

ECS 315: Probability and Random Processes

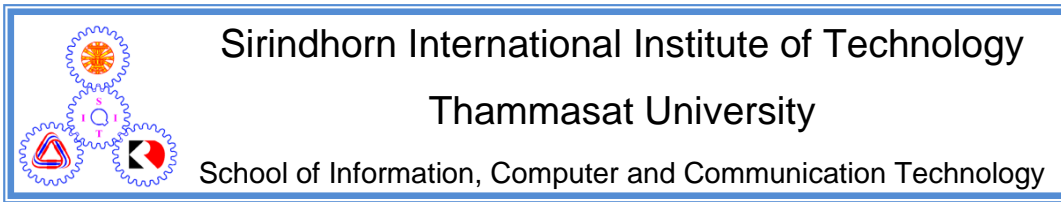
Asst. Prof. Dr. Prapun Suksompong
prapun@siit.tu.ac.th

September 26, 2014

This note covers fundamental concepts in probability and a brief touch on random processes for undergraduate students in electronics and communication engineering (EC). It is designed for a one-semester course at Sirindhorn International Institute of Technology (SIIT).

Contents

1	Probability and You	3
1.1	Randomness	3
1.2	Background on Some Frequently Used Examples	5
1.3	A Glimpse at Probability Theory	7
2	Review of Set Theory	11
3	Classical Probability	16
4	Enumeration / Combinatorics / Counting	19
4.1	Four Principles	19
4.2	Four Kinds of Counting Problems	25
4.3	Binomial Theorem and Multinomial Theorem	35
4.4	Famous Example: Galileo and the Duke of Tuscany	37
4.5	Application: Success Runs	38
5	Probability Foundations	43
6	Event-based Independence and Conditional Probability	52
6.1	Event-based Conditional Probability	52
6.2	Event-based Independence	65
6.3	Bernoulli Trials	71
7	Random variables	77
8	Discrete Random Variables	83
8.1	PMF: Probability Mass Function	84
8.2	CDF: Cumulative Distribution Function	87
8.3	Families of Discrete Random Variables	90
8.4	Some Remarks	101



ECS315 2014/1 Part I.1 Dr.Prapun

1 Probability and You

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?". [15, 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. [15, 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; see also Ex. 1.7.)
- (c) our unwillingness to carry out exact analysis because it is not worth the trouble

Example 1.4. Communication Systems [18]: 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 uncertainty 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. [15, 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. [17, 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 “**heads**” (**H**) and “**tails**” (**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, . . . , 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.

1.3 A Glimpse at Probability Theory

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*.” [15, 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 3.5.

[15, 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 dice, 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.
- Any¹ event is a subset of Ω .
 - Intuitively, an event is a statement about the outcome(s) of an experiment.

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 subsets of Ω .

Example 1.16. The statement “when a coin is tossed, the probability to get heads is $1/2$ (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.

¹For 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.

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. [15, 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 a theorem called the **law of large numbers** (LLN) states that, in the long run, the event A will happen approximately $100p\%$ 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 $P(A)$.

Example 1.18. Return to the coin tossing experiment in Ex. 1.16:

Definition 1.19. 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

is called the **relative frequency** of the event A in these n trials.

1.20. The long-run frequency interpretation mentioned in 1.17 can be restated as

$$P(A) \text{ “=” } \lim_{n \rightarrow \infty} \frac{N(A, n)}{n}.$$

1.21. Another interpretation: The probability of an outcome can be interpreted as our subjective probability, or degree of belief, that the outcome will occur. Different individuals will no doubt assign different probabilities to the same outcomes.

1.22. 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. (See Ex. 3.4.)

[17, p 4]

2 Review of Set Theory

Example 2.1. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$

2.2. Venn diagram is very useful in set theory. It is often used to portray relationships between sets. Many identities can be read out simply by examining Venn diagrams.

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

Definition 2.4. 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 .

2.5. Basic Set Identities:

- Idempotence: $(A^c)^c = A$
- Commutativity (symmetry):

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

- 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**

- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$

2.6. 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 [8, p. 9] if and only if $A_i \cap A_j = \emptyset$ when $i \neq j$.

Example 2.7. Sets A , B , and C are pairwise disjoint if

2.8. 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 2.9. Given a set S , a collection $\Pi = (A_\alpha : \alpha \in I)$ of subsets² 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.

Example 2.10 (Slide:maps).

Example 2.11. Let E be the set of students taking ECS315

Definition 2.12. 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.

- A *finite* set is a set that has a finite number of elements.
- A set that is not finite is called *infinite*.
- Countable sets:

²In this case, the subsets are indexed or labeled by α taking values in an index or label set I

- Empty set and finite sets are automatically countable.
 - 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, \dots
- 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 .

2.13. We can categorize sets according to their cardinality:

Example 2.14. Examples of countably infinite sets:

- the set $\mathbb{N} = \{1, 2, 3, \dots\}$ 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,

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

Example 2.15. Example of uncountable sets³:

- $\mathbb{R} = (-\infty, \infty)$
- interval $[0, 1]$
- interval $(0, 1]$
- $(2, 3) \cup [5, 7)$

Definition 2.16. 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.

2.17. 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.

³We 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. [15, 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

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

[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 dice, 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.
- In many problems, when the finite sample space does not contain equally likely outcomes, we can redefine the sample space to make the outcome equipossible.

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 verified⁴ the properties below.

- $P(A) \geq 0$
- $P(\Omega) = 1$
- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 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 $\Omega = \{\omega_1, \dots, \omega_n\}$ 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

⁴Because we will not rely on Definition 3.1 beyond this section, we will not worry about how to prove these properties. In Section 5, we will prove the same properties in a more general setting.

Example 3.4 (Slides). When rolling two dice, there are 36 (equiprobable) possibilities.

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

Though one of the finest minds of his age, Leibniz was not immune to blunders: he thought it just as easy to throw 12 with a pair of dice as to throw 11. The truth is...

$$\begin{aligned} P[\text{sum of the two dice} = 11] &= \\ P[\text{sum of the two dice} = 12] &= \end{aligned}$$

Definition 3.5. 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.

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

Example 3.7. *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.”

ECS315 2014/1 Part I.2 Dr.Prapun

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 Four Principles

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

- When there are m cases such that the i th case has n_i options, for $i = 1, \dots, m$, and no two of the cases have any options in common, the total number of options is $n_1 + n_2 + \dots + n_m$.
- In set-theoretic terms, suppose that a finite set S can be partitioned⁵ into (pairwise disjoint parts) S_1, S_2, \dots, S_m . Then,

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

⁵The 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 S into parts each containing only one element, then applying the

In words, if you can count the number of elements in all of the parts of a partition of S , then $|S|$ is simply the sum of the number of elements in all the parts.

Example 4.2. We may find the number of people living in a country by adding up the number from each province/state.

Example 4.3. [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 4.4. [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.

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

4.6. Tree diagrams: When a set can be constructed in several steps or stages, we can represent each of the n_1 ways of completing the first step as a branch of a tree. Each of the ways of completing the second step can be represented as n_2 branches starting from

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]

the ends of the original branches, and so forth. The size of the set then equals the number of branches in the last level of the tree, and this quantity equals

$$n_1 \times n_2 \times \cdots$$

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

- When a procedure/operation 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 $i - 1$ ($i = 2, \dots, m$) there are n_i options for step i (for each way of completing step $i - 1$), the number of ways of performing the procedure is $n_1 n_2 \cdots n_m$.
- In set-theoretic terms, if sets S_1, S_2, \dots, S_m are finite, then $|S_1 \times S_2 \times \cdots \times S_m| = |S_1| \times |S_2| \times \cdots \times |S_m|$.
- For m finite sets A_1, A_2, \dots, A_m , there are $|A_1| \times |A_2| \times \cdots \times |A_m|$ m -tuples of the form (a_1, a_2, \dots, a_m) where each $a_i \in A_i$.

Example 4.8. 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 $2 \times 2 \times 2 = 8$ 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. [15, p 33]

Example 4.9 (Slides). 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. [15, p 34]

Example 4.10. There are 2^n binary strings/sequences of length n .

Example 4.11. For a finite set A , the cardinality of its power set 2^A is

$$|2^A| = 2^{|A|}.$$

Example 4.12. (Slides) Jack is so busy that he's always throwing his socks into his top drawer without pairing them. One morning Jack oversleeps. In his haste to get ready for school, (and still a bit sleepy), he reaches into his drawer and pulls out 2 socks. Jack knows that 4 blue socks, 3 green socks, and 2 tan socks are in his drawer.

- (a) What are Jack's chances that he pulls out 2 blue socks to match his blue slacks?

- (b) What are the chances that he pulls out a pair of matching socks?

Example 4.13. [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

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

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

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

- When S is the same as Ω , we have $|A| = |\Omega| - |A^c|$
- 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 .

Example 4.15. *Chevalier de Mere's Scandal of Arithmetic:*

Which is more likely, obtaining at least one six in 4 tosses of a fair dice (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 money⁶ 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 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). [17, p 3]

4.16. 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.

⁶Even 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.

4.2 Four Kinds of Counting Problems

4.17. Choosing objects from a collection is called **sampling**, and the chosen objects are known as a **sample**. The four kinds of counting problems are [8, 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 4.33 for “bars and stars” argument.

Many counting problems can be simplified/solved by realizing that they are equivalent to one of these counting problems.

4.18. Ordered Sampling: Given a set of n distinct items/objects, select a distinct **ordered**⁷ sequence (word) of length r drawn from this set.

- (a) **Ordered sampling with replacement:** $\mu_{n,r} = n^r$
 - Ordered sampling of r out of n items with replacement.
 - The “with replacement” part means “an object can be chosen repeatedly.”
 - Example: From a deck of n cards, we draw r cards with replacement; i.e., we draw a 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? [8, Ex. 1.30]

⁷Different sequences are distinguished by the order in which we choose objects.

(b) *Ordered sampling without replacement:*

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

- Ordered sampling of $r \leq n$ out of n items without replacement.
- The “without replacement” means “once we choose an object, we remove that object from the collection and we cannot choose it again.”
- In Excel, use PERMUT(n, r).
- Sometimes referred to as “the number of possible r -permutations of n distinguishable objects”
- Example: The number of sequences⁸ of size r drawn from an alphabet of size n without replacement.

$(3)_2 = 3 \times 2 = 6$ is the number of sequences 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?

⁸Elements in a sequence are ordered.

- For integers r, n such that $r > n$, we have $(n)_r = 0$.
- We define $(n)_0 = 1$. (This makes sense because we usually take the empty product to be 1.)
- $(n)_1 = n$
- $(n)_r = (n - (r - 1))(n)_{r-1}$. For example, $(7)_5 = (7 - 4)(7)_4$.
- $(1)_r = \begin{cases} 1, & \text{if } r = 1 \\ 0, & \text{if } r > 1 \end{cases}$
- Extended definition: The definition in product form

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

can be extended to *any real number* n and a non-negative integer r .

Example 4.19. (Slides) The Seven Card Hustle: Take five red cards and two black cards from a pack. Ask your friend to shuffle them and then, without looking at the faces, lay them out in a row. Bet that they can't turn over three red cards. The probability that they CAN do it is

Definition 4.20. 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 .

(a) $0! = 1! = 1$

(b) $n! = n(n - 1)!$

(c) $n! = \int_0^{\infty} e^{-t} t^n dt$

(d) Computation:

- (i) **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.
 - (ii) Google's web search box built-in calculator: Use `n!`
- (e) Approximation: Stirling's Formula [5, p. 52]:

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

In some references, the sign \approx is replaced by \sim to emphasize that the ratio of the two sides converges to unity as $n \rightarrow \infty$.

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

- 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 create 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 4.22. 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

Example 4.23. (Slides) Finger-Smudge on Touch-Screen Devices

Example 4.24. (Slides) *Probability of coincidence birthday*: Probability that there is at least two people who have the same birthday⁹ in a group of r persons:

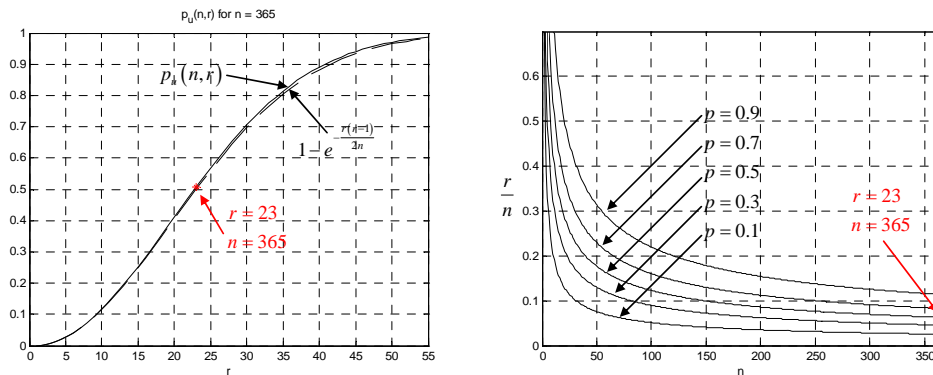


Figure 1: $p_u(n, r)$: The probability of the event that at least one element appears twice in random sample of size r with replacement is taken from a population of n elements.

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

⁹We ignore February 29 which only comes in leap years.

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 year¹⁰) is about 0.5.

- At first glance it is surprising that the probability of 2 people having the same birthday is so large¹¹, 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]

Example 4.26. Another variant of the birthday coincidence paradox: 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.
- 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

¹⁰In 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.

¹¹In other words, it was surprising that the size needed to have 2 people with the same birthday was so small.

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. [17, p 76]

4.27. Now, let’s revisit ordered sampling of r out of n different items without replacement. This is also called the number of possible **r -permutations** of n different items. One way to look at the sampling is to first consider the $n!$ permutations of the n items. Now, use only the first r positions. Because we do not care about the last $n - r$ positions, we will group the permutations by the first r positions. The size of each group will be the number of possible permutations of the $n - r$ items that has not already been used in the first r positions. So, each group will contain $(n - r)!$ members. By the division principle, the number of groups is $n!/(n - r)!$.

4.28. The number of permutations of $n = n_1 + n_2 + \dots + n_r$ objects of which n_1 are of one type, n_2 are of one type, n_2 are of second type, \dots , and n_r are of an r th type is

$$\frac{n!}{n_1!n_2!\dots n_r!}$$

Example 4.29. The number of permutations of AABC

Example 4.30. Bar Codes: A part is labeled by printing with four thick lines, three medium lines, and two thin lines. If each ordering of the nine lines represents a different label, how many different labels can be generated by using this scheme?

4.31. 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 distinct 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 **r -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) Excel: `combin(n,r)`

(iii) Mathcad: `combin(n,r)`

(iv) Maple: $\binom{n}{r}$

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

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

Example 4.32. 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)

4.33. The bars and stars argument:

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

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

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

- There are $\binom{n+r-1}{r} = \binom{n+r-1}{n-1}$ distinct n -tuples (x_1, x_2, \dots, x_n) of nonnegative integers such that $x_1 + x_2 + \dots + x_n = r$. 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

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

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

4.34. 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 $\{a_1, a_2, \dots, a_n\}$. Let x_i be the number of a_i in the drawn sequence. Because we sample r times, we know that, for every sample, $x_1 + x_2 + \dots + x_n = r$ where the x_i are non-negative integers. Hence, there are $\binom{n+r-1}{r}$ possible unordered samples with replacement.

4.3 Binomial Theorem and Multinomial Theorem

4.35. 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,

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

(Slide) 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 $r = 0, 1, 2, \dots, n$. 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 x s from the n sums (so that the other $n - r$ terms in the product are y s). 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:

$$\sum_{r=0}^n \binom{n}{r} = 2^n. \quad (4)$$

4.36. *Multinomial Counting:* The *multinomial coefficient*

$$\binom{n}{n_1 \ n_2 \ \cdots \ n_r}$$

is defined as

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

We have seen this before in (4.28). 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.

4.37. *Multinomial Theorem:*

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

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

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

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

ECS315 2014/1 Part I.3 Dr.Prapun

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

4.4 Famous Example: Galileo and the Duke of Tuscany

Example 4.39. 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 i th dice and S_n be the sum $X_1 + X_2 + \cdots + X_n$.

- $P[S_3 = 9] = P[S_3 = 12] = \frac{25}{6^3} < \frac{27}{6^3} = P[S_3 = 10] = P[S_3 = 11]$. Note that the difference between the two probabilities is only $\frac{1}{108}$.
- The range of S_n is from n to $6n$. So, there are $6n - n + 1 = 5n + 1$ possible values.
- The pmf of S_n is symmetric around its expected value at $\frac{n+6n}{2} = \frac{7n}{2}$.
 - $P[S_n = m] = P[S_n = 7n - m]$.

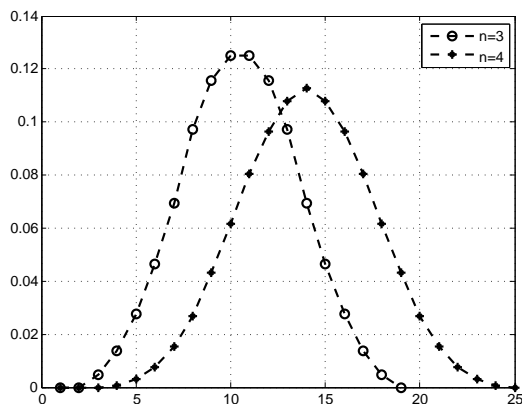


Figure 2: pmf of S_n for $n = 3$ and $n = 4$.

4.5 Application: Success Runs

Example 4.40. 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 Tversky and Thomas Gilovich examined¹² 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]

4.41. 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 4.42. 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

¹²“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 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 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? [17, Ex. 7.1 p 42–43]

The answer is later provided in Example 4.48.

Definition 4.43. A **run** is a sequence of more than one consecutive identical outcomes, also known as a **clump**.

Definition 4.44. 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 4.45. 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

4.46. 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 $k \in \{1, 2, \dots, x + 1\}$, the favorable sequences are in the form

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

where, to keep the sequences in $\mathcal{A}_n(x)$, the last $n - k$ positions¹³ 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

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

[16]. 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);
```

4.47. Similar technique can be used to construct $\mathcal{B}_n(x)$ 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

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

[16].

¹³Strictly 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 4.48. Continue from Example 4.42. 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! [17, Ex. 7.1 p 42–43]

ECS315 2014/1 Part II Dr.Prapun

5 Probability Foundations

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. [17, p 223]

We will try to avoid several technical details¹⁴ ¹⁵ in this class. Therefore, the definition given below is not the “complete” definition. Some parts are modified or omitted to make the definition easier to understand.

¹⁴To study formal definition of probability, we start with the *probability space* (Ω, \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 (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} .

¹⁵The 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**¹⁶ is a real-valued set function¹⁷ that satisfies

P1 Nonnegativity:

$$P(A) \geq 0.$$

P2 Unit normalization:

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

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

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

- The number $P(A)$ is called the **probability** of the event A
- 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 $P(A) = 1$.

From the three axioms¹⁸, we can derive many more properties of probability measure. These properties are useful for calculating probabilities.

¹⁶Technically, probability measure is defined on a σ -algebra \mathcal{A} of Ω . The *triple* (Ω, \mathcal{A}, P) is called a **probability measure space**, or simply a **probability space**

¹⁷A real-valued set function is a function that maps sets to real numbers.

¹⁸Remark: The axioms do not determine probabilities; the probabilities are assigned based on our knowledge of the system under study. (For example, one approach is to base probability assignments on the simple concept of equally likely outcomes.) The axioms enable us to easily calculate the probabilities of some events from knowledge of the probabilities of other events.

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

5.3. Finite additivity¹⁹: If A_1, \dots, A_n are disjoint events, then

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

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

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

¹⁹It is not possible to go backwards and use finite additivity to derive countable additivity (P3).

5.4. The probability of a finite or countable event equals the sum of the probabilities of the outcomes in the event.

(a) In particular, if A is countable, e.g. $A = \{a_1, a_2, \dots\}$, then

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

(b) Similarly, if A is finite, e.g. $A = \{a_1, a_2, \dots, a_{|A|}\}$, then

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

- This greatly simplifies²⁰ construction of probability measure.

Remark: Note again that the set A under consideration here is finite or countably infinite. You can not apply the properties above to uncountable set.²¹

²⁰ Recall that a probability measure P is a set function that assigns number (probability) to all set (event) in \mathcal{A} . When Ω is countable (finite or countably infinite), we may let $\mathcal{A} = 2^\Omega =$ the power set of the sample space. In other words, in this situation, it is possible to assign probability value to all subsets of 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, 5.4 implies that we only need to define P for all the singletons (when Ω is countable).

²¹In Section ??, we will start talking about (absolutely) continuous random variables. In such setting, we have $P(\{\alpha\}) = 0$ for any α . However, it is possible to have an uncountable set A with $P(A) > 0$. This does not contradict the properties that we discussed in 5.4. If A is finite or countably infinite, we can still write

$$P(A) = \sum_{\alpha \in A} P(\{\alpha\}) = \sum_{\alpha \in A} 0 = 0.$$

For event A that is uncountable, the properties in 5.4 are not enough to evaluate $P(A)$.

Example 5.5. A random experiment can result in one of the outcomes $\{a, b, c, d\}$ with probabilities 0.1, 0.3, 0.5, and 0.1, respectively. Let A denote the event $\{a, b\}$, B the event $\{b, c, d\}$, and C the event $\{d\}$.

- $P(A) =$
- $P(B) =$
- $P(C) =$
- $P(A^c) =$
- $P(A \cap B) =$
- $P(A \cap C) =$

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

Example 5.7. 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.

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

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

5.10. $P(A \cap B)$ 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 5.11 (Slides). Experiments by psychologists Kahneman and Tversky.

Example 5.12. 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. T’s 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* [10] 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 discomfiting 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. [15, p 11–12]

5.13. Complement Rule:

$$P(A^c) = 1 - P(A).$$

- “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 [9]

5.14. Probability of a union (not necessarily disjoint):

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

- $P(A \cup B) \leq P(A) + P(B)$.
- Approximation: If $P(A) \gg P(B)$ then we may approximate $P(A \cup B)$ by $P(A)$.

Example 5.15 (Slides). Combining error probabilities from various sources in DNA testing

Example 5.16. 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 5.14:

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

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. [15, p 12]

5.17. Probability of a union of three events:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C) \end{aligned}$$

5.18. Two bounds:

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

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

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

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

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

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

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

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

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

5.20. Connection to classical probability theory: Consider an experiment with **finite** sample space $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ 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 finite²² event A , we can apply 5.4 to get

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

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.

²²In classical probability, the sample space is finite; therefore, any event is also finite.

6 Event-based Independence and Conditional Probability

Example 6.1. Roll a dice...

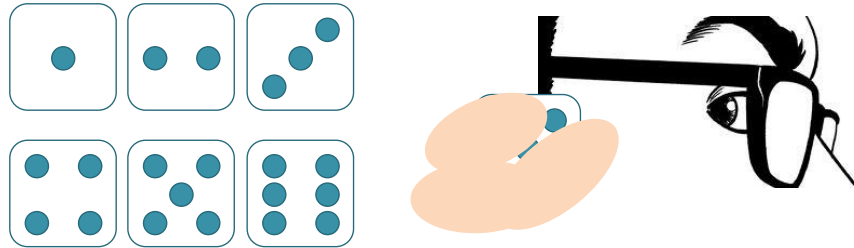


Figure 3: Conditional Probability Example: Sneak Peek

Example 6.2 (Slides). Diagnostic Tests.

6.1 Event-based Conditional Probability

Definition 6.3. *Conditional Probability*: The conditional probability $P(A|B)$ of event A , given that event $B \neq \emptyset$ occurred, is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (6)$$

- Some ways to say²³ or express the conditional probability, $P(A|B)$, are:
 - the “probability of A , given B ”
 - the “probability of A , knowing B ”
 - the “probability of A happening, knowing B has already occurred”

²³Note also that although the symbol $P(A|B)$ itself is practical, its 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.”

- 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.

6.4. 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).

Example 6.5. Back to Example 6.1

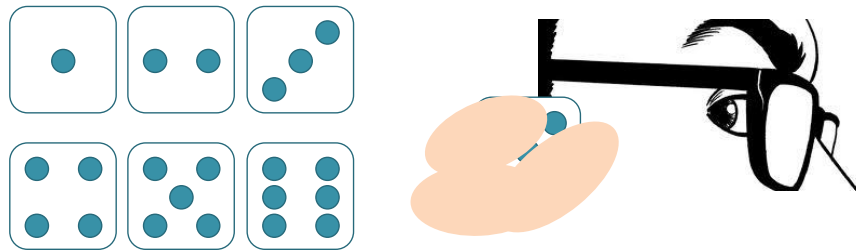


Figure 4: Sneak Peek: A Revisit

Example 6.6. In diagnostic tests Example 6.2, 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 $P(D|T_P) \approx 0.01$.

6.7. “Prelude” to the concept of “independence”:

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

$$P(A|B) = P(A) \tag{7}$$

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.”

We will soon define “independence” in Section 6.2. Property (7) can be regarded as a “practical” definition for independence. However, there are some “technical” issues²⁴ that we need to deal with when we actually define independence.

²⁴Here, the statement assume $P(B) > 0$ because it considers $P(A|B)$. The concept of independence to be defined in Section 6.2 will not rely directly on conditional probability and therefore it will include the case where $P(B) = 0$.

6.8. 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,

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

- In particular, if $A_1 \perp A_2$,

$$P(A_1 \cup A_2 | B) = P(A_1 | B) + P(A_2 | B)$$

6.9. More Properties:

- $P(A|\Omega) = P(A)$
- $P(A^c|B) = 1 - P(A|B)$

- $P(A \cap B|B) = P(A|B)$
- $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$.
- $P(A \cap B) \leq P(A|B)$

6.10. 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.

Exercise 6.11. Someone has rolled a fair dice twice. You know that one of the rolls turned up a face value of six. What is the probability that the other roll turned up a six as well?

Ans: $\frac{1}{11}$ (not $\frac{1}{6}$). [17, Example 8.1, p. 244]

6.12. 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 6.13. 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

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

[17, p 243]

Question: What about the unconditional probability $P(B)$?

Example 6.14. 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.

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

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

$$P(A) = \sum_i P(A|B_i)P(B_i). \quad (8)$$

This is a formula²⁵ for computing the probability of an event that can occur in different ways.

6.16. 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 6.14.

²⁵The tree diagram is useful for helping you understand the process. However, when the number of possible cases is large (many B_i for the partition), drawing the tree diagram may be too time-consuming and therefore you should also learn how to apply the total probability theorem directly without the help of the tree diagram.

Example 6.17. Continue from the “Diagnostic Tests” Example 6.2 and Example 6.6.

$$\begin{aligned} P(T_P) &= P(T_P \cap D) + P(T_P \cap D^c) \\ &= P(T_P | D) P(D) + P(T_P | D^c) P(D^c). \end{aligned}$$

For conciseness, we define

$$p_d = P(D)$$

and

$$p_{TE} = P(T_P | D^c) = P(T_P^c | D).$$

Then,

$$P(T_P) = (1 - p_{TE})p_d + p_{TE}(1 - p_d).$$

6.18. 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 $\{B_1, B_2, \dots\}$ is a partition of Ω , then

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

- 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 6.19. Continue from the “Disease Testing” Examples 6.2, 6.6, and 6.17:

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

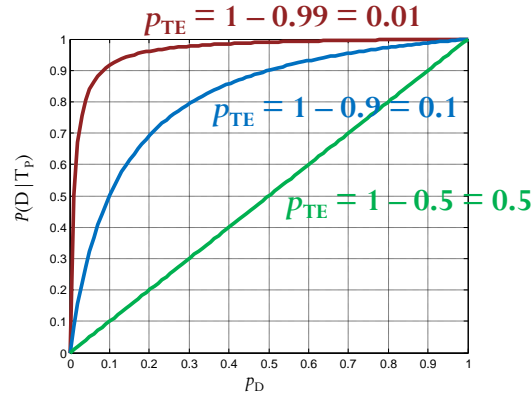


Figure 5: Probability $P(D|T_P)$ that a person will have the disease given that the test result is positive. The conditional probability is evaluated as a function of P_D which tells how common the disease is. The values of test error probability p_{TE} are shown.

Example 6.20. Medical Diagnostic: Because a new medical procedure has been shown to be effective in the early detection of an illness, a medical screening of the population is proposed. The probability that the test correctly identifies someone with the illness as positive is 0.99, and the probability that the test correctly identifies someone without the illness as negative is 0.95. The incidence of the illness in the general population is 0.0001. You take the test, and the result is positive. What is the probability that you have the illness? [14, Ex. 2-37]

Example 6.21. Bayesian networks are used on the Web sites of high-technology manufacturers to allow customers to quickly diagnose problems with products. An oversimplified example is presented here.

A printer manufacturer obtained the following probabilities from a database of test results. Printer failures are associated with three types of problems: hardware, software, and other (such as connectors), with probabilities 0.1, 0.6, and 0.3, respectively. The probability of a printer failure given a hardware problem is 0.9, given a software problem is 0.2, and given any other type of problem is 0.5. If a customer enters the manufacturers Web site to diagnose a printer failure, what is the most likely cause of the problem?

Let the events H , S , and O denote a hardware, software, or other problem, respectively, and let F denote a printer failure.

Example 6.22 (Slides). The Murder of Nicole Brown

6.23. In practice, here is how we use the total probability theorem and Bayes' theorem:

Usually, we work with a system, which of course has input and output. There can be many possibilities for inputs and there can be many possibilities for output. Normally, for deterministic system, we may have a specification that tells what would be the output given that a specific input is used. Intuitively, we may think of this as a table of mapping between input and output. For system with

random component(s), when a specific input is used, the output is not unique. This means we need conditional probability to describe the output (given an input). Of course, this conditional probability can be different for different inputs.

We will assume that there are many cases that the input can happen. The event that the i th case happens is denoted by B_i . We assume that we consider all possible cases. Therefore, the union of these B_i will automatically be Ω . If we also define the cases so that they do not overlap, then the B_i partitions Ω .

Similarly, there are many cases that the output can happen. The event that the j th case happens is denoted by A_j . We assume that the A_j also partitions Ω .

In this way, the system itself can be described by the conditional probabilities of the form $P(A_j|B_i)$. This replaces the table mentioned above as the specification of the system. Note that even when this information is not available, we can still obtain an approximation of the conditional probability by repeating trials of inputting B_i into the system to find the relative frequency of the output A_j .

Now, when the system is used in actual situation. Different input cases can happen with different probabilities. These are described by the prior probabilities $P(B_i)$. Combining this with the conditional probabilities $P(A_j|B_i)$ above, we can use the total probability theorem to find the probability of occurrence for output and, even more importantly, for someone who cannot directly observe the input, Bayes' theorem can be used to infer the value (or the probability) of the input from the observed output of the system.

In particular, the total probability theorem deals with the calculation of the output probabilities $P(A_j)$:

$$P(A_j) = \sum_i P(A_j \cap B_i) = \sum_i P(A_j|B_i) P(B_i).$$

Bayes' theorem calculates the probability that B_k was the input event when the observer can only observe the output of the system

and the observed value of the output is A_j :

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

Example 6.24. 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. [15, p 22]

6.25. Chain rule of conditional probability [8, p 58]:

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

Example 6.26. 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. [15, p 23–24]

Example 6.27. 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 dice, the probability that you get heads is $1/2$ regardless of the outcome of my dice. Events that are unrelated to each other in this way are called *independent*.

Definition 6.28. Two events A , B are called (statistically²⁶) *independent* if

$$P(A \cap B) = P(A)P(B) \quad (9)$$

- Notation: $A \perp\!\!\!\perp B$
- Read “ A and B are independent” or “ A is independent of B ”
- We call (9) the *multiplication rule* for probabilities.
- If two events are not independent, they are *dependent*. Intuitively, if two events are dependent, the probability of one changes with the knowledge of whether the other has occurred.

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

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B). \quad (10)$$

In other words, “the unconditional and the conditional probabilities are the same”. We can almost use (10) as the definitions for independence. This is what we mentioned in 6.7. However, we use (9) instead because it (1) also works with events whose probabilities are zero and (2) also has clear symmetry in the expression (so that $A \perp\!\!\!\perp B$ and $B \perp\!\!\!\perp A$ can clearly be seen as the same). In fact, in 6.33, we show how (10) can be used to define independence with extra condition that deals with the case when zero probability is involved.

²⁶Sometimes our definition for independence above does not agree with the everyday-language use of the word “independence”. Hence, many authors use the term “statistically independence” to distinguish it from other definitions.

Example 6.30. [19, Ex. 5.4] Which of the following pairs of events are independent?

(a) The card is a club, and the card is black.

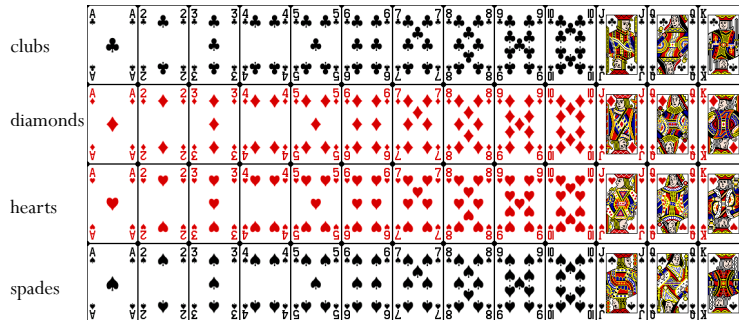


Figure 6: A Deck of Cards

(b) The card is a king, and the card is black.

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

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

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

6.33. 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.

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

6.35. The following four statements are equivalent:

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

Example 6.36. If $P(A|B) = 0.4$, $P(B) = 0.8$, and $P(A) = 0.5$, are the events A and B independent? [14]

6.37. 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. [15, p 12]
- To check whether A and B are disjoint, you only need to look at the sets themselves and see whether they have shared element(s). This can be answered without knowing probabilities.

To check whether A and B are independent, you need to look at the probabilities $P(A)$, $P(B)$, and $P(A \cap B)$.

- 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 .

- 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.

Example 6.38. 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 $B = \{HH, TH\}$. Note that even though the events A and B are not disjoint, they are independent.

Example 6.39 (Slides). *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 sudden infant death syndrome (SIDS), 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? 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 $1/850,000$. 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 [15, 22–23].

Definition 6.40. 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 the first three equations hold, we say that the three events are *pairwise independent*.
- (b) We may use the term “mutually independence” to further emphasize that we have “independence” instead of “pairwise independence”.

Definition 6.41. The events A_1, A_2, \dots, A_n are *independent* if and only if for any subcollection $A_{i_1}, A_{i_2}, \dots, A_{i_k}$,

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \times P(A_{i_2}) \times \dots \times P(A_{i_k}).$$

- Note that part of the requirement is that

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \times P(A_2) \times \dots \times P(A_n).$$

Therefore, if someone tells us that the events A_1, A_2, \dots, A_n are independent, then one of the properties that we can conclude is that

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \times P(A_2) \times \dots \times P(A_n).$$

- Equivalently, this is the same as the requirement that

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

- Note that the case when $j = 1$ automatically holds. The case when $j = 0$ can be regarded as the \emptyset event case, which is also trivially true.

6.42. Four events A, B, C, D are pairwise independent if and only if they satisfy the following six conditions:

$$\begin{aligned} P(A \cap B) &= P(A)P(B), \\ P(A \cap C) &= P(A)P(C), \\ P(A \cap D) &= P(A)P(D), \\ P(B \cap C) &= P(B)P(C), \\ P(B \cap D) &= P(B)P(D), \text{ and} \\ P(C \cap D) &= P(C)P(D). \end{aligned}$$

They are independent if and only if they are pairwise independent (satisfy the six conditions above) and also satisfy the following five more conditions:

$$\begin{aligned} P(B \cap C \cap D) &= P(B)P(C)P(D), \\ P(A \cap C \cap D) &= P(A)P(C)P(D), \\ P(A \cap B \cap D) &= P(A)P(B)P(D), \\ P(A \cap B \cap C) &= P(A)P(B)P(C), \text{ and} \\ P(A \cap B \cap C \cap D) &= P(A)P(B)P(C)P(D). \end{aligned}$$

6.3 Bernoulli Trials

Example 6.43. Consider the following random experiments

- (a) Flip a coin 10 times. We are interested in the number of heads obtained.
- (b) Of all bits transmitted through a digital transmission channel, 10% are received in error. We are interested in the number of bits in error in the next five bits transmitted.
- (c) A multiple-choice test contains 10 questions, each with four choices, and you guess at each question. We are interested in the number of questions answered correctly.

These examples illustrate that a general probability model that includes these experiments as particular cases would be very useful.

Example 6.44. Each of the random experiments in Example 6.43 can be thought of as consisting of a series of repeated, random trials. In all cases, we are interested in the number of trials that meet a specified criterion. The outcome from each trial either meets the criterion or it does not; consequently, each trial can be summarized as resulting in either a success or a failure.

Definition 6.45. 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. We can just as well use “H and T”, “A and B”, or “1 and 0”.

Example 6.46. 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 6.47. (Independent) *Bernoulli Trials* = a Bernoulli trial is repeated many times.

- (a) It is usually assumed that the trials are independent. This implies that the outcome from one trial has no effect on the outcome to be obtained from any other trial.
- (b) Furthermore, it is often reasonable to assume that the probability of a success in each trial is constant.

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

Example 6.49. 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 Bernoulli trial. 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 $(1 - p)^2 p^3$. There are 10 of them. Hence, the total probability is $10(1 - p)^2 p^3$

6.50. 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 6.51. At least one occurrence of a 1-in- n -chance event in n repeated trials:

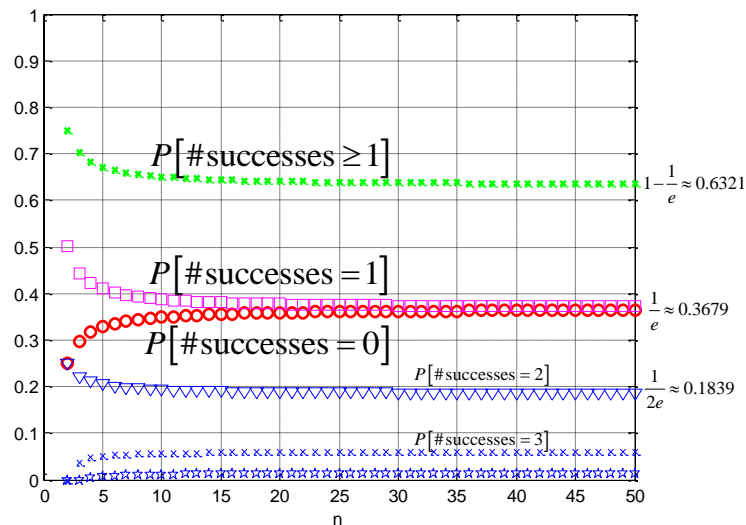


Figure 7: Number of occurrences of 1-in- n -chance event in n repeated Bernoulli trials

Example 6.52. *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***. Example of such code is 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” was transmitted,
 - (ii) if $\#0 > \#1$, then it claims that “0” was transmitted.

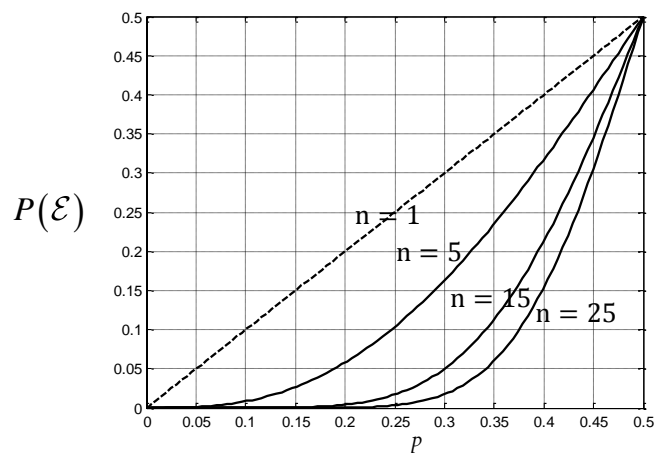
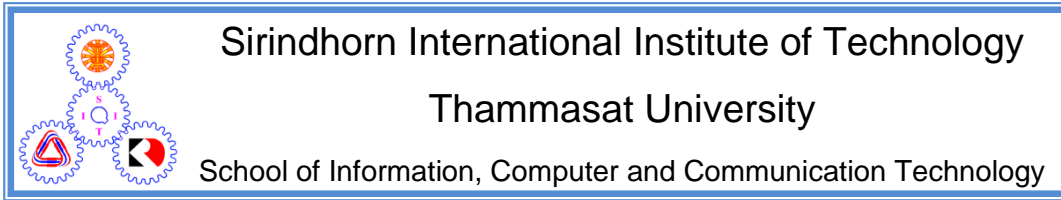


Figure 8: Bit error probability for a simple system that uses repetition code at the transmitter (repeat each bit n times) and majority vote at the receiver. The channel is assumed to be binary symmetric with bit error probability p .

Exercise 6.53 (F2011). Kakashi and Gai are eternal rivals. Kakashi is a little stronger than Gai and hence for each time that they fight, the probability that Kakashi wins is 0.55. In a competition, they fight n times (where n is odd). Assume that the results of the fights are independent. The one who wins more will win the competition.

Suppose $n = 3$, what is the probability that Kakashi wins the competition.



ECS315 2014/1 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.

Example 7.1. Take this course and observe your grades.

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

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

Definition 7.3. A real-valued function $X(\omega)$ defined for all points ω in a sample space Ω is called a *random variable* (r.v. or RV)²⁷.

- So, a random variable is a rule that assigns a numerical value to each possible outcome of a chance experiment.

- Intuitively, a random variable is a variable that takes on its values by chance.
- The convention is to use capital letters such as X , Y , Z to denote random variables.

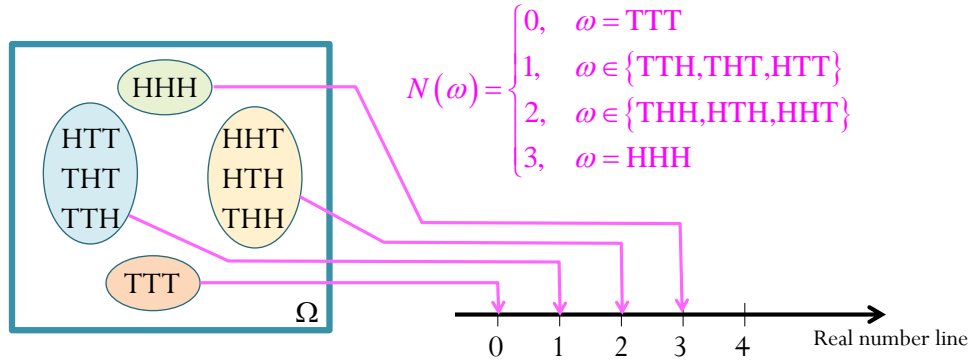
Example 7.4. Roll a fair dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

²⁷The term “random variable” is a misnomer. 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][19, p 254]

- 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 7.5 (Three Coin Tosses). Counting the number of heads in a sequence of three coin tosses.

$$\Omega = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$$



Example 7.6 (Sum of Two Dice). If S is the sum of the dots when rolling one fair dice twice, the random variable S assigns the numerical value $i+j$ to the outcome (i, j) of the chance experiment.

Example 7.7. Continue from Example 7.4,

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

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

Definition 7.8. Events involving random variables:

- $[\text{some condition(s) on } X] = \text{the set of outcomes in } \Omega \text{ such that } X(\omega) \text{ satisfies the conditions.}$
- $[X \in B] = \{ \omega \in \Omega : X(\omega) \in B \}$
- $[a \leq X < b] = [X \in [a, b)] = \{ \omega \in \Omega : a \leq X(\omega) < b \}$
- $[X > a] = \{ \omega \in \Omega : X(\omega) > a \}$
- $[X = x] = \{ \omega \in \Omega : X(\omega) = x \}$
 - We usually use the corresponding lowercase letter²⁸ to denote
 - (a) a possible value (realization) of the random variable
 - (b) the value that the random variable takes on
 - (c) the running values for the random variable

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

Example 7.9. Continue from Examples 7.4 and 7.7,

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

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

7.10. Event of the form “[some condition(s) on X]” or “[some statement(s) about X]” can be written in the form $[X \in B]$ for some appropriate B .

²⁸This is the same as writing $[X = c]$ where c is a constant. Basically, it is a generic notation for $[X = 5]$, $[X = 1.6]$, $[X = \pi]$, etc. We use this when

- (a) we don't want to specify the constant in the expression yet or
- (b) we want to say that the statement/equation/property containing it is valid for any value of c .

It turns out that, later on, we will have to deal with many random variables and hence it is convenient to have the name of the constant c match the name of the corresponding random variable. So, we talk about the events $[X = x]$, $[Y = y]$, and $[Z = z]$ instead of having to find new name for the constant corresponding to each one of them, say, $[X = c]$, $[Y = d]$, and $[Z = h]$.

You may think we can use constants c_1, c_2, \dots . However, we also will have to deal with random variables $X_1, X_2, \dots, Y_1, Y_2, \dots, Z_1, Z_2, \dots$. So, again, will have to come up with new names for a lot of constants.

Example 7.11. Express each event below in the form $[X \in B]$.

(a) $[5 \leq X < 8]$

(b) $[|X| < 3]$

(c) $[X > 2]$

(d) $[X = 1]$

Definition 7.12. We also have another notation for $P[X \in B]$:

$$P^X(B) \equiv P[X \in B].$$

Observe that this function P^X is a set function. It maps subsets of real numbers into their probability values. Technically, we call this function the **law** or **distribution** of the random variable X . However, later on, we shall see that there are many functions that are also referred to as the “distribution” of X as well. They are all equivalent in the sense that they (almost surely) give the same information about probability concerning X .

Definition 7.13. To avoid double use of brackets (round brackets over square brackets), we write $P[X \in B]$ when we mean $P([X \in B])$. 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 7.14. In Example 7.5 (Three Coin Tosses), if the coin is fair, then

$$P[N < 2] =$$

7.15. 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 7.16. 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 .
- Practically, we define a support of a random variable X to be the set of all the “possible” values of X .²⁹
- For any random variable, the set \mathbb{R} of all real numbers is always a support; however, it is not that useful because it does not further limit the possible values of the random variable.
- Recall that a support of a probability measure P is any set $A \subset \Omega$ such that $P(A) = 1$.

Definition 7.17. The **probability distribution** is a description of the probabilities associated with the random variable.

7.18. There are three types of random variables. The first type, which will be discussed in Section 8, is called **discrete random variable**. To tell whether a random variable is discrete, one simple way is to consider the “possible” values of the random variable. If it is limited to only a finite or countably infinite number of possibilities, then it is discrete. We will later discuss **continuous random variables** whose possible values can be anywhere in some intervals of real numbers.

²⁹Later on, you will see that 1) a default support of a discrete random variable is the set of values where the pmf is strictly positive and 2) a default support of a continuous random variable is the set of values where the pdf is strictly positive.

8 Discrete Random Variables

Intuitively, to tell whether a random variable is discrete, we simply consider the possible values of the random variable. If the random variable is limited to only a finite or countably infinite number of possibilities, then it is discrete.

Example 8.1. Voice Lines: A voice communication system for a business contains 48 external lines. At a particular time, the system is observed, and some of the lines are being used. Let the random variable X denote the number of lines in use. Then, X can assume any of the integer values 0 through 48. [14, Ex 3-1]

Definition 8.2. A random variable X is said to be a *discrete random variable* if there exists a countable number of distinct real numbers x_k such that

$$\sum_k P[X = x_k] = 1. \quad (11)$$

In other words, X is a discrete random variable if and only if X has a countable support.

Example 8.3. For the random variable N in Example 7.5 (Three Coin Tosses),

For the random variable S in Example 7.6 (Sum of Two Dice),

8.4. Although the support S_X of a random variable X is defined as any set S such that $P[X \in S] = 1$. For discrete random variable, S_X is usually set to be $\{x : p_X(x) > 0\}$, the set of all “possible values” of X .

Definition 8.5. Important Special Case: An *integer-valued random variable* is a discrete random variable whose x_k in (11) above are all integers.

8.6. Recall, from 7.17, that the *probability distribution* of a random variable X is a description of the probabilities associated with X .

For a discrete random variable, the distribution is often characterized by just a list of the possible values (x_1, x_2, x_3, \dots) along with the probability of each:

$$(P[X = x_1], P[X = x_2], P[X = x_3], \dots, \text{ respectively}).$$

In some cases, it is convenient to express the probability in terms of a formula. This is especially useful when dealing with a random variable that has an unbounded number of outcomes. It would be tedious to list all the possible values and the corresponding probabilities.

8.1 PMF: Probability Mass Function

Definition 8.7. When X is a discrete random variable satisfying (11), we define its *probability mass function* (pmf) by³⁰

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

- 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 (x) of a pmf ranges over all real numbers. Hence, the pmf is defined for x that is not among the x_k in (11). In such case, the pmf is simply 0. This is usually expressed as “ $p_X(x) = 0$, otherwise” when we specify a pmf for a particular random variable.

³⁰Many references (including [14] and MATLAB) does not distinguish the pmf from another function called probability density function (pdf). These references use the function $f_X(x)$ to represent both pmf and pdf. 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 RV).

Example 8.8. Continue from Example 7.5. N is the number of heads in a sequence of three coin tosses.

8.9. Graphical Description of the Probability Distribution: Traditionally, we 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.

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

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

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

When you are asked to verify that a function is a pmf, check these two properties.

8.11. Finding probability from pmf: 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,

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

8.12. Steps to find probability of the form P [some condition(s) on X] when the pmf $p_X(x)$ is known.

- (a) Find the support of X .
- (b) Consider only the x inside the support. Find all values of x that satisfies the condition(s).
- (c) Evaluate the pmf at x found in the previous step.
- (d) Add the pmf values from the previous step.

Example 8.13. 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) $P[X > 3]$

8.14. 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, \dots such that $\sum_k p(x_k) = 1$ and $p(x) = 0$ for other x

is a pmf of some discrete random variable.

8.2 CDF: Cumulative Distribution Function

Definition 8.15. The (*cumulative*) *distribution function* (*cdf*) of a random variable X is the function $F_X(x)$ defined by

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

- The argument (x) of a cdf ranges over all real numbers.
- 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 .

8.16. From pmf to cdf: In general, for any discrete random variable with possible values x_1, x_2, \dots , the cdf of X is given by

$$F_X(x) = P[X \leq x] = \sum_{x_k \leq x} p_X(x_k).$$

Example 8.17. Continue from Examples 7.5, 7.14, and 8.8 where N is defined as the number of heads in a sequence of three coin tosses. We have

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

(a) $F_N(0)$

(b) $F_N(1.5)$

(c) Sketch of cdf

8.18. Facts:

- For any discrete r.v. X , F_X is a right-continuous, *staircase* function of x with jumps at a countable set of points x_k .
- When you are given the cdf of a discrete random variable, you can derive its pmf from the locations and sizes of the jumps. If a jump happens at $x = c$, then $p_X(c)$ is the same as the amount of jump at c . At the location x where there is no jump, $p_X(x) = 0$.

Example 8.19. Consider a discrete random variable X whose cdf $F_X(x)$ is shown in Figure 9.

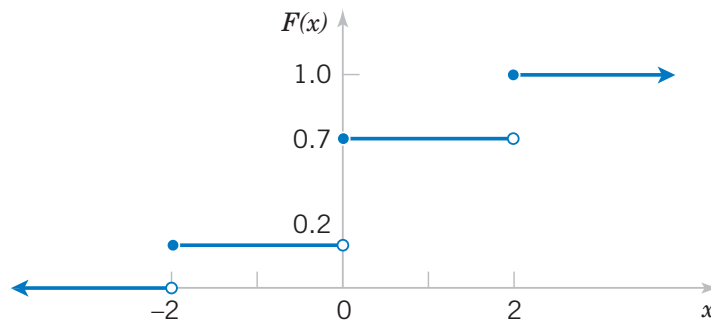


Figure 9: CDF for Example 8.19

Determine the pmf $p_X(x)$.

8.20. Characterizing³¹ properties of cdf:

CDF1 F_X is non-decreasing (monotone increasing)

CDF2 F_X is right continuous (continuous from the right)

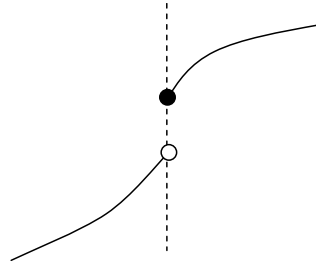


Figure 10: Right-continuous function at jump point

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

8.21. For discrete random variable, the cdf 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.

³¹These properties hold for any type of random variables. Moreover, for any function F that satisfies these three properties, there exists a random variable X whose CDF is F .

ECS315 2014/1 Part III.2 Dr.Prapun

8.3 Families of Discrete Random Variables

Many physical systems can be modeled by the same or similar random experiments and random variables. In this subsection, we present the analysis of several discrete random variables that frequently arise in applications.³²

Definition 8.22. X is *uniformly distributed* on a finite set S if

$$p_X(x) = P[X = x] = \begin{cases} \frac{1}{|S|}, & x \in S, \\ 0, & \text{otherwise,} \end{cases}$$

- We write $X \sim \mathcal{U}(S)$ or $X \sim \text{Uniform}(S)$.
- Read “ X is uniform on S ” or “ X is a uniform random variable on set S ”.
- The pmf is usually referred to as the uniform discrete distribution.
- Simulation: When the support S contains only consecutive integers³³, it can be generated by the command `randi` in MATLAB (R2008b).

³²As mention in 7.15, we often omit a discussion of the underlying sample space of the random experiment and directly describe the distribution of a particular random variable.

³³or, with minor manipulation, only uniformly spaced numbers

Example 8.23. X is uniformly distributed on $1, 2, \dots, n$ if

In MATLAB, X can be generated by `randi(n)`.

Example 8.24. 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
- Fair gaming devices (well-balanced coins and dice, well-shuffled decks of cards)

Example 8.25. Roll a fair dice. Let X be the outcome.

Definition 8.26. X is a ***Bernoulli*** random variable if

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

- Write $X \sim \mathcal{B}(1, p)$ or $X \sim \text{Bernoulli}(p)$
- X takes only two values: 0 or 1

Definition 8.27. X is a ***binary*** random variable if

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

- X takes only two values: a or b

Definition 8.28. X is a *binomial* random variable with size $n \in \mathbb{N}$ and parameter $p \in (0, 1)$ if

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x \in \{0, 1, 2, \dots, n\}, \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

- Write $X \sim \mathcal{B}(n, p)$ or $X \sim \text{binomial}(n, p)$.
 - Observe that $\mathcal{B}(1, p)$ is Bernoulli with parameter p .
- To calculate $p_X(x)$, can use `binopdf(x,n,p)` in MATLAB.
- Interpretation: X is the number of successes in n independent Bernoulli trials.

Example 8.29. An optical inspection system is to distinguish among different part types. The probability of a correct classification of any part is 0.98. Suppose that three parts are inspected and that the classifications are independent.

- (a) Let the random variable X denote the number of parts that are correctly classified. Determine the probability mass function of X . [14, Q3-20]
- (b) Let the random variable Y denote the number of parts that are incorrectly classified. Determine the probability mass function of Y .

Solution:

- (a) X is a binomial random variable with $n = 3$ and $p = 0.98$. Hence,

$$p_X(x) = \begin{cases} \binom{3}{x} 0.98^x (0.02)^{3-x}, & x \in \{0, 1, 2, 3\}, \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

In particular, $p_X(0) = 8 \times 10^{-6}$, $p_X(1) = 0.001176$, $p_X(2) = 0.057624$, and $p_X(3) = 0.941192$. Note that in MATLAB, these probabilities can be calculated by evaluating `binopdf(0:3,3,0.98)`.

(b) Y is a binomial random variable with $n = 3$ and $p = 0.02$. Hence,

$$p_Y(y) = \begin{cases} \binom{3}{y} 0.02^y (0.98)^{3-y}, & y \in \{0, 1, 2, 3\}, \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

In particular, $p_Y(0) = 0.941192$, $p_Y(1) = 0.057624$, $p_Y(2) = 0.001176$, and $p_Y(3) = 8 \times 10^{-6}$. Note that in MATLAB, these probabilities can be calculated by evaluating `binopdf(0:3,3,0.02)`.

Alternatively, note that there are three parts. If X of them are classified correctly, then the number of incorrectly classified parts is $n - X$, which is what we defined as Y . Therefore, $Y = 3 - X$. Hence, $p_Y(y) = P[Y = y] = P[3 - X = y] = P[X = 3 - y] = p_X(3 - y)$.

Example 8.30. 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%. [17, Ex 4.1]

Definition 8.31. A geometric random variable X is defined by the fact that for some constant $\beta \in (0, 1)$,

$$p_X(k+1) = \beta \times p_X(k)$$

for all $k \in S$ where S can be either \mathbb{N} or $\mathbb{N} \cup \{0\}$.

(a) When its support is $\mathbb{N} = \{1, 2, \dots\}$,

$$p_X(x) = \begin{cases} (1 - \beta) \beta^{x-1}, & x = 1, 2, \dots \\ 0, & \text{otherwise.} \end{cases}$$

- In MATLAB, to evaluate $p_X(x)$, use `geopdf(x-1, 1-beta)`.
- Interpretation: X is the number of trials required in Bernoulli trials to achieve the first success.

In particular, in a series of Bernoulli trials (independent trials with constant probability p of a success), let the random variable X denote the number of trials until the first success. Then X is a geometric random variable with parameter $\beta = 1 - p$ and

$$\begin{aligned} p_X(x) &= \begin{cases} (1 - \beta) \beta^{x-1}, & x = 1, 2, \dots \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} p(1 - p)^{x-1}, & x = 1, 2, \dots \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

- Write $X \sim \mathcal{G}_1(p)$ or `geometric1(p)`.

(b) When its support is $\mathbb{N} \cup \{0\}$,

$$\begin{aligned} p_X(x) &= \begin{cases} (1 - \beta) \beta^x, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} p(1 - p)^x, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

- Write $X \sim \mathcal{G}_0(p)$ or `geometric0(p)`.
- In MATLAB, to evaluate $p_X(x)$, use `geopdf(x, 1-beta)`.
- Interpretation: X is the number of failures in Bernoulli trials before the first success occurs.

8.32. 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*). [17]

Definition 8.33. X is a **Poisson** random variable with **parameter** $\alpha > 0$ if

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

- In MATLAB, use `poisspdf(x,alpha)`.
- Write $X \sim \mathcal{P}(\alpha)$ or $\text{Poisson}(\alpha)$.
- We will see later in Example ?? 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

Example 8.34. 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]³⁴.

³⁴I. 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 8.35. 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.

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

Definition 8.37. 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.

Example 8.38. 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

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

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

Definition 8.40. For *homogeneous Poisson process*, there is only one parameter that describes the whole process. This number is called the *rate* and usually denoted by λ .

Example 8.41. 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.

8.42. One important fact which we will revisit later is that, 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 8.43. 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 8.44. 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.

8.45. *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. [17, p 109]

(b) More specifically, suppose $X \sim \mathcal{B}(n, p_n)$. If $p_n \rightarrow 0$ and $np_n \rightarrow \alpha$ as $n \rightarrow \infty$, then for $x = 0, 1, 2, \dots$, we have³⁵

$$P[X = x] = \binom{n}{x} p_n^x (1 - p_n)^{n-x} \xrightarrow{n \rightarrow \infty} e^{-\alpha} \frac{\alpha^x}{x!}.$$

Example 8.46. Consider $X \sim \mathcal{B}(n, 1/n)$. (We have already seen this in Example 6.51.) For $x = 0, 1, 2, \dots$, we have

$$P[X = x] = \binom{n}{x} \left(\frac{1}{n}\right)^x \left(1 - \frac{1}{n}\right)^{n-x} \xrightarrow{n \rightarrow \infty} \frac{1}{x!e}.$$

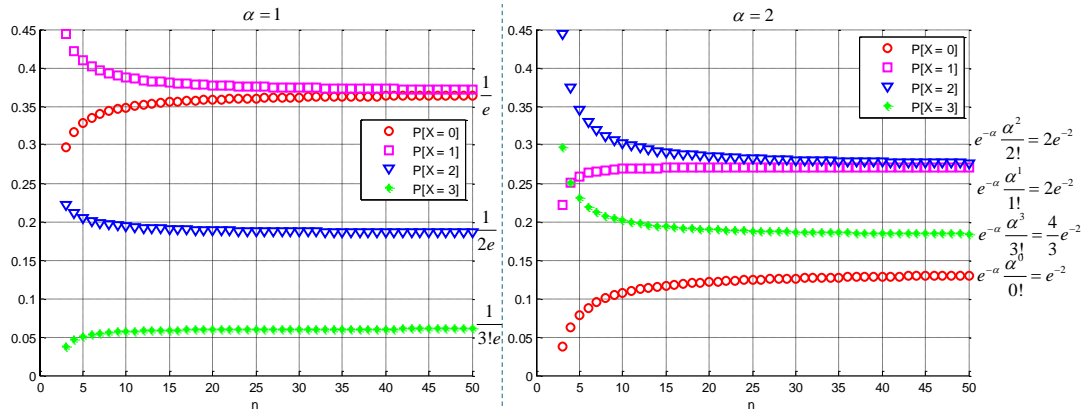


Figure 11: Pointwise convergence of the Binomial($n, \frac{\alpha}{n}$) pmf to the Poisson pmf when $\alpha = 1$ and $\alpha = 2$.

Example 8.47. Consider $X \sim \mathcal{B}(n, \alpha/n)$. For $x = 0, 1, 2, \dots$, we have

$$P[X = x] = \binom{n}{x} \left(\frac{\alpha}{n}\right)^x \left(1 - \frac{\alpha}{n}\right)^{n-x} \xrightarrow{n \rightarrow \infty} e^{-\alpha} \frac{\alpha^x}{x!}.$$

³⁵To see this, note that the first x (largest) terms of $n!$ can be bounded by $n - x \leq n - k \leq n$. Therefore, $\frac{(n-x)^x}{x!} \leq \binom{n}{x} \leq \frac{n^x}{x!}$ and

$$P[X = x] = \underbrace{\binom{n}{x}}_{\rightarrow \frac{1}{x!}} \underbrace{\frac{1}{n^x} (np_n)^x}_{\rightarrow \alpha^x} \underbrace{(1 - p_n)^n}_{=(1 - \frac{np_n}{n})^n \rightarrow e^{-\alpha}} \underbrace{(1 - p_n)^{-x}}_{\rightarrow 1}.$$

Example 8.48. In one of the New York state lottery games, a number is chosen at random between 0 and 999. Suppose you play this game 250 times. Use the Poisson approximation to estimate the probability that you will never win and compare this with the exact answer. [3, Q2.41]

Solution: Let W be the number of wins. Then, $W \sim \text{Binomial}(250, p)$ where $p = 1/1000$. Hence, $P[W = 0] = \binom{250}{0} p^0 (1-p)^{250} \approx 0.7787$.

If we approximate W by $\Lambda \sim \mathcal{P}(\alpha)$. Then we need to set $\alpha = np = \frac{250}{1000} = \frac{1}{4}$. In which case, $P[\Lambda = 0] = e^{-\alpha} \frac{\alpha^0}{0!} = e^{-\alpha} \approx 0.7788$ which is very close to the answer from direct calculation.

Example 8.49. 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).

8.50. Summary:

$X \sim$	Support S_X	$p_X(x) =$
Uniform $\mathcal{U}(S)$	S	$\begin{cases} \frac{1}{ S }, & x \in S, \\ 0, & \text{otherwise.} \end{cases}$
Bernoulli(p)	$\{0, 1\}$	$\begin{cases} 1-p, & x = 0, \\ p, & x = 1, \\ 0, & \text{otherwise.} \end{cases}$
Binomial $\mathcal{B}(n, p)$	$\{0, 1, \dots, n\}$	$\begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x \in \{0, 1, 2, \dots, n\}, \\ 0, & \text{otherwise.} \end{cases}$
Geometric $\mathcal{G}_0(p)$	$\mathbb{N} \cup \{0\}$	$\begin{cases} p(1-p)^x, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise.} \end{cases}$
Geometric $\mathcal{G}_1(p)$	\mathbb{N}	$\begin{cases} p(1-p)^{x-1}, & x = 1, 2, \dots \\ 0, & \text{otherwise.} \end{cases}$
Poisson $\mathcal{P}(\alpha)$	$\mathbb{N} \cup \{0\}$	$\begin{cases} e^{-\alpha} \frac{\alpha^x}{x!}, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$

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

8.4 Some Remarks

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

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

Example 8.52. For $\mathcal{B}(3, \frac{1}{3})$, we may define

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

and

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

8.53. 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 8.54. If $p_X(c) = 1$, that is $P[X = c] = 1$, for some constant c , then X is called a *degenerated* random variable.

References

- [1] Richard A. Brualdi. *Introductory Combinatorics*. Prentice Hall, 5 edition, January 2009. 4.3, 4.4, 5, 4.13
- [2] F. N. David. *Games, Gods and Gambling: A History of Probability and Statistical Ideas*. Dover Publications, unabridged edition, February 1998. 4.39
- [3] Rick Durrett. *Elementary Probability for Applications*. Cambridge University Press, 1 edition, July 2009. 4.25, 4.26, 8.34, 8.48
- [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. 5
- [6] Terrence L. Fine. *Probability and Probabilistic Reasoning for Electrical Engineering*. Prentice Hall, 2005. 3.1, 8.34
- [7] Martin Gardner. *Entertaining mathematical puzzles*. Dover, 1986. 1.13
- [8] John A. Gubner. *Probability and Random Processes for Electrical and Computer Engineers*. Cambridge University Press, 2006. 2.6, 4.17, 1, 6.25
- [9] John Haigh. *Taking Chances: Winning with Probability*. Oxford University Press, 2003. 5.13
- [10] Daniel Kahneman, Paul Slovic, and Amos Tversky. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1 edition, April 1982. 5.12
- [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

- [13] Leonard Mlodinow. *The Drunkard's Walk: How Randomness Rules Our Lives*. Pantheon; 8th Printing edition, 2008. 1.8, 1.9, 4.40, 4.41, 6.27, 6.39
- [14] Douglas C. Montgomery and George C. Runger. *Applied Statistics and Probability for Engineers*. Wiley, 2010. 6.20, 6.36, 8.1, 30, 1
- [15] 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, 4.8, 4.9, 5.12, 5.16, 6.24, 6.26, 6.37, 6.39
- [16] Mark F. Schilling. The longest run of heads. *The College Mathematics Journal*, 21(3):196–207, 1990. 4.46, 4.47
- [17] Henk Tijms. *Understanding Probability: Chance Rules in Everyday Life*. Cambridge University Press, 2 edition, August 2007. 1.8, 1.22, 4.15, 4.26, 4.42, 4.48, 5, 6.11, 3, 8.30, 8.32, 1
- [18] John M. Wozencraft and Irwin Mark Jacobs. *Principles of Communication Engineering*. Waveland Press, June 1990. 1.4
- [19] Rodger E. Ziemer and William H. Tranter. *Principles of Communications*. John Wiley & Sons Ltd, 2010. 6.30, 27