_X(x) - F_X(x^-)$ = the jump or saltus in F at x.

Theorem 11.8. If you find a function F that satisfies CDF1, CDF2, and CDF3 above, then F is a cdf of some random variable. See also 18.2.

Example 11.9. Consider an input X to a device whose output Y will be the same as the input if the input level does not exceed 5. For input level that exceeds 5, the output will be saturated at 5. Suppose $X \sim \mathcal{U}(0, 6)$. Find $F_Y(y)$.

11.10. We can categorize random variables into three types according to its cdf:

(a) If $F_X(x)$ is piecewise flat with discontinuous jumps, then X is **discrete**.

(b) If $F_X(x)$ is a continuous function, then X is **continuous**.

(c) If $F_X(x)$ is a piecewise continuous function with discontinuities, then X is **mixed**.

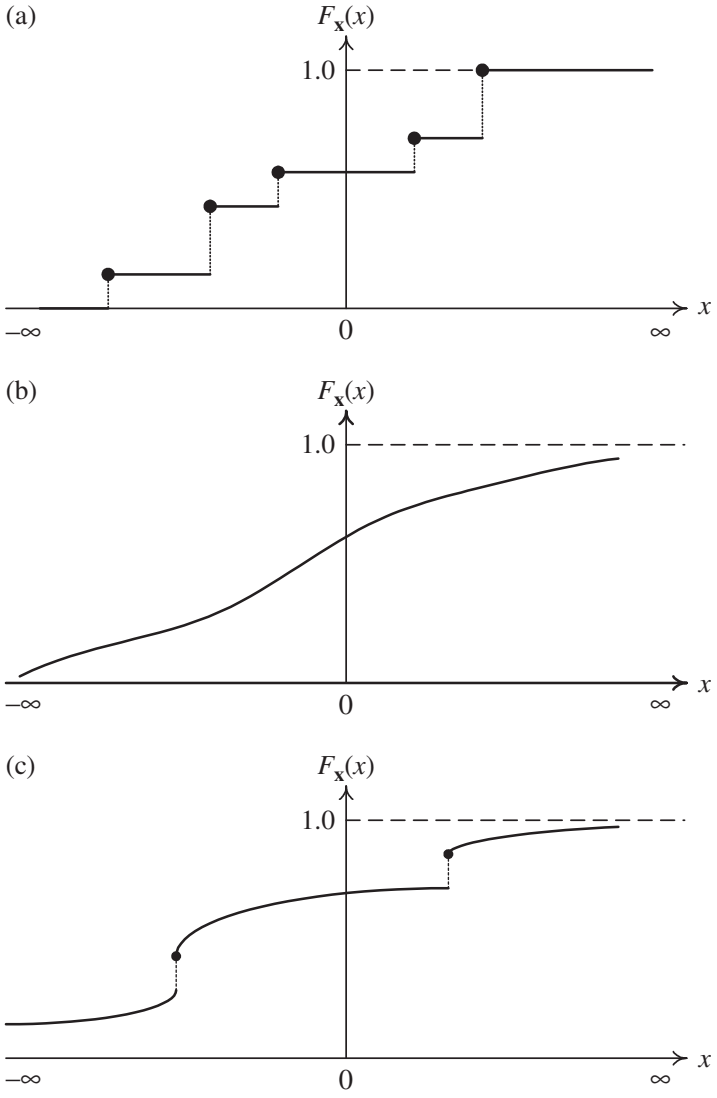

Figure 21: Typical cdfs: (a) a discrete random variable, (b) a continuous random variable, and (c) a mixed random variable [16, Fig. 3.2].

11.11. (**PDF of Mixed Random Variables**) We have seen how to get the pdf of random variable for the first two types. What about the last type? Going back to idea that $f_X(x) = \frac{d}{dx}F_X(x)$,

we will combine properties from the first two types of random variables into the recipe for finding pdf of the mixed random variable.

(i) For $x = c$ at which $F_X(x)$ is continuous, we shall follow what we have been using for continuous random variable; that is we try to actually find the derivative of $F_X(x)$. If it exists, this is the value of $f_X(x)$. If the derivative does not exist (left and right slopes are not the same), then we can set $f_X(x)$ to be any value. (0 is frequently used).

(ii) For $x = c$ at which $F_X(x)$ has a jump, we know that

$$P[X = c] = F_X(c) - F_X(c^-) > 0.$$

We shall follow what we have just used with discrete random variable. That is, for each jump, we add the term $p_X(c)\delta(x-c)$ into the pdf expression.

Example 11.12. In Example 11.9, the pdf is given by

We have seen in Example 11.9 that some function can turn a continuous random variable into a mixed random variable. Next, we will work on an example where a continuous random variable is turned into a discrete random variable.

Example 11.13. Let $X \sim \mathcal{U}(0,1)$ and $Y = g(X)$ where

$$g(x) = \begin{cases} 1, & x < 0.6 \\ 0, & x \geq 0.6. \end{cases}$$

Before going deeply into the math, it is helpful to think about the nature of the derived random variable Y. The definition of $g(x)$ tells us that Y has only two possible values, $Y = 0$ and $Y = 1$. Thus, Y is a discrete random variable.

Example 11.14. In MATLAB, we have the `rand` command to generate $\mathcal{U}(0,1)$. If we want to generate a Bernoulli random variable with parameter p, what can we do?

Example 11.15. Observe someone making a call on a mobile phone and record the duration of the call. In a simple model of the experiment, 1/3 of the calls never begin either because no one answers or the signal is poor. The duration of these calls is 0 minutes. Otherwise, with probability 2/3, the call duration is assumed (for simplicity) to be uniformly distributed between 0 and 3 minutes. Let X denote the call duration. Find the cdf $F_X(x)$ and the pdf $f_X(x)$. [25, Ex 3.21]

For this example, from the way that it is stated, you probably realize that there will be conditional probability involved. Our goal is to find $F_X(x) = P[X \leq x]$. So, let $A = [X \leq x]$ and hence we want to find $P(A)$.

Let B be the event that the call never begins (as stated in the example).

12 Conditional Probability: Conditioning by a Random Variable

We have seen how the concept of conditional probability can be applied to random variables. Here we first review the concept of conditional probability and then introduce the concept of conditional expectation and also consider continuous random variables.

Definition 12.1. Recall that the (event-based) conditional probability $P(A|B)$ is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{37}$$

where $P(B) > 0$.

For random variable X, if you have a new knowledge about X, you can (and should) update its pdf to reflect this new knowledge. In which case, the new pdf will not be the same as your original pdf anymore.

Example 12.2. Suppose you are considering a random variable X with some pmf. The updated information could be that another random variable Y takes the value $Y = 1$. Of course, it will be

useless if X and Y are independent. However, if X and Y are not independent, their dependency will be captured in the joint pmf. Hence, we would like to use the joint pmf to update the marginal pmf of X when we know that the event $Y = 1$ occurs.

Example 12.3. Consider the scores of 20 students below:

$$\underbrace{10, 9, 10, 9, 9, 10, 9, 10, 10, 9}_{\text{Room \#1}}, \underbrace{1, 3, 4, 6, 5, 5, 3, 3, 1, 3.}_{\text{Room \#2}}$$

The first ten scores are from (ten) students in room #1. The last 10 scores are from (ten) students in room #2.

Suppose we have the a score report card for each of the student. Then, in total, we have 20 report cards.

Figure 22: Picking a report cards randomly from a pile of cards

I pick one of the report cards up randomly. Let S be the score on the card. What is the chance that $S > 5$? (Ans: $P[S > 5] = 11/20$.) What is the chance that $S = 10$? (Ans: $p_S(10) = P[S = 10] = 5/20 = 1/4$.)

Now suppose someone informs me that the report card that I picked up is from a student in room #1. (He may be able to tell this by the color of the report card of which I have no knowledge.) If we let the random variable Y denote the room# of the student whose report card is picked up by me, then I now have an extra information that $Y = 1$. Given this new information, if I want to answer the same question of what is the chance that $S > 5$, it is probably not a good idea to ignore the extra information that $Y = 1$. I must update my probability to $P[S > 5|Y = 1]$. Of course, we need no calculation at all for this because all of the students in room #1 have scores that are > 5. So, $P[S > 5|Y = 1] =$

1. We can also calculate $P[S = 10|Y = 1] = 5/10 = 1/2$. The concise notation for this conditional probability $P[S = 10|Y = 1]$ is $p_{S|Y}(10|1)$ which is formally defined in Definition 12.5 below.

Without the information about Y, I would also say that the expected value of S is $\mathbb{E}S = 6.45$. However, given the new information that $Y = 1$, there are only two possible values of the score: either 9 or 10. In fact, they are equally likely. My expected value of S will not be 6.45 anymore; it increases to 9.5! This new expectation value is denoted by $\mathbb{E}[S|Y = 1]$.

Similarly, we can also update the variance and standard deviation of my random variable S. The original standard deviation of S is approximately 3.3. However, when I know that my card comes from room #1, my uncertainty about the value of S reduces significantly. The new standard deviation is only 0.5. This new standard deviation and its corresponding variance are denoted by $\sigma_{S|Y=1}$ and $\mathrm{Var}[S|Y = 1]$, respectively.

Example 12.4. Suppose you have to guess the age X (number of years) of a particular person. If you don't have any further information, you may have a particular pmf/pdf in your mind. However, if someone gives you more information, for example, that this person is studying at SIIT. You then have a new pmf/pdf for the age of this person. Next, if you know that this person is taking ECS315 at SIIT, you can update the pmf/pdf again to reflect this new knowledge.

Of course, your pmf/pdf will be very different if you are told that that person is now attending Little-Bear Kindergarten.

Definition 12.5 (Conditional pmf)**.** Recall that, for discrete random variable, the ***conditional pmf*** of X given Y is defined as

$$p_{X|Y}(x|y) = P[X = x|Y = y]$$

which gives

$$p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x).$$

Example 12.6. For our discussion in Example 12.3 above, we change from using the pmf $p_S(s)$ to $p_{S|Y}(s|1)$. You may also check that this new function of s is a pmf.

12.7. When $X \perp\!\!\!\perp Y$, we can write $p_{X|Y}(x|y) = p_X(x)$ and we say that we "drop the conditioning".

12.8 (***Conditional Expectation***)**.** The ***conditional expected value*** of X given $Y = y$ is

$$\mathbb{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y).$$

which can be generalized into

$$\mathbb{E}[g(X)|Y = y] = \sum_x g(x) p_{X|Y}(x|y).$$

12.9. The ***conditional variance*** of X given $Y = y$ is

$$\mathrm{Var}[X|Y = y] = \sum_x ((x - m(y))^2 p_{X|Y}(x|y) \tag{38}$$

$$= \mathbb{E}\left[X^2|Y = y\right] - (\mathbb{E}[X|Y = y])^2 \tag{39}$$

where $m(y) = \mathbb{E}[X|Y = y]$.

Note that (39) holds even for continuous or mixed random variables.

Definition 12.10. Note that $\mathbb{E}[X|Y = y]$, $\mathbb{E}[g(X)|Y = y]$, and $\mathrm{Var}[X|Y = y]$ are all deterministic functions of a number y. If we change the value of y (the extra information), the values of $\mathbb{E}[X|Y = y]$, $\mathbb{E}[g(X)|Y = y]$, and $\mathrm{Var}[X|Y = y]$ will change.

Extending the above insight, one may also write $\mathbb{E}[X|Y]$, $\mathbb{E}[g(X)|Y]$, and $\mathrm{Var}[X|Y]$. All of these are deterministic functions of a random variable Y. In which case, we can find the expectation of all these functions as

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_y \mathbb{E}[X|Y = y] \times p_Y(y) \tag{40}$$

$$\mathbb{E}[\mathbb{E}[g(X)|Y]] = \sum_y \mathbb{E}[g(X)|Y = y] \times p_Y(y) \tag{41}$$

$$\mathbb{E}[\mathrm{Var}[X|Y]] = \sum_y \mathrm{Var}[X|Y = y] \times p_Y(y) \tag{42}$$

12.11. ***Law of Iterated Expectation***:

$$\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}\left[X\right] \tag{43}$$

$$\mathbb{E}\left[\mathbb{E}\left[g(X)|Y\right]\right] = \mathbb{E}\left[g(X)\right] \tag{44}$$

These properties hold even for continuous or mixed random variables However, in general

$$\mathbb{E}\left[\mathrm{Var}[X|Y]\right] \neq \mathrm{Var}\, X$$

12.12. Now that we have talked about conditional pmf and condition expectation for discrete random variables, it is easy to write down the same concepts for continuous random variables. These are shown in Table 10.

	Discrete	Continuous
Conditional	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
$\mathbb{E}\left[X\|Y=y\right]$	$\sum_x x p_{X\|Y}(x\|y)$	$\int_{-\infty}^{+\infty} x f_{X\|Y}(x\|y)dx$
$\mathbb{E}\left[g(X)\|Y=y\right]$	$\sum_x g(x) p_{X\|Y}(x\|y)$	$\int_{-\infty}^{+\infty} g(x) f_{X\|Y}(x\|y)dx$
$\mathrm{Var}[X\|Y=y]$	$\sum_x ((x-m(y))^2 p_{X\|Y}(x\|y)$	$\int_{-\infty}^{+\infty} ((x-m(y))^2 f_{X\|Y}(x\|y)dx$
$\mathbb{E}\left[\mathbb{E}\left[X\|Y\right]\right]$	$\sum_y \mathbb{E}\left[X\|Y=y\right] p_Y(y)$	$\int_{-\infty}^{+\infty} \mathbb{E}\left[X\|Y=y\right] f_Y(y)dy$
$\mathbb{E}\left[\mathbb{E}\left[g(X)\|Y\right]\right]$	$\sum_y \mathbb{E}\left[g(X)\|Y=y\right] p_Y(y)$	$\int_{-\infty}^{+\infty} \mathbb{E}\left[g(X)\|Y=y\right] f_Y(y)dy$
$\mathbb{E}\left[\mathrm{Var}[X\|Y]\right]$	$\sum_y \mathrm{Var}[X\|Y=y] p_Y(y)$	$\int_{-\infty}^{+\infty} \mathrm{Var}[X\|Y=y] f_Y(y)dy$
$\mathbb{E}\left[\mathrm{Var}[X\|Y]\right]$	$\sum_y \mathrm{Var}[X\|Y=y] p_Y(y)$	$\int_{-\infty}^{+\infty} \mathrm{Var}[X\|Y=y] f_Y(y)dy$

Table 10: Conditional PMF/PDF and Condition Expectation. In this table, $m(y) = \mathbb{E}\left[X|Y=y\right]$.

12.13. If $\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}\left[X\right]$ looks strange to you, try to think about it as combining the mean of the scores for two sections of a class into the total mean. Suppose we have two sections for our class: section 1 and section 2. There are 10 students in section 1 and 20 students in section 2. The mean for section 1 is 70 and the mean for section 2 is 80. To get the mean of all of the 30 students, you find

$$\frac{10 \times 70 + 20 \times 80}{10 + 20}.$$

This can be rewritten as

$$\frac{10}{30} \times 70 + \frac{20}{30} \times 80.$$

12.14. *Substitution Law*: For a (real-valued) function of two random variables,

(a) $P[g(X,Y)=z|Y=y]=P[g(X,y)=z|Y=y]$ [9, eq. 3.13 p 123]

(b) $\mathbb{E}[g(X,Y)|Y=y]=\mathbb{E}[g(X,y)|Y=y]=\sum_x g(x,y)p_{X|Y}(x|y)$

Note that $g(X,y)$ is a function of only one random variable X. The value of y is already conditioned and hence we can treat it as a deterministic constant.

Example 12.15. For bivariate normal whose joint pdf is $f_{X,Y}(x,y)$ is

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{\left(\frac{x-\mathbb{E}X}{\sigma_X}\right)^2-2\rho\left(\frac{x-\mathbb{E}X}{\sigma_X}\right)\left(\frac{y-\mathbb{E}Y}{\sigma_Y}\right)+\left(\frac{y-\mathbb{E}Y}{\sigma_Y}\right)^2}{2\left(1-\rho^2\right)}\right\},$$

(a) given $Y=y$, the conditional pdf of X is

$$\mathcal{N}\left(\rho\frac{\sigma_X}{\sigma_Y}(y-\mathbb{E}Y)+\mathbb{E}X,\sigma_X^2(1-\rho^2)\right)$$

and

(b) given $X=x$, the conditional pdf of Y is

$$\mathcal{N}\left(\rho\frac{\sigma_Y}{\sigma_X}(x-\mathbb{E}X)+\mathbb{E}Y,\sigma_Y^2(1-\rho^2)\right).$$

ECS315 2011/1 Part VI Dr.Prapun

13 Transform methods: Characteristic Functions and Moment Generating functions

Definition 13.1. The ***moment generating function*** (mgf) of a real-valued random variable X is defined by

$$M_X(s) = \mathbb{E}\left[e^{sX}\right].$$

Remarks:

(a) For continuous random variable,

$$\mathbb{E}\left[e^{sX}\right] = \int_{-\infty}^{+\infty} e^{sx} f_X(x)dx,$$

which is the Laplace *transform* of the pdf f_X evaluated at $-s$.

(b) This function is called moment generating function because

$$M_X^{(k)}(0) = \mathbb{E}\left[X^k\right].$$

Definition 13.2. The characteristic function of a random variable X is defined by

$$\varphi_X(v) = \mathbb{E}\left[e^{jvX}\right].$$

Remarks:

(a) If X is a continuous random variable with density f_X, then

$$\varphi_X(v) = \int_{-\infty}^{+\infty} e^{jvx} f_X(x)dx,$$

which is the *Fourier transform* of f_X evaluated at $-v$. More precisely,

$$\varphi_X(v) = \mathcal{F}\{f_X\}(\omega)|_{\omega=-v}. \tag{45}$$

(b) Many references use u or t instead of v.

Example 13.3. You may have learned that the Fourier transform of a Gaussian waveform is a Gaussian waveform. In fact, you have already shown in one of the HW questions that when $X \sim \mathcal{N}(m, \sigma^2)$,

$$\mathcal{F}\{f_X\}(\omega) = \int_{-\infty}^{\infty} f_X(x)\, e^{-j\omega x} dx = e^{-j\omega m - \frac{1}{2}\omega^2\sigma^2}.$$

Using (45), we have

$$\varphi_X(v) = e^{jvm - \frac{1}{2}v^2\sigma^2}.$$

Example 13.4. For $X \sim \mathcal{E}(\lambda)$, we have $\varphi_X(v) = \frac{\lambda}{\lambda - jv}$.

As with the Fourier transform, we can build a large list of commonly used characteristic functions. (You probably remember that rectangular function in time domain gives a sinc function in frequency domain.) When you see a random variable that has the same form of characteristic function as the one that you know, you can quickly make a conclusion about the type and property of that random variable.

Example 13.5. Suppose a random variable X has the characteristic function $\varphi_X(v) = \frac{2}{2 - jv}$. You can quickly conclude that it is an exponential random variable with parameter 2.

For many random variables, it is easy to find its expected value or any moments via the characteristic function. This can be done via the following result.

13.6. $\varphi^{(k)}(v) = j^k \mathbb{E}\left[X^k e^{jvX}\right]$ and $\varphi^{(k)}(0) = j^k \mathbb{E}\left[X^k\right]$.

Example 13.7. When $X \sim \mathcal{E}(\lambda)$,

(a) $\mathbb{E}X = \frac{1}{\lambda}$.

(b) $\operatorname{Var} X = \frac{1}{\lambda^2}$.

One very important properties of characteristic function is that it is very easy to find the characteristic function of a sum of independent random variables.

13.8. Suppose $X \perp\!\!\!\perp Y$. Let $Z = X + Y$. Then, the characteristic function of Z is the product of the characteristic functions of X and Y:

$$\varphi_Z(v) = \varphi_X(v)\varphi_Y(v)$$

Remark: Can you relate this property to the property of the Fourier transform?

Example 13.9. We can use characteristic function to show that the sum of two independent Gaussian random variables is still a Gaussian random variable:

14 Limiting Theorems

Example 14.1. Suppose a chance that you will find a particular rare disease in a person is 1/1,000,000. In a group of 1,000,000 persons, what is the chance that you will find at least one person that has this disease?

Does the probability of $p = 1/1,000,000$ means that in a group of $n = 1,000,000$ persons, you would find roughly 1 persons who has the disease? The answer is YES. Let N be the number of perons who have this disease. Then N will be binomial r.v. with parameter (n, p). The expected number of persons who have the disease will be $n \times p$. However, the probability of having at least one persons who have the disease is only

$$1 - p_N(0) = 1 - (1-p)^n \approx 0.6321.$$

What about something that has 1/2 chance of occurrence? If you perform two trials, then the chance that you find at least one occurrence is $1 - (1 - 1/2)^2 = 3/4$. In general, for n trials of something that has $1/n$ chance of occurrence, the probability that there is at least one occurrence is

$$1 - \left(1 - \frac{1}{n}\right)^n.$$

In this section, we will talk more about the interpretation or the meaning of probability. What does it mean, from a practical point of view, when we say the chances are 1 in 6 a dice will land on number 2? It doesn't mean that in any series of tosses the die will land on number 2 exactly 1 time in 6.

Example 14.2. Toss a (balanced) coin 10 times and you might observe 7 heads, but toss it 1 zillion times and you'll most likely get very near 50 percent. In the 1940s a South African mathematician named John Kerrich decided to test this out in a practical experiment, tossing a coin what must have seemed like 1 zillion times - actually it was 10,000 - and recording the results of each toss. You might think Kerrich would have had better things to do, but he was a war prisoner at the time, having had the bad luck of being a visitor in Copenhagen when the Germans invaded Denmark in April 1940. According to Kerrichs data, after 100 throws he had only 44 percent heads, but by the time he reached 10,000, the number was much closer to half: 50.67 percent. [13, p 95]

From the example above, we see that we need a lot of samples in order to get a meaningful statement about probability. So, our first step would be to review the way that we can describe more than two random variables.

Definition 14.3. As you might expect, to talk about N random variables where $N > 2$, we will need a multivariate pmf or pdf. Let's be more specific by considering n random variables:

$$X_1, X_2, X_3, \ldots, X_n.$$

If all of them are discrete random variables, then they are characterized by their joint pmf:

$$p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P\left[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right].$$

Similarly, we can also talk about the joint pdf of a collection of random variables:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$$

Recall that when $N = 1$ the pdf gives probability per unit length. When $N = 2$, the joint pdf gives probability per unit area. You can then correctly guess that when $N = 3$, the joint pdf gives probability per unit volume.

The definition for expectation can also be extended in the obvious way. For example, to find $\mathbb{E}\left[g(X_1, X_2, \ldots, X_n)\right]$, we evaluate

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(x_1, x_2, \ldots, x_n) f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n.$$

14.4. Most of the time, we will not have to deal directly with the joint pmf/pdf. For example, the following property of expectation will allow you to focus on the expectation of each random variable:

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}\left[X_i\right].$$

Similarly, for the variance,

$$\operatorname{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \operatorname{Var} X_i + 2 \sum_{i \neq j} a_j a_j \operatorname{Cov}\left[X_i, X_j\right].$$

In particular, if the X_i are uncorrelated random variables, then

$$\operatorname{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \operatorname{Var} X_i.$$

Definition 14.5. The concept of independence among many random variables can also be extended from what you know about independence between two random variables. In fact, the following statements are equivalent:

(a) The random variables $X_1, X_2, \ldots, X_n$ are independent.

(b) $\forall B_1 \cdots \forall B_n$

$$P\left[X_1 \in B_1, \ldots, X_n \in B_n\right] = P\left[X_1 \in B_1\right] \times \cdots \times P\left[X_n \in B_n\right]$$

(c) $\forall x_1 \cdots \forall x_n$

$$f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \times \cdots \times f_{X_n}(x_n)$$

(d) $\forall x_1 \cdots \forall x_n$

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = F_{X_1}(x_1) \times \cdots \times F_{X_n}(x_n)$$

14.1 Law of Large Numbers

Definition 14.6. Let $X_1, X_2, \ldots, X_n$ be a collection of random variables with a common mean $\mathbb{E}[X_i] = m$ for all i. In practice, since we do not know m, we use the numerical average, or sample mean,

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

in place of the true, but unknown value, m.

Q: Can this procedure of using M_n as an estimate of m be justified in some sense?

A: This can be done via the law of large number.

14.7. The law of large number basically says that if you have a sequence of i.i.d random variables $X_1, X_2, \ldots$. Then the sample means $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ will converge to the actual mean as $n \to \infty$.

Of course, the above statement is not quite meaningful. We need to be more specific about the word "converge".

14.8. The ***strong law of large numbers*** (SLLN): For a sequence of i.i.d random variables $X_1, X_2, \ldots$, the sample means $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ will converge ***almost surely*** to the actual mean as $n \to \infty$.

The above law is very powerful (strong). It is difficult to prove but its application is vast.

14.9. Application of Strong Law of Large Numbers: If a certain chance experiment is repeated an unlimited number of times under exactly the same conditions, and if the repetitions are independent of each other, then the fraction of times that a given event A occurs will converge with probability 1 to a number that is equal to the probability that A occurs in a single repetition of the experiment. [22, p 20]

- This result is also the mathematical basis for the widespread application of computer simulations to solve practical probability problems. In these applications, the (unknown) probability of a given event in a chance experiment is estimated by the relative frequency of occurrence of the event in a large number of computer simulations of the experiment.

- The mathematical basis for the strong law of large numbers was given for the first time by the famous Russian mathematician Kolmogorov in the twentieth century.

However, because it is stated using limit, it does not tell us much about the case when we have finite number of samples.

Another version of the law of large numbers (LLN) is call the weak law of large numbers (WLLN). It is very easy to prove and in fact was discovered long before SLLN. The implication of WLLN is very similar to SLLN: $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ will converge to the common mean m. They are different in the type of convergence.[26] There is also a probability bound associated with WLLN and hence it is possible to estimate how far in the sequence (how large should n be) for M_n to be "close enough" to the common mean m.

14.10. The ***weak law of large numbers*** (WLLN): Consider a sequence of uncorrelated random variables $X_1, X_2, \ldots$ with common mean m and variance σ^2. Define the sample mean M_n by $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for any given positive value of ε,

$$\lim_{n\to\infty} P\left[|M_n - m| > \varepsilon\right] = 0. \tag{46}$$

In fact,

$$\boxed{P\left[|M_n - m| > \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2}.} \tag{47}$$

- Remark: WLLN is formulated by the Swiss mathematician Jakob Bernoulli in his masterpiece *Ars Conjectandi*.

Let's look closely at (46) and (47). For any kind of LLN, the statement is that M_n will converge to m. Now, suppose we have

[26] In probability theory, there are many type of convergence: almost sure convergence, convergence in probability, convergence in distribution, convergence in mean of order p, etc.

a limited resource (which is the case for any practical application) and we can't go as large as ∞ or have time to wait for n to get to ∞. Then we expect that there will be some difference between M_n and m. So, instead of claiming that M_n and m will be the same, we will allow some small difference; that is, if the difference is not more than ε, we will say that M_n is "close enough" to m. Of course, by the randomness in the M_n, it is possible that the difference will exceed ε. WLLN says that "the probability that M_n is "close enough" to m is very high for large enough value of n." Or, equivalently, "the probability that M_n is NOT "close enough" to m is very low for large enough value of n."

WLLN, as expressed in (47), is then a relationship between three quantities:

(a) ε: how close do you want M_n to be with m,

(b) how confident you want to be when you say that M_n is within $m \pm \varepsilon$,

(c) n: how large the value of n (the number of X_i that are being averaged) should be to guarantee the above two quantities.

Example 14.11. Given ε and σ^2, determine how large n should be so the probability that M_n is within ε of m is at least 0.9.

14.12. *Proof of WLLN*: WLLN is easy to prove via ***Chebyshev's Inequality*** which states that any random variable Z must satisfies

$$P\left[|Z - \mathbb{E}Z| > \varepsilon\right] \leq \frac{\operatorname{Var} Z}{\varepsilon^2}.$$

For us, we shall use $Z = M_n$. Because

$$\mathbb{E}\left[M_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i = m$$

and

$$\text{Var}[M_n] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\, X_i = \frac{1}{n}\sigma^2, \tag{48}$$

we have WLLN as shown earlier in (47).

Remarks:

(a) For (48) to hold, it is sufficient to have uncorrelated X_i's.

(b) Even without the Chebyshev's inequality, by seeing that the variance of M_n decreases to 0 as $n \to \infty$, we can sense that its value would be very close to its mean when n is large.

(c) From (48), we also have

$$\sigma_{M_n} = \frac{1}{\sqrt{n}}\sigma. \tag{49}$$

In words, "when uncorrelated (or independent) random variables each having the same distribution are averaged together, the standard deviation is reduced according to the square root law." [22, p 142].

14.13. Applications

(a) Investment: It is common wisdom in finance that diversification of a portfolio of stocks generally reduces the total risk exposure of the investment. (Recall Example 9.52.)

(b) Inventory control: Aggregation of independent demands at similar retail outlets by replacing the outlets with a single large outlet reduces the total required safety stock.

14.2 Central Limit Theorem (CLT)

In practice, there are many random variables that arise as a sum of many other random variables. In this section, we consider the sum

$$S_n = \sum_{i=1}^{n} X_i \tag{50}$$

where the X_i are i.i.d. with common mean m and common variance σ^2.

- Note that when we talk about X_i being i.i.d., the definition is that they are independent and identically distributed. It is then convenient to talk about a random variable X which shares the same distribution (pdf/pmf) with these X_i. This allow us to write
$$X_i \overset{\text{i.i.d.}}{\sim} X, \tag{51}$$
which is much more compact than saying that the X_i are i.i.d. with the same distribution (pdf/pmf) as X. Moreover, we can also use $\mathbb{E}X$ and σ_X^2 for the common expected value and variance of the X_i.

Q: How does S_n behave?

For the S_n defined above, there are many cases for which we know the pmf/pdf of S_n.

Example 14.14. When the X_i are i.i.d. Bernoulli(p),

Example 14.15. When the X_i are i.i.d. $\mathcal{N}(m, \sigma^2)$,

It is not difficult to find the characteristic function of S_n if we know the common characteristic function $\varphi_X(v)$ of the X_i:
$$\varphi_{S_n}(v) = (\varphi_X(v))^n .$$
If we are lucky, as in the case for the sum of Gaussian random variables in Example 14.15 above, we get $\varphi_{S_n}(v)$ that is of the form that we know. However, $\varphi_{S_n}(v)$ will usually be something we haven't seen before or difficult to find the inverse transform. This is one of the reason why having a way to approximate the sum S_n would be very useful.

There are also some situations where the distribution of the X_i is unknown or difficult to find. In which case, it would be amazing if we can say something about the distribution of S_n.

In the previous section, we consider the sample mean of identically distributed random variables. More specifically, we consider the random variable $M_n = \frac{1}{n}S_n$. We found that M_n will converge to m as n increases to ∞. Here, we don't want to rescale the sum S_n by the factor $\frac{1}{n}$.

14.16 (Approximation of densities and pmfs using the CLT)**.** The actual statement of the CLT is a bit difficult to state. So, we first give you the interpretation/insight from CLT which is very easy to remember and use:

For n large enough, we can approximate S_n by a Gaussian random variable with the same mean and variance as S_n.

Note that the mean and variance of S_n is nm and $n\sigma^2$, respectively. Hence, for n large enough we can approximate S_n by $\mathcal{N}\left(nm, n\sigma^2\right)$. In particular,

(a) $F_{S_n}(s) \approx \Phi\left(\frac{s-nm}{\sigma\sqrt{n}}\right)$.

(b) If the X_i are continuous random variable, then

$$f_{S_n}(s) \approx \frac{1}{\sqrt{2\pi}\sigma\sqrt{n}} e^{-\frac{1}{2}\left(\frac{x-nm}{\sigma\sqrt{n}}\right)^2}.$$

(c) If the X_i are integer-valued, then

$$P[S_n = k] = P\left[k - \frac{1}{2} < S_n \leq k + \frac{1}{2}\right] \approx \frac{1}{\sqrt{2\pi}\sigma\sqrt{n}} e^{-\frac{1}{2}\left(\frac{k-nm}{\sigma\sqrt{n}}\right)^2}.$$

[9, eq (5.14), p. 213].

The approximation is best for k near nm [9, p. 211].

Example 14.17. Approximation for Binomial Distribution: For $X \sim \mathcal{B}(n, p)$, when n is large, binomial distribution becomes difficult to compute directly because of the need to calculate factorial terms.

(a) When p is not close to either 0 or 1 so that the variance is also large, we can use CLT to approxmiate

$$P[X = k] \approx \frac{1}{\sqrt{2\pi \operatorname{Var} X}} e^{-\frac{(k-\mathbb{E}X)^2}{2 \operatorname{Var} X}} \tag{52}$$

$$= \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}. \tag{53}$$

This is called Laplace approximation to the Binomial distribution [26, p. 282].

(b) When p is small, the binomial distribution can be approximated by $\mathcal{P}(np)$ as discussed in Section 8.2.2.

(c) If p is very close to 1, then $n - X$ will behave approximately Poisson.

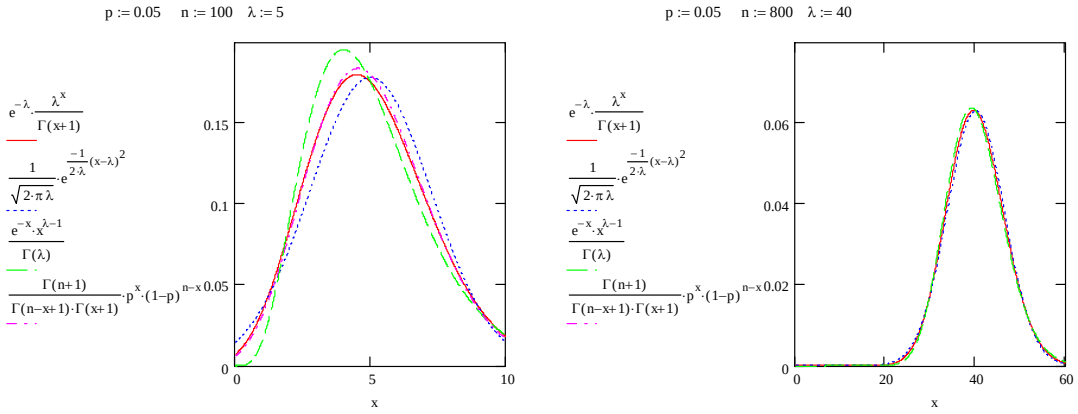

Figure 23: Gaussian approximation to Binomial, Poisson distribution, and Gamma distribution.

Example 14.18. ***An Even Split at Coin Tossing***: Let N be the number of heads that show up when a fair coin is tossed $2n$ times. Then, $N \sim \mathcal{B}(2n, 1/2)$. Let $p_n = P[N = n]$, the probability of getting exactly n heads (and hence exactly n tails). Then,

$$p_n = \frac{\binom{2n}{n}}{2^{2n}}.$$

(a) Sometimes, to work theoretically with large factorials, we use Stirling's Formula:

$$n! \approx \sqrt{2\pi n} n^n e^{-n} = \left(\sqrt{2\pi e}\right) e^{\left(n+\frac{1}{2}\right)\ln\left(\frac{n}{e}\right)}. \tag{54}$$

By Stirling's Formula, we have

$$\binom{2n}{n} = \frac{(2n)!}{n!n!} \approx \frac{\sqrt{2\pi 2n}(2n)^{2n}e^{-2n}}{\left(\sqrt{2\pi n} n^n e^{-n}\right)^2} = \frac{4^n}{\sqrt{\pi n}}.$$

Hence [14, Problem 18],

$$p_n \approx \frac{1}{\sqrt{\pi n}}. \tag{55}$$

See Figure 24 for comparison of p_n and its approximation via Stirling's formula.

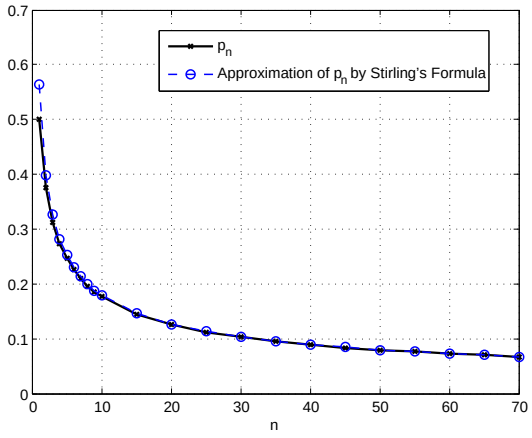

Figure 24: Comparison of p_n and its approximation via Stirling's formula

(b) We can also use CLT to approximate p_n. Note that $\mathbb{E}N = n$ and $\operatorname{Var} N = n/2$. From (52),

$$P[N = n] \approx \frac{1}{\sqrt{2\pi\frac{n}{2}}} e^{-\frac{(n-n)^2}{2\times\frac{n}{2}}} = \frac{1}{\sqrt{\pi n}}.$$

(c) Note that Poisson approximation is not applicable here because $p = 1/2$ is not small. When Poisson approximation is

mistakenly used, we have

$$P[N=n] \approx e^{-n}\frac{n^n}{n!} \approx e^{-n}\frac{n^n}{\sqrt{2\pi n}n^n e^{-n}} = \frac{1}{\sqrt{2\pi n}}$$

which is off by a factor of $1/\sqrt{2}$.

Example 14.19. Normal Approximation to Poisson Distribution with large λ: Let $X \sim \mathcal{P}(\lambda)$. X can be though of as a sum of i.i.d. $X_i \sim \mathcal{P}(\lambda_n)$, i.e., $X = \sum_{i=1}^{n} X_i$, where $n\lambda_n = \lambda$. Hence X is approximately normal $\mathcal{N}(\lambda, \lambda)$ for λ large.

Some says that the normal approximation is good when $\lambda > 5$.

Example 14.20. In statistics, the central limit theorem is the basis for constructing **confidence intervals**.

14.21. A more precise statement for CLT can be expressed via the convergence of the characteristic function. In particular, suppose that $(X_k)_{k\geq 1}$ is a sequence of i.i.d. random variables with mean m and variance $0 < \sigma^2 < \infty$. Let $S_n = \sum_{k=1}^{n} X_k$. It can be shown that

(a) the characteristic function of $\frac{S_n - mn}{\sigma\sqrt{n}}$ converges pointwise to the characteristic function of $\mathcal{N}(0,1)$ and that

(b) the characteristic function of $\frac{S_n - mn}{\sqrt{n}}$ converges pointwise to the characteristic function of $\mathcal{N}(0,\sigma)$.

To see this, let $Z_k = \frac{X_k - m}{\sigma} \overset{\text{iid}}{\sim} Z$ and $Y_n = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_k$. Then, $\mathbb{E}Z = 0$, $\operatorname{Var} Z = 1$, and $\varphi_{Y_n}(t) = \left(\varphi_Z(\frac{t}{\sqrt{n}})\right)^n$. By approximating $e^x \approx 1 + x + \frac{1}{2}x^2$. We have $\varphi_X(t) \approx 1 + jt\mathbb{E}X - \frac{1}{2}t^2\mathbb{E}\left[X^2\right]$ and

$$\varphi_{Y_n}(t) = \left(1 - \frac{1}{2}\frac{t^2}{n}\right)^n \to e^{-\frac{t^2}{2}},$$

which is the characteristic function of $\mathcal{N}(0,1)$.

- The case of Bernoulli(1/2) was derived by Abraham de Moivre around 1733. The case of Bernoulli(p) for $0 < p < 1$ was considered by Pierre-Simon Laplace [9, p. 208].

ECS315 2011/1 Part VII.1 Dr.Prapun

15 Random Vector

In Section 14, we have introduced the way to deal with more than two random variables. In particular, we introduce the concepts of joint pmf:

$$p_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = P\left[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right]$$

and joint pdf:

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$$

of a collection of random variables.

Definition 15.1. You may notice that it is tedious to write the n-tuple $(X_1, X_2, \ldots, X_n)$ every time that we want to refer to this collection of random variables. A more convenient notation uses a ***column vector*** $\mathbf{X}$ to represent all of them at once, keeping in mind that the ith component of $\mathbf{X}$ is the random variable X_i. This allows us to express

(a) $p_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$ as $p_{\mathbf{X}}(\mathbf{x})$ and

(b) $f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$ as $f_{\mathbf{X}}(\mathbf{x})$.

When the random variables are separated into two groups, we may label those in a group as $X_1, X_2, \ldots, X_n$ and those in another group as $Y_1, Y_2, \ldots, Y_m$. In which case, we can express

(a) $p_{X_1,\ldots,X_n,Y_1,\ldots,Y_m}(x_1,\ldots,x_n,y_1,\ldots,y_m)$ as $p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$ and

(b) $f_{X_1,\ldots,X_n,Y_1,\ldots,Y_m}(x_1,\ldots,x_n,y_1,\ldots,y_m)$ as $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$.

Definition 15.2. Random vectors $\mathbf{X}$ and $\mathbf{Y}$ are ***independent*** if and only if

(a) Discrete: $p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = p_{\mathbf{X}}(\mathbf{x})p_{\mathbf{Y}}(\mathbf{y})$.

(b) Continuous: $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})f_{\mathbf{Y}}(\mathbf{y})$.

Definition 15.3. A random vector $\mathbf{X}$ contains many random variables. Each of these random variables has its own expected value. We can represent the expected values of all these random variables in the form of a vector as well by using the notation $\mathbb{E}[\mathbf{X}]$. This is a vector whose ith component is $\mathbb{E}X_i$.

- In other words, the expectation $\mathbb{E}[\mathbf{X}]$ of a random vector $\mathbf{X}$ is defined to be the vector of expectations of its entries.
- $\mathbb{E}[\mathbf{X}]$ is usually denoted by $\mu_{\mathbf{X}}$ or $m_{\mathbf{X}}$.

15.4. For non-random matrix A, B, C and a random vector $\mathbf{X}$,

$$\mathbb{E}[A\mathbf{X}B + C] = A(\mathbb{E}\mathbf{X})B + C.$$

Definition 15.5. Recall that a random vector is simply a vector containing random variables as its components. We can also talk about ***random matrix*** which is simply a matrix whose entries are random variables. In which case, we define the expectation of a random matrix to be a matrix whose entries are expectation of the corresponding random variables in the random matrix.

Correlation and covariance are important quantities that capture linear dependency between two random variables. When we have many random variables, there are many possible pairs to find correlation $\mathbb{E}[X_iX_j]$ and covariance $\mathrm{Cov}[X_i, X_j]$. All of the correlation values can be expressed at once using the correlation matrix.

Definition 15.6. The ***correlation matrix*** $R_{\mathbf{X}}$ of a random vector $\mathbf{X}$ is defined by

$$R_{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right].$$

Note that it is symmetric and that the ij-entry of $R_{\mathbf{X}}$ is simply $\mathbb{E}[X_i X_j]$.

Example 15.7. Consider $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$.

Definition 15.8. Similarly, all of the covariance values can be expressed at once using the covariance matrix. The covariance matrix $C_{\mathbf{X}}$ of a random vector $\mathbf{X}$ is defined as

$$\begin{aligned} C_{\mathbf{X}} &= \mathbb{E}\left[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^T\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] - (\mathbb{E}\mathbf{X})(\mathbb{E}\mathbf{X})^T \\ &= R_{\mathbf{X}} - (\mathbb{E}\mathbf{X})(\mathbb{E}\mathbf{X})^T. \end{aligned}$$

Note that it is symmetric and that the ij-entry of $C_{\mathbf{X}}$ is simply $\mathrm{Cov}[X_i, X_j]$.

- In some references, $\Lambda_{\mathbf{X}}$ or $\Sigma_{\mathbf{X}}$ is used instead of $C_{\mathbf{X}}$.

15.9. Properties of covariance matrix:

(a) For i.i.d. X_i each with variance σ^2, $C_{\mathbf{X}} = \sigma^2 I$.

(b) $\mathrm{Cov}[A\mathbf{X} + b] = AC_{\mathbf{X}}A^T$.

In addition to the correlations and covariances of the elements of one random vector, it is useful to refer to the correlations and covariances of elements of two random vectors.

Definition 15.10. If $\mathbf{X}$ and $\mathbf{Y}$ are both random vectors (not necessarily of the same dimension), then their cross-correlation matrix is

$$R_{\mathbf{XY}} = \mathbb{E}\left[\mathbf{X}\mathbf{Y}^T\right].$$

and their cross-covariance matrix is

$$C_{\mathbf{XY}} = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T\right].$$

Example 15.11. ***Jointly Gaussian random vector*** $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Lambda)$:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Lambda)}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \Lambda^{-1}(\mathbf{x}-\mathbf{m})}.$$

(a) $\mathbf{m} = \mathbb{E}\mathbf{X}$ and $\Lambda = C_{\mathbf{X}} = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^T\right]$.

(b) For bivariate normal, $X_1 = X$ and $X_2 = Y$. We have

$$\Lambda = \begin{pmatrix} \sigma_X^2 & \mathrm{Cov}[X,Y] \\ \mathrm{Cov}[X,Y] & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

ECS315 2011/1 Part VII.2 Dr.Prapun

16 Introduction to Stochastic Processes (Random Processes)

In the previous section, we talked about how to deal with finite number of random variables, say $X_1, X_2, \ldots, X_n$. In this section, we will consider infinitely many random variables. Most of the time, these random variables will be indexed by time. So, the obvious notation for random process would be $X(t)$. As in the signals-and-systems class, time can be discrete or continuous. When time is discrete, it may be more appropriate to use $X_1, X_2, \ldots$ or $X[1], X[2], X[3], \ldots$ to denote a random process.

16.1. Two perspectives:

(a) We can view a random process as a collection of many random variables indexed by t.

(b) We can also view a random process as the outcome of a random experiment, where the outcome of each trial is a waveform (a function) that is a function of t.

16.2. Formal definition of random process requires going back to the probability space $(\Omega, \mathcal{A}, P)$.

Recall that a random variable X is in fact a deterministic function of the outcome ω from Ω. So, we should have been writing it as $X(\omega)$. However, as we get more familiar with the concept of random variable, we usually drop the "(ω)" part and simply refer to it as X.

For random process, we have $X(t, \omega)$. This two-argument expression corresponds to the two perspectives that we have just discussed earlier.

(a) When you fix the time t, you get a random variable from a random process.

(b) When you fix ω, you get a deterministic function of time from a random process.

As we get more familiar with the concept of random processes, we again drop the ω argument.

Definition 16.3. A ***sample function*** $x(t, \omega)$ is the time function associated with the outcome ω of an experiment.

Example 16.4. Consider the random process defined by

$$X(t) = A \times \cos(1000t)$$

where A is a random variable. For example, A could be a Bernoulli random variable with parameter p.

This is a good model for a digital transmission via amplitude modulation.

(a) Consider the time $t = 2$ ms. $X(t)$ is a random variable taking the value $1\cos(2) = -0.4161$ with probability p and value $0\cos(2) = 0$ with probability $1 - p$.

If you consider $t = 4$ ms. $X(t)$ is a random variable taking the value $1\cos(4) = -0.6536$ with probability p and value $0\cos(4) = 0$ with probability $1 - p$.

(b) From another perspective, we can look at the process $X(t)$ as two possible waveforms $\cos(1000t)$ and 0. The first one happens with probability p; the second one happens with probability $1 - p$. In this view, notice that each of the waveforms is not random. They are deterministic. Randomness in this situation is associated not with the waveform but with the uncertainty as to which waveform will occur in a given trial

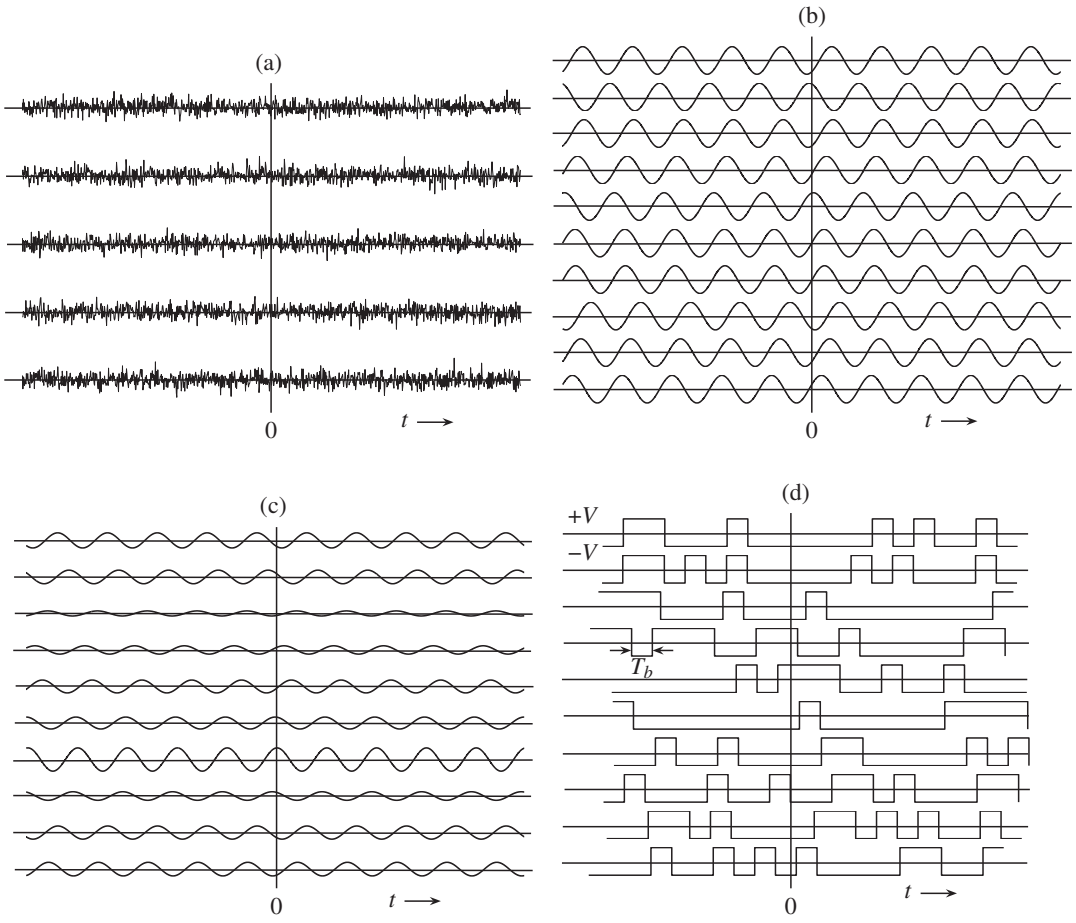

Figure 25: Typical ensemble members for four random processes commonly encountered in communications: (a) thermal noise, (b) uniform phase, (c) Rayleigh fading process, and (d) binary random data process. [16, Fig. 3.8]

Definition 16.5. At any particular time t, because we have a random variable, we can also find its expected value. The function $m_X(t)$ captures these expected values as a function of time:

$$m_X(t) = \mathbb{E}\left[X(t)\right].$$

16.1 Autocorrelation Function and WSS

Definition 16.6. ***Autocorrelation Function***: One of the most important characteristics of a random process is its autocorrelation function, which leads to the spectral information of the random process. The frequency content process depends on the rapidity of the amplitude change with time. This can be measured correlating the values of the process at two time instances t_l and t_2. The autocorrelation function $R_X(t_1, t_2)$ for a random process $X(t)$ is defined by

$$R_X(t_1, t_2) = \mathbb{E}\left[X(t_1)X(t_2)\right].$$

Example 16.7. The random process $x(t)$ is a slowly varying process compared to the process $y(t)$ in Figure 26. For $x(t)$, the values at t_l and t_2 are similar; that is, have stronger correlation. On the other hand, for $y(t)$, values at t_l and t_2 have little resemblance, that is, have weaker correlation.

Example 16.8. Consider a random process

$$x(t) = 5\cos(7t + \Theta)$$

where Θ ia a uniform random variable on the interval $(0, 2\pi)$.

$$\begin{aligned} m_X(t) = \mathbb{E}\left[X(t)\right] &= \int_{-\infty}^{+\infty} 5\cos(7t+\theta) f_\Theta(\theta) d\theta \\ &= \int_0^{2\pi} 5\cos(7t+\theta)\frac{1}{2\pi} d\theta = 0. \end{aligned}$$

and

$$\begin{aligned} R_X(t_1, t_2) &= \mathbb{E}\left[X(t_1)X(t_2)\right] \\ &= \mathbb{E}\left[5\cos(7t_1+\Theta) \times 5\cos(7t_2+\Theta)\right] \\ &= \frac{25}{2}\cos\left(7(t_2 - t_1)\right). \end{aligned}$$

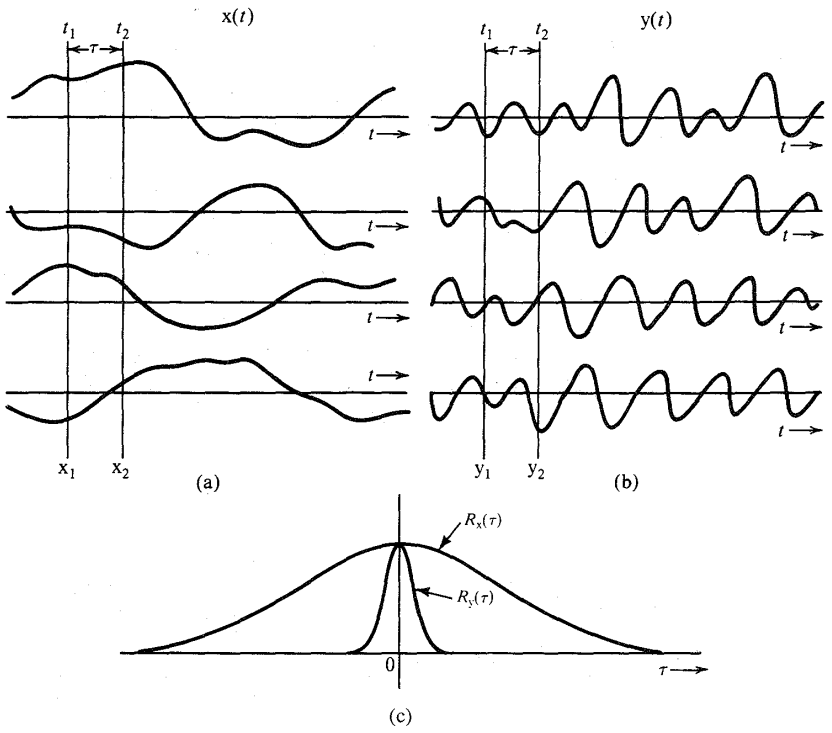

Figure 26: Autocorrelation functions for a slowly varying and a rapidly varying random process [12, Fig. 11.4]

Definition 16.9. A random process whose statistical characteristics do not change with time is classified as a ***stationary*** random process. For a stationary process, we can say that a shift of time origin will be impossible to detect; the process will appear to be the same.

Example 16.10. The random process representing the temperature of a city is an example of a ***nonstationary*** process, because the temperature statistics (mean value, for example) depend on the time of the day.

On the other hand, the noise process is stationary, because its statistics (the mean ad the mean square values, for example) do not change with time.

16.11. In general, it is not easy to determine whether a process is stationary. In practice, we can ascertain stationary if there is no change in the signal-generating mechanism. Such is the case for the noise process.

A process may not be stationary in the strict sense. A more relaxed condition for stationary can also be considered.

Definition 16.12. A random process $X(t)$ is ***wide-sense stationary (WSS)*** if

(a) $m_X(t)$ is a constant

(b) $R_X(t_1, t_2)$ depends only on the time difference $t_2 - t_1$ and does not depend on the specific values of t_1 and t_2.

In which case, we can write the correlation function as $R_X(\tau)$ where $\tau = t_2 - t_1$.

- One important consequence is that $\mathbb{E}\left[X^2(t)\right]$ will be a constant as well.

Example 16.13. The random process defined in Example 16.7 is WSS with

$$R_X(\tau) = \frac{25}{2}\cos(7\tau).$$

16.2 Power Spectral Density (PSD)

An electrical engineer instinctively thinks of signals and linear systems in terms of their frequency-domain descriptions. Linear systems are characterized by their frequency response (the transfer function), and signals are expressed in terms of the relative amplitudes and phases of their frequency components (the Fourier transform). transform). From a knowledge of the input spectrum and transfer function, the response of a linear system to a given signal can be obtained in terms of the frequency content of that signal. This is an important procedure for deterministic signals. We may wonder if similar methods may be found for random processes.

In the study of stochastic processes, the power spectral density function, $S_X(f)$, provides a frequency-domain representation of the time structure of $X(t)$. Intuitively, $S_X(f)$ is the expected value of the squared magnitude of the Fourier transform of a sample function of $X(t)$.

You may recall that not all functions of time have Fourier transforms. For many functions that extend over infinite time, the Fourier transform does not exist. Sample functions $x(t)$ of a stationary stochastic process $X(t)$ are usually of this nature. To work with these functions in the frequency domain, we begin with $X_T(t)$, a truncated version of $X(t)$. It is identical to $X(t)$ for $-T \leq t \leq T$ and 0 elsewhere. We use $\mathcal{F}\{X_T\}(f)$ to represent the Fourier transform of $X_T(t)$ evaluated at the frequency f.

Definition 16.14. Consider a WSS process $X(t)$. The ***power spectral density*** (PSD) is defined as

$$\begin{aligned} S_X(f) &= \lim_{t\to\infty} \frac{1}{2T}\mathbb{E}\left[|\mathcal{F}\{X_T\}(f)|^2\right] \\ &= \lim_{t\to\infty} \frac{1}{2T}\mathbb{E}\left[\left|\int_{-T}^{T} X(t)e^{-j2\pi ft}dt\right|^2\right] \end{aligned}$$

We refer to $S_X(f)$ as a density function because it can be interpreted as the amount of power in $X(t)$ in the small band of frequencies from f to $f + df$.

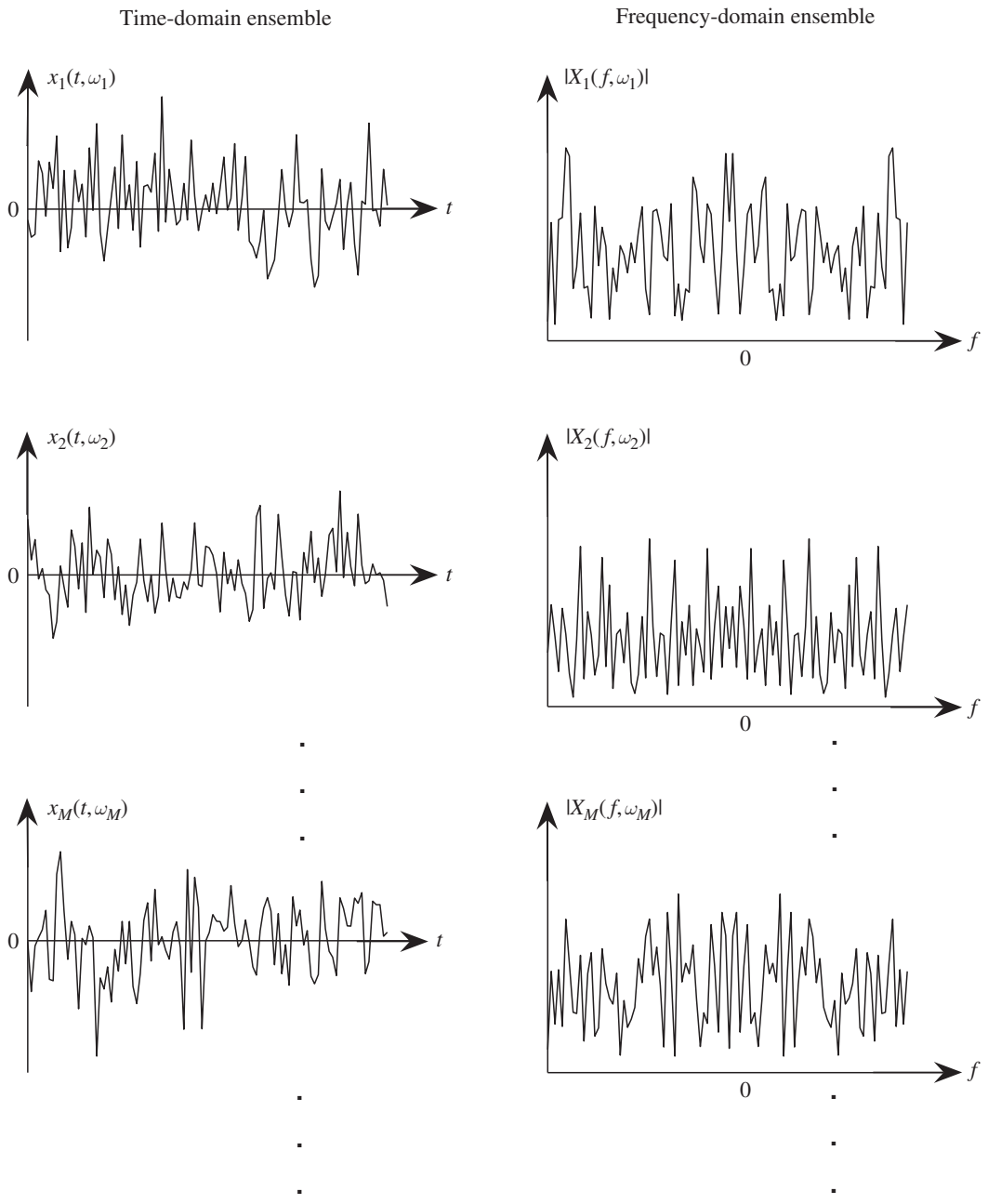

Figure 27: Fourier transforms of member functions of a random process. For simplicity, only the magnitude spectra are shown. [16, Fig. 3.9]

16.15. ***Wiener-Khinchine theorem***: the PSD of a WSS random process is the Fourier transform of its autocorrelation function:

$$S_X(f) = \int_{-\infty}^{+\infty} R_X(\tau) e^{-j2\pi f\tau} d\tau$$

and

$$R_X(\tau) = \int_{-\infty}^{+\infty} S_X(f) e^{j2\pi f\tau} df.$$

One important consequence is

$$R_X(0) = \mathbb{E}\left[X^2(t)\right] = \int_{-\infty}^{+\infty} S_X(f) df.$$

Theorem 16.16. When we input $X(t)$ through an LTI system whose frequency response is $H(f)$. Then, the PSD of the output $Y(t)$ will be given by

$$S_Y(f) = S_X(f)|H(f)|^2.$$

Example 16.17. In Theorem 16.16, suppose $H(f)$ is a bandpass filter whose passband is from f_1 to f_2, find $\mathbb{E}\left[Y^2(t)\right]$ from the $S_X(f)$.

17 Poisson Processes

Now that we have learned more about expectation, independence, and continuous random variables, we can say more about the Poisson processes.

17.1. Recall from 8.35 that for a homogeneous Poisson process, the number N of arrivals during a time interval of duration T is a Poisson random variable with parameter $\alpha = \lambda T$. Note also that α is the expected value of N [Example 8.54].

17.2. Generalization of 17.1: If we consider n non-overlapping intervals. Denote that number of arrivals in these intervals by $N_1, N_2, \ldots, N_n$. Then, $N_1, N_2, \ldots, N_n$ are independent Poisson random variables with $\mathbb{E}[N_i] = \lambda \times$ the length of the corresponding interval.

17.3. Another fact about the Poisson process is that if you consider any ***inter-arrival time***, that is the time from the moment that one customer arrives to the moment that the next customer arrive, it **will be exponential random variable with parameter** λ. This is why we also call λ as the rate for the exponential random variable.

17.4. Poisson process has both a discrete component (Poisson distribution for the number of arrivals in a specific time interval) and a continuous component (exponential distribution for the inter-arrival times).

Example 17.5. The emission of alpha-particles by a piece of radioactive material can be described by a Poisson process: the number of particles emitted in any fixed time interval is a discrete random variable with a Poisson distribution and the times between successive emissions are continuous random variables with an exponential distribution.

Figure 28 summarizes the two key properties of the Poisson process. If we use the same notation as in Figure 28, we see that

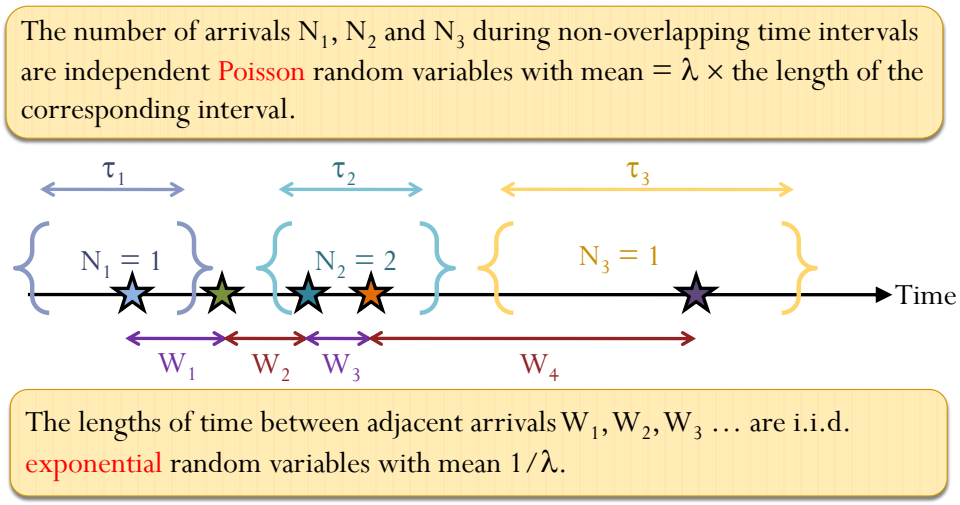

Figure 28: Key Properties of Poisson Process

during the time $W_1 + W_2 + \cdots + W_n$, there are n arrivals. So, on average, the arrival rate is

$$\frac{n}{W_1 + W_2 + \cdots + W_n} = \left(\frac{W_1 + W_2 + \cdots + W_n}{n} \right)^{-1}.$$

With the help of LLN which we will discuss later in this class,

$$\frac{W_1 + W_2 + \cdots + W_n}{n} \to \mathbb{E}W_1 \text{ as } n \to \infty$$

and therefore the arrival rate is $1/\mathbb{E}W_1$. Recall that for exponential random variable with parameter λ, the expected value is $1/\lambda$. So, plugging $\mathbb{E}W_1 = 1/\lambda$ into our calculation of the rate, we see that the arrival rate is

$$\frac{1}{\mathbb{E}W_1} = \frac{1}{1/\lambda} = \lambda.$$

Therefore, calling the parameter λ in the exponential random variable and the Poisson process as "rate" now makes sense.

18 Generation of Random Variable

18.1. Left-continuous inverse: $g^{-1}(y) = \inf\{x \in \mathbb{R} : g(x) \geq y\}$, $y \in (0,1)$

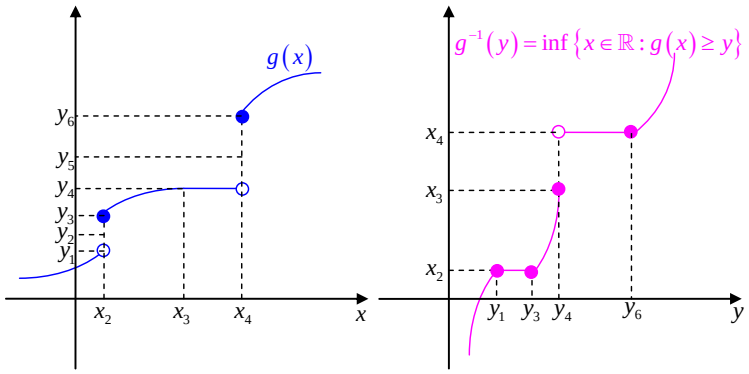

Figure 29: Left-continuous inverse on (0,1)

- *Trick*: Just flip the graph along the line $x = y$, then make the graph left-continuous.
- If g is a cdf, then only consider $y \in (0,1)$. It is called the inverse CDF [6, Def 8.4.1, p. 238] or quantile function.
  - In [23, Def 2.16, p. 25], the inverse CDF is defined using strict inequality ">" rather than "≥".
- See Table 11 for examples.

Distribution	F	F^{-1}
Exponential	$1 - e^{-\lambda x}$	$-\frac{1}{\lambda}\ln(u)$
Extreme value	$1 - e^{-e^{\frac{x-a}{b}}}$	$a + b \ln\ln u$
Geometric	$1 - (1-p)^i$	$\left\lceil \frac{\ln u}{\ln(1-p)} \right\rceil$
Logistic	$1 - \frac{1}{1+e^{\frac{x-\mu}{b}}}$	$\mu - b\ln\left(\frac{1}{u} - 1\right)$
Pareto	$1 - x^{-a}$	$u^{-\frac{1}{a}}$
Weibull	$1 - e^{\left(\frac{x}{a}\right)^b}$	$a(\ln u)^{\frac{1}{b}}$

Table 11: Left-continuous inverse

18.2. Inverse-Transform Method: To generate a random variable X with CDF F, set $X = F^{-1}(U)$ where U is uniform on $(0, 1)$. Here, F^{-1} is the left-continuous inverse of F.

Example 18.3. For example, to generate $X \sim \mathcal{E}(\lambda)$, set $X = -\frac{1}{\lambda} \ln(U)$

A Math Review

A.1. By definition,

- $\sum_{n=1}^{\infty} a_n = \sum_{n\in\mathbb{N}} a_n = \lim_{N\to\infty} \sum_{n=1}^{N} a_n$ and
- $\prod_{n=1}^{\infty} a_n = \prod_{n\in\mathbb{N}} a_n = \lim_{N\to\infty} \prod_{n=1}^{N} a_n$.

A.1 Summations

A.2. Basic formulas:

By the sixth century b.c.e., the Pythagoreans already knew how to find a sum of consecutive natural numbers:

$$\sum_{k=1}^{n} k = \frac{n\,(n+1)}{2}$$

Archimedes of Syracuse (c. 287-212 B.C.E.), the greatest mathematician of antiquity, also discovered how to calculate a sum of squares. Translated into contemporary symbolism, his work shows that

$$\sum_{k=1}^{n} k^2 = \frac{n\,(n+1)\,(2n+1)}{6} = \frac{1}{6}\left(2n^3+3n^2+n\right)$$

Throughout the next two millennia, the search for general formulas for $\sum_{k=1}^{n} k^r$, a sum of consecutive rth powers for any fixed natural number r, became a recurring theme of study, primarily because such sums could be used to find areas and volumes.

- $\sum\limits_{k=1}^{n} k^3 = \left(\sum\limits_{k=1}^{n} k\right)^2 = \frac{1}{4}n^2\,(n+1)^2 = \frac{1}{4}\left(n^4+2n^3+n^2\right)$

A nicer formula is given by

$$\sum_{k=1}^{n} k\,(k+1)\cdots(k+d) = \frac{1}{d+2}n\,(n+1)\cdots(n+d+1) \qquad (56)$$

A.3. Let $g\,(n) = \sum\limits_{k=0}^{n} h\,(k)$ where h is a polynomial of degree d. Then, g is a polynomial of degree $d+1$; that is $g\,(n) = \sum\limits_{m=1}^{d+1} a_m x^m$.

- To find the coefficients a_m, evaluate $g(n)$ for $n = 1, 2, \ldots, d+1$. Note that the case when $n = 0$ gives $a_0 = 0$ and hence the sum starts with $m = 1$.
- Alternative, first express $h(k)$ in terms of summation of polynomials:

$$h(k) = \left(\sum_{i=0}^{d-1} b_i k(k+1) \cdots (k+i) \right) + c. \tag{57}$$

To do this, substitute $k = 0, -1, -2, \ldots, -(d-1)$.

- $k^3 = k(k+1)(k+2) - 3k(k+1) + k$

Then, to get $g(n)$, use (56).

A.4. Geometric Sums:

(a) $\sum_{i=0}^{\infty} \rho^i = \frac{1}{1-\rho}$ for $|\rho| < 1$

(b) $\sum_{i=k}^{\infty} \rho^i = \frac{\rho^k}{1-\rho}$

(c) $\sum_{i=a}^{b} \rho^i = \frac{\rho^a - \rho^{b+1}}{1-\rho}$

(d) $\sum_{i=0}^{\infty} i\rho^i = \frac{\rho}{(1-\rho)^2}$

(e) $\sum_{i=a}^{b} i\rho^i = \frac{\rho^{b+1}(b\rho - b - 1) - \rho^a(a\rho - a - \rho)}{(1-\rho)^2}$

(f) $\sum_{i=k}^{\infty} i\rho^i = \frac{k\rho^k}{1-\rho} + \frac{\rho^{k+1}}{(1-\rho)^2}$

(g) $\sum_{i=0}^{\infty} i^2\rho^i = \frac{\rho + \rho^2}{(1-\rho)^3}$

A.5. Double Sums:

(a) $\left(\sum_{i=1}^{n} a_i \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j$

(b) $\sum_{j=1}^{\infty}\sum_{i=j}^{\infty} f(i,j) = \sum_{i=1}^{\infty}\sum_{j=1}^{i} f(i,j) = \sum_{(i,j)} 1[i \geq j] f(i,j)$

A.6. Exponential Sums:

- $e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots$
- $\lambda e^{\lambda} + e^{\lambda} = 1 + 2\lambda + 3\frac{\lambda^2}{2!} + 4\frac{\lambda^3}{3!} + \ldots = \sum_{k=1}^{\infty} k\frac{\lambda^{k-1}}{(k-1)!}$

A.7. Suppose h is a polynomial of degree d, then

$$\sum_{k=0}^{\infty} h(k)\frac{\lambda^k}{k!} = g(\lambda)e^{\lambda},$$

where g is another polynomial of the same degree. For example,

$$\sum_{k=0}^{\infty} k^3\frac{\lambda^k}{k!} = \left(\lambda^3 + 3\lambda^2 + \lambda\right)e^{\lambda}. \tag{58}$$

This result can be obtained by several techniques.

(a) Start with $e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$. Then, we have

$$\sum_{k=0}^{\infty} k^3\frac{\lambda^k}{k!} = \lambda\frac{d}{d\lambda}\left(\lambda\frac{d}{d\lambda}\left(\lambda\frac{d}{d\lambda}e^{\lambda}\right)\right).$$

(b) We can expand

$$k^3 = k(k-1)(k-2) + 3k(k-1) + k. \tag{59}$$

similar to (57). Now note that

$$\sum_{k=0}^{\infty} k(k-1)\cdots(k-(\ell-1))\frac{\lambda^k}{k!} = \lambda^{\ell}e^{\lambda}.$$

Therefore, the coefficients of the terms in (59) directly becomes the coefficients in (58)

A.8. ***Zeta function*** $\xi(s)$ is defined for any complex number s with $\mathrm{Re}\{s\} > 1$ by the Dirichlet series: $\xi(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$.

- For real-valued nonnegative x

  (a) $\xi(x)$ converges for $x > 1$

  (b) $\xi(x)$ diverges for $0 < x \leq 1$

  [9, Q2.48 p 105].

- $\xi(1) = \infty$ corresponds to harmonic series.

A.9. ***Abel's theorem***: Let $a = (a_i : i \in \mathbb{N})$ be any sequence of real or complex numbers and let

$$G_a(z) = \sum_{i=0}^{\infty} a_i z^i,$$

be the power series with coefficients a. Suppose that the series $\sum_{i=0}^{\infty} a_i$ converges. Then,

$$\lim_{z \to 1^-} G_a(z) = \sum_{i=0}^{\infty} a_i. \tag{60}$$

In the special case where all the coefficients a_i are nonnegative real numbers, then the above formula (60) holds also when the series $\sum_{i=0}^{\infty} a_i$ does not converge. I.e. in that case both sides of the formula equal $+\infty$.

A.2 Inequalities

A.10. Inequalities involving exponential and logarithm.

(a) For any x,

$$e^x \leq 1 + x$$

with equality if and only if $x = 0$.

(b) If we consider $x > -1$, then we have $\ln(x+1) \leq x$. If we replace $x + 1$ by x, then we have $\ln(x) \leq x - 1$ for $x > 0$. If we replace x by $\frac{1}{x}$, we have $\ln(x) \geq 1 - \frac{1}{x}$. This give the fundamental inequalities of information theory:

$$1 - \frac{1}{x} \leq \ln(x) \leq x - 1 \text{ for } x > 0$$

with equality if and only if $x = 1$. Alternative forms are listed below.

(i) For $x > -1$, $\frac{x}{1+x} \leq \ln(1+x) < x$ with equality if and only if $x = 0$.

(ii) For $x < 1$, $x \leq -\ln(1-x) \leq \frac{x}{1-x}$ with equality if and only if $x = 0$.

A.11. For $|x| \leq 0.5$, we have

$$e^{x-x^2} \leq 1 + x \leq e^x. \tag{61}$$

This is because

$$x - x^2 \leq \ln(1+x) \leq x, \tag{62}$$

which is semi-proved by the plot in Figure 30.

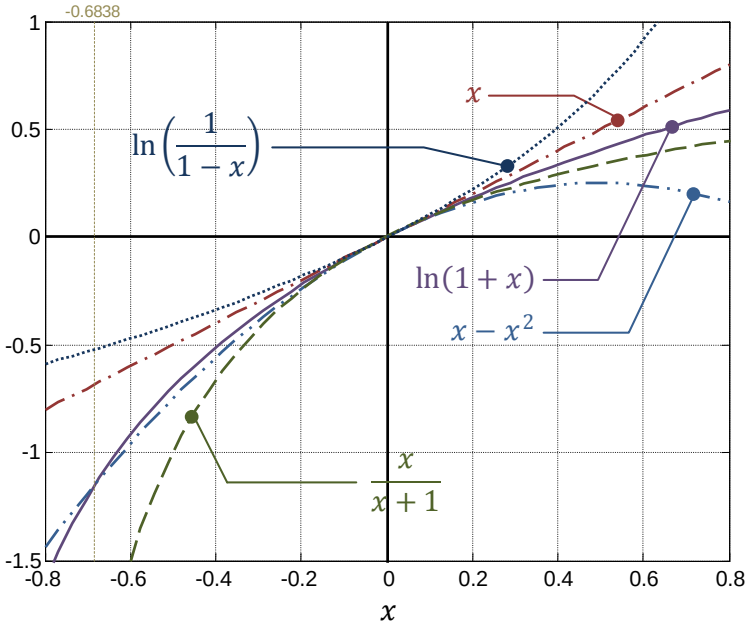

Figure 30: Bounds for $\ln(1+x)$ when x is small.

A.12. Consider a triangular array of real numbers $(x_{n,k})$. Suppose (i) $\sum_{k=1}^{r_n} x_{n,k} \to x$ and (ii) $\sum_{k=1}^{r_n} x_{n,k}^2 \to 0$. Then,

$$\prod_{k=1}^{r_n} (1 + x_{n,k}) \to e^x.$$

Moreover, suppose the sum $\sum_{k=1}^{r_n} |x_{n,k}|$ converges as $n \to \infty$ (which automatically implies that condition (i) is true for some x). Then,

condition (ii) is equivalent to condition (iii) where condition (iii) is the requirement that $\max_{k\in[r_n]} |x_{k,n}| \to 0$ as $n \to \infty$.

Proof. When n is large enough, conditions (ii) and (iii) each implies that $|x_{n,k}| \le 0.5$. (For (ii), note that we can find n large enough such that $|x_{n,k}|^2 \le \sum_k x_{n,k}^2 \le 0.5^2$.) Hence, we can apply (A.11) and get

$$e^{\sum_{k=1}^{r_n} x_{n,k} - \sum_{k=1}^{r_n} x_{n,k}^2} \le \prod_{k=1}^{r_n} (1 + x_{n,k}) \le e^{\sum_{k=1}^{r_n} x_{n,k}}. \tag{63}$$

Suppose $\sum_{k=1}^{r_n} |x_{n,k}| \to x_0$. To show that (iii) implies (ii), let $a_n = \max_{k\in[r_n]} |x_{k,n}|$. Then,

$$0 \le \sum_{k=1}^{r_n} x_{n,k}^2 \le a_n \sum_{k=1}^{r_n} |x_{n,k}| \to 0 \times x_0 = 0.$$

On the other hand, suppose we have (ii). Given any $\varepsilon > 0$, by (ii), $\exists n_0$ such that $\forall n \ge n_0$, $\sum_{k=1}^{r_n} x_{n,k}^2 \le \varepsilon^2$. Hence, for any k, $x_{n,k}^2 \le \sum_{k=1}^{r_n} x_{n,k}^2 \le \varepsilon^2$ and hence $|x_{n,k}| \le \varepsilon$ which implies $a_n \le \varepsilon$. □

Note that when the $x_{k,n}$ are non-negative, condition (i) already implies that the sum $\sum_{k=1}^{r_n} |x_{n,k}|$ converges as $n \to \infty$. Alternative versions of A.12 are as followed.

(a) Suppose (ii) $\sum_{k=1}^{r_n} x_{n,k}^2 \to 0$ as $n \to \infty$. Then, as $n \to \infty$ we have

$$\prod_{k=1}^{r_n} (1 + x_{n,k}) \to e^x \text{ if and only if } \sum_{k=1}^{r_n} x_{n,k} \to x. \tag{64}$$

Proof. We already know from A.12 that the RHS of (64) implies the LHS. Also, condition (ii) allows the use of A.11 which implies

$$\prod_{k=1}^{r_n} (1 + x_{n,k}) \le e^{\sum_{k=1}^{r_n} x_{n,k}} \le e^{\sum_{k=1}^{r_n} x_{n,k}^2} \prod_{k=1}^{r_n} (1 + x_{n,k}). \tag{63b}$$

□

(b) Suppose the $x_{n,k}$ are nonnegative and (iii) $a_n \to 0$ as $n \to \infty$. Then, as $n \to \infty$ we have

$$\prod_{k=1}^{r_n} (1 - x_{n,k}) \to e^{-x} \text{ if and only if } \sum_{k=1}^{r_n} x_{n,k} \to x. \tag{65}$$

Proof. We already know from A.12 that the RHS of (64) implies the LHS. Also, condition (iii) allows the use of A.11 which implies

$$\prod_{k=1}^{r_n} (1 - x_{n,k}) \leq e^{-\sum_{k=1}^{r_n} x_{n,k}} \leq e^{\sum_{k=1}^{r_n} x_{n,k}^2} \prod_{k=1}^{r_n} (1 - x_{n,k}). \tag{63c}$$

Furthermore, by (62), we have

$$\sum_{k=1}^{r_n} x_{n,k}^2 \leq a_n \left(-\sum_{k=1}^{r_n} \ln (1 - x_{n,k}) \right) \to 0 \times x = 0.$$

□

A.13. Let α_i and β_i be complex numbers with $|\alpha_i| \leq 1$ and $|\beta_i| \leq 1$. Then,

$$\left| \prod_{i=1}^{m} \alpha_i - \prod_{i=1}^{m} \beta_i \right| \leq \sum_{i=1}^{m} |\alpha_i - \beta_i|.$$

In particular, $|\alpha^m - \beta^m| \leq m |\alpha - \beta|$.

A.14. Suppose $\lim_{n\to\infty} a_n = a$. Then $\lim_{n\to\infty} \left(1 - \frac{a_n}{n}\right)^n = e^{-a}$ [9, p 584].

Proof. Use (A.12) with $r_n = n$, $x_{n,k} = -\frac{a_n}{n}$. Then, $\sum_{k=1}^{n} x_{n,k} = -a_n \to -a$ and $\sum_{k=1}^{n} x_{n,k}^2 = a_n^2 \frac{1}{n} \to a \cdot 0 = 0$. □

Alternatively, from L'Hôpital's rule, $\lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^n = e^{-a}$. (See also [20, Theorem 3.31, p 64]) This gives a direct proof for the case when $a > 0$. For n large enough, note that both $\left|1 - \frac{a_n}{n}\right|$ and $\left|1 - \frac{a}{n}\right|$ are ≤ 1 where we need $a > 0$ here. Applying (A.13), we get $\left|\left(1 - \frac{a_n}{n}\right)^n - \left(1 - \frac{a}{n}\right)^n\right| \leq |a_n - a| \to 0$.

For $a < 0$, we use the fact that, for $b_n \to b > 0$, (1) $\left(\left(1+\frac{b}{n}\right)^{-1}\right)^n = \left(\left(1+\frac{b}{n}\right)^n\right)^{-1} \to e^{-b}$ and (2) for n large enough, both $\left|\left(1+\frac{b}{n}\right)^{-1}\right|$ and $\left|\left(1+\frac{b_n}{n}\right)^{-1}\right|$ are ≤ 1 and hence

$$\left|\left(\left(1+\frac{b_n}{n}\right)^{-1}\right)^n - \left(\left(1+\frac{b}{n}\right)^{-1}\right)^n\right| \leq \frac{|b_n - b|}{\left(1+\frac{b_n}{n}\right)\left(1+\frac{b}{n}\right)} \to 0.$$

References

[1] Richard A. Brualdi. *Introductory Combinatorics*. Prentice Hall, 5 edition, January 2009. 4.1, 4.2, 4.3, 4.8

[2] F. N. David. *Games, Gods and Gambling: A History of Probability and Statistical Ideas*. Dover Publications, unabridged edition, February 1998. 4.27

[3] Rick Durrett. *Elementary Probability for Applications*. Cambridge University Press, 1 edition, July 2009. 4.25, 8.27

[4] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. John Wiley & Sons, 1971. 1.10

[5] William Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*. Wiley, 3 edition, 1968. 4.13

[6] Terrence L. Fine. *Probability and Probabilistic Reasoning for Electrical Engineering*. Prentice Hall, 2005. 3.1, 8.27, 18.1

[7] Martin Gardner. *Entertaining mathematical puzzles*. Dover, 1986. 1.13

[8] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Processes*. Oxford University Press, 3 edition, 2001. 8.66

[9] John A. Gubner. *Probability and Random Processes for Electrical and Computer Engineers*. Cambridge University Press, 2006. 2.5, 4.11, 6.21, 6.32, 8.66, 8.82, 5, 6, 9.33, 7, 10.20, 10.22, 1, 3, 14.16, 14.21, A.8, A.14

[10] Samuel Karlin and Howard E. Taylor. *A First Course in Stochastic Processes*. Academic Press, 1975. 7.2

[11] A.N. Kolmogorov. *The Foundations of Probability*. 1933. 5.1

[12] B. P. Lathi. *Modern Digital and Analog Communication Systems*. Oxford University Press, 1998. 1.3, 1.15, 26

[13] Leonard Mlodinow. *The Drunkard's Walk: How Randomness Rules Our Lives*. Pantheon; 8th Printing edition, 2008. 1.8, 1.9, 4.29, 4.30, 6.23, 6.41, 8.56, 8.57, 14.2

[14] Frederick Mosteller. *Fifty Challenging Problems in Probability with Solutions*. Dover Publications, 1987. 1

[15] Paul J. Nahin. *Digital Dice: Computational Solutions to Practical Probability Problems*. Princeton University Press, March 2008. 10.71

[16] Ha H. Nguyen and Ed Shwedyk. *A First Course in Digital Communications*. Cambridge University Press, 1 edition, June 2009. 10, 11, 12, 14, 18, 21, 25, 27

[17] Peter Olofsson. *Probabilities: The Little Numbers That Rule Our Lives*. Wiley, 2006. 1.1, 1.2, 1.7, 1.13, 1.14, 1.16, 3, 4.6, 4.7, 5.10, 5.13, 6.20, 6.22, 6.34, 6.41

[18] Nabendu Pal, Chun Jin, and Wooi K. Lim. *Handbook of Exponential and Related Distributions for Engineers and Scientists*. Chapman & Hall/CRC, 2005. 10.56

[19] Athanasios Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Companies, 1991. 6.42

[20] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976. A.14

[21] Mark F. Schilling. The longest run of heads. *The College Mathematics Journal*, 21(3):196–207, 1990. 4.35, 4.36

[22] Henk Tijms. *Understanding Probability: Chance Rules in Everyday Life*. Cambridge University Press, 2 edition, August 2007. 1.8, 1.20, 3.6, 4.26, 4.31, 4.37, 5, 6.10, 3, 6.42, 8.22, 1, 8.40, 8.41, 8.42, 8.66, 8.74, 8.80, 9.33, 9.52, 9.53, 10.2, 10.24, 14.9, 3

[23] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004. 8.66, 1, 9.33, 18.1

[24] John M. Wozencraft and Irwin Mark Jacobs. *Principles of Communication Engineering*. Waveland Press, June 1990. 1.4

[25] Roy D. Yates and David J. Goodman. *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*. Wiley, 2 edition, May 2004. 11.2, 11.15

[26] Rodger E. Ziemer and William H. Tranter. *Principles of Communications*. John Wiley & Sons Ltd, 2010. 6.27, 1