ECS315 2019/1 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. [21, p 223]

We will try to avoid several technical details[14] [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.

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

The theory of probability focuses on collections of events, called event ***σ-algebras***, typically denoted by $\mathcal{A}$ (or $\mathcal{F}$), that contain all the events of interest (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 $\mathcal{A}$.

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

Definition 5.1. Kolmogorov's Axioms for Probability [12]: A **probability measure**[16] is a real-valued set function[17] 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,

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

- The number $P\left(A\right)$ is called the ***probability*** of the event A

From the three axioms[18], we can derive many more properties of probability measure. These properties are useful for calculating probabilities.

Definition 5.2. Some definitions involving events whose probability = 1.

- The entire sample space Ω is called the ***sure event*** or the ***certain event***.
- If an event A satisfies $P(A) = 1$, we say that A is an ***almost-sure event***.
- A ***support*** of P is any set A for which $P\left(A\right) = 1$.

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

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

[18]Remark: 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.

Example 5.3. "Direct" construction of a probability measure: Consider a sample space $\Omega = \{1, 2, 3\}$.

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

5.5. Finite additivity[19]: If $A_1, \ldots, A_n$ are disjoint events, then

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

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

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

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

5.6. The probability of a countable event equals the sum of the probabilities of the outcomes in the event.

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

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

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

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

- This greatly simplifies[20] construction of probability measure.

Remark: Note again that the set A under consideration here is finite or countably infinite. You cannot 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, 5.6 implies that we only need to define P for all the singletons (when Ω is countable).

[21]In Section 10, 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 5.6. 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 5.6 are not enough to evaluate $P(A)$.

Example 5.7. 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 $\{d\}$.

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

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

Example 5.9. 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.

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

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

5.12. $P(A \cap B)$ cannot 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 5.13 (Slides)**.** Experiments by psychologists Kahneman and Tversky.

Example 5.14. 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* [11] by Economics Nobel laureate Daniel Kahneman and co-authors Paul Slovic and Amos Tversky. They show empirically how people often make similar mistakes when they are asked to choose the most probable among a set of statements. It certainly helps to know the rules of probability. A more discomforting aspect is that the more you explain something in detail, the more likely you are to be wrong. If you want to be credible, be vague. [17, p 11–12]

5.15. Complement Rule:

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

- "The probability that something does not occur can be computed as one minus the probability that it does occur."
- Named "probability's Trick Number One" in [10]

5.16. 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 5.17 (Slides)**.** Combining error probabilities from various sources in DNA testing

Example 5.18. 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 5.16:

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

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

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

5.19. Probability of a union of three events:

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

5.20. Two bounds:

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

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

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

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

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

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

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

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

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

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

We must have

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

Now, given any event finite[22] event A, we can apply 5.6 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.

[22] In 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...

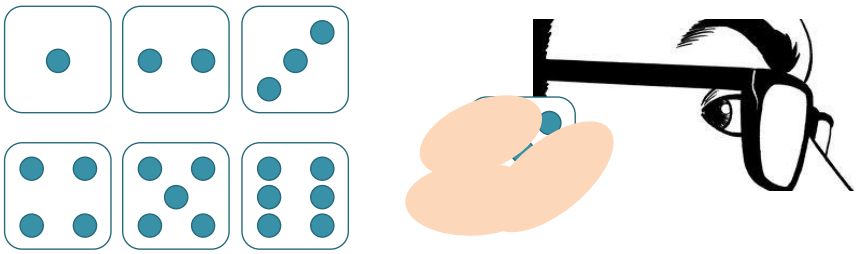


Figure 10: Conditional Probability Example: Sneak Peek

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

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

- Some ways to say[23] or express the conditional probability, $P(A|B)$, are:
  - the "(conditional) probability of A, given B"
  - the "(conditional) probability of A, knowing B"
  - the "(conditional) probability of A happening, knowing B has already occurred"
  - the "(conditional) probability of A, given that B occurred"
  - the "(conditional) probability of an event A under the knowledge that the outcome will be in event B"

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

- 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***: It is sometimes useful to interpret $P(A)$ as our knowledge of the occurrence of event A *before* the experiment takes place. Conditional probability[24] $P(A|B)$ 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 did occur).

Definition 6.5. Sometimes, we refer to $P(A)$ as

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

Example 6.6. Back to Example 6.1. Roll a dice. Let X be the outcome.

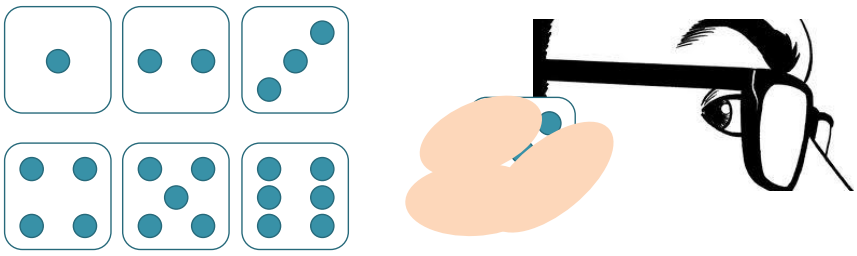


Figure 11: Sneak Peek: A Revisit

[24]In general, $P(A)$ and $P(A|B)$ are not the same. However, in the next section (Section 6.2), we will consider the situation in which they are the same.

Example 6.7. 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 $P(D|T_P) \approx 0.01$.

6.8. "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" issues[25] that we need to deal with when we actually define independence.

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

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

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

[25]Here, 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 $P(B) = 0$.

Exercise 6.10. 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 $\frac{1}{6}$). [21, Example 8.1, p. 244]

Example 6.11. Consider the following sequences of 1s and 0s which summarize the data obtained from 15 testees.

```
D: 0 1 1 0 0 0 0 1 1 1 1 0 1 0 1

TP: 1 0 0 1 1 0 0 0 0 0 1 1 0 1 1
```

The "D" row indicates whether each of the testees actually has the disease under investigation. The "TP" row indicates whether each of the testees is tested positive for the disease.

Numbers "1" and "0" correspond to "True" and "False", respectively.

Suppose we randomly pick a testee from this pool of 15 persons. Let D be the event that this selected person actually has the disease. Let T_P be the event that this selected person is tested positive for the disease.

Find the following probabilities.

(a) $P(D)$

(b) $P(D^c)$

(c) $P(T_P)$

(d) $P(T_P^c)$

(e) $P(T_P|D)$

(f) $P(T_P|D^c)$

(g) $P(T_P^c|D)$

(h) $P(T_P^c|D^c)$

6.12. Similar properties to the three probability axioms:

(a) Nonnegativity: $P(A \quad) \geq 0$

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

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,

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

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

$$P\left(A_1 \cup A_2 \quad\right) = P(A_1 \quad) + P(A_2 \quad)$$

6.13. More Properties:

- $P(A|\Omega) = P(A)$
- $P(A^c|B) = 1 - P(A|B)$
- $P(A \cap B|B) = P(A|B)$
- $P\left(A_1 \cup A_2|B\right) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$.
- $P(A \cap B) \leq P(A|B)$

6.14. Probability of compound events

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

(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 conditioning part, we often use "," instead of "$\cap$".

Example 6.15. 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

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

[21, p 243]

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

Example 6.16. 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?

6.17. Tree Diagram and Conditional Probability: Conditional probabilities can be represented on a tree diagram as shown in Figure 12.

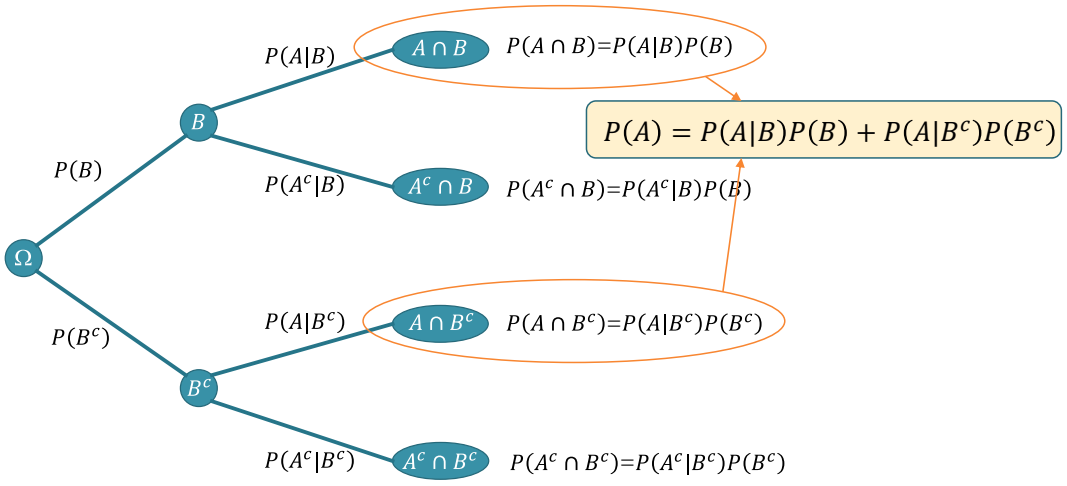


Figure 12: Tree Diagram and Conditional Probabilities

A more compact representation is shown in Figure 13.

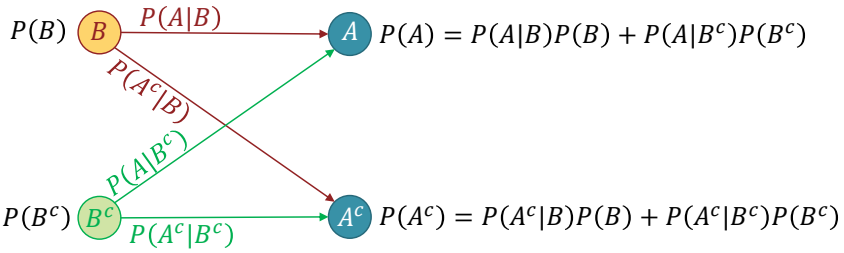


Figure 13: Compact Diagram for Conditional Probabilities

Example 6.18. A simple digital communication channel called **binary symmetric channel** (BSC) is shown in Figure 6.58. This channel can be described as a channel that introduces random bit errors with probability p.

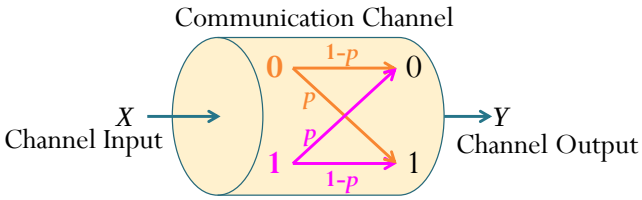


Figure 14: Binary Symmetric Channel (BSC)

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

$$P(A) = \sum_i P(A|B_i)P(B_i). \tag{8}$$

This is a formula[26] for computing the probability of an event that can occur in different ways. Observe that it follows directly from 5.21 and Definition 6.3.

[26]The 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.

- 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 6.16.

Example 6.20. Continue from the "Diagnostic Tests" Example 6.2 and Example 6.7.

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

For conciseness, we define

$$p_d = P(D)$$

and

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

Then,

$$P\left(T_P\right) = (1 - p_{TE})p_D + p_{TE}(1 - p_D).$$

6.21. Bayes' Theorem:

(a) Form 1:

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

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

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

- 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".

6.22. Summary:

(a) An easy but crucial property:

(b) Key setup: find a partition of the sample space

(c) Total probability theorem:

(d) Bayes' theorem:

- Special case: When there are only two cases: B_1 and B_2, we can think of them as B and B^c, respectively:
  - $P(A) =$
  - $P(B|A) =$
  - $P(B|A^c) =$

Example 6.23. Suppose $\Omega = \{a, b, c, d, e\}$. Define four events

$$A = \{a, b, c\},\ B = \{a, b\},\ C = \{c, d\}, \text{ and } D = \{e\}.$$

Let

$$P(\{a\}) = P(\{b\}) = 0.2, \quad \text{and} \quad P(\{c\}) = P(\{d\}) = 0.1.$$

Calculate the following probabilities:

(a) $P(\{e\})$

(b) $P(B)$, $P(C)$,

$P(D)$

(c) $P(A|B)$

$P(A|C)$

$P(A|D)$

(d) $P(A)$

Check: Observe that the collection $\{B, C, D\}$ partitions Ω. Use the total probability theorem to find $P(A)$.

(e) $P(B|A)$

Example 6.24. Continue from the "Disease Testing" Examples 6.2, 6.7, and 6.20:

$$P(D|T_P) = \frac{P(D \cap T_P)}{P(T_P)} = \frac{P(T_P|D)\,P(D)}{P(T_P)}$$
$$= \frac{(1-p_{TE})p_D}{(1-p_{TE})p_D + p_{TE}(1-p_D)}$$

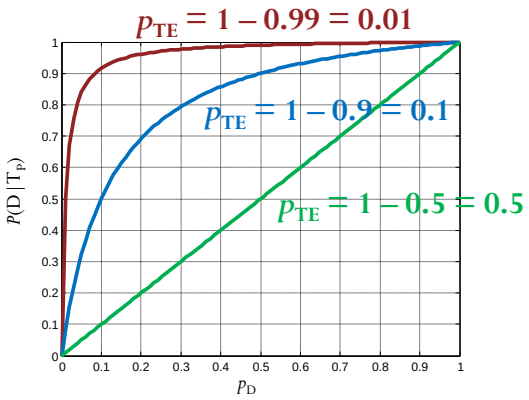


Figure 15: Probability $P(D|T_P)$ that a person will have the disease given that the test result is positive. The conditional probability is evaluated as a function of P_D which tells how common the disease is. Thee values of test error probability p_{TE} are shown.

Example 6.25. 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? [15, Ex. 2-37]

Example 6.26. 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.

$$
\begin{aligned}
P(H|F) &= \frac{P(H \cap F)}{P(F)} &&= \frac{P(F|H)P(H)}{P(F)} \\
P(S|F) &= \frac{P(S \cap F)}{P(F)} &&= \frac{P(F|S)P(S)}{P(F)} \\
P(O|F) &= \frac{P(O \cap F)}{P(F)} &&= \frac{P(F|O)P(O)}{P(F)}
\end{aligned}
$$

Example 6.27 (Slides)**.** The Murder of Nicole Brown

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

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

6.29. 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 mean we needs 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 ith 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 jth case happens is depenoted 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 replace 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 in to 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, total probability theorem deals with the calculation of the output probabilities $P(A_j)$:

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

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

Let us solve this quickly. Just to have some numbers, let us say that there are a total of $N = 100,000$ tickets and that $n = 1,000$ of them are chosen at random to be in the final drawing. If the drawing was from all tickets, your chance to win would be $1/N = 1/100,000$. The way it is actually done, you need to both survive the first drawing to get your ticket into the bag and then get your ticket drawn from the bag. The probability to get your entry into the bag is $n/N = 1,000/100,000$. The conditional probability to be drawn from the bag, given that your entry is in it, is $1/n = 1/1,000$. Multiply to get $1/N = 1/100,000$ once more. There were no riots in the streets. [17, p 22]

Example 6.31. Suppose your professor tells the class that there will be a surprise quiz next week. On one day, Monday-Friday, you will be told in the morning that a quiz is to be given on that day. You quickly realize that the quiz 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 quiz. 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 quiz. Thursday is ruled out, but then Wednesday becomes impossible, then Tuesday, then Monday, and you conclude: There is no such thing as a surprise quiz! But the professor decides to give the quiz 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 quiz is chosen randomly among the five days of the week. Now start a new school week. What is the probability that you get the test on Monday? Obviously 1/5 because this is the probability that Monday is chosen. If the test was not given on Monday. what is the probability that it is given on Tuesday? The probability that Tuesday is chosen to start with is 1/5, but we are now asking for the conditional probability that the test is given on Tuesday, given that it was not given on Monday. As there are now four days left, this conditional probability is 1/4. Similarly, the conditional probabilities that the test is given

on Wednesday, Thursday, and Friday conditioned on that it has not been given thus far are 1/3, 1/2, and 1, respectively.

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

Example 6.32. 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. [14, 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 6.33. Two events A, B are called (statistically[27]) ***independent*** if

$$P(A \cap B) = P(A)\,P(B) \tag{9}$$

- 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.

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

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

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 6.8. 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 6.37, we show how (10) can be used to define independence with extra condition that deals with the case when zero probability is involved.

[27]Sometimes 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 6.35. [25, Ex. 5.4] Which of the following pairs of events are independent?

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

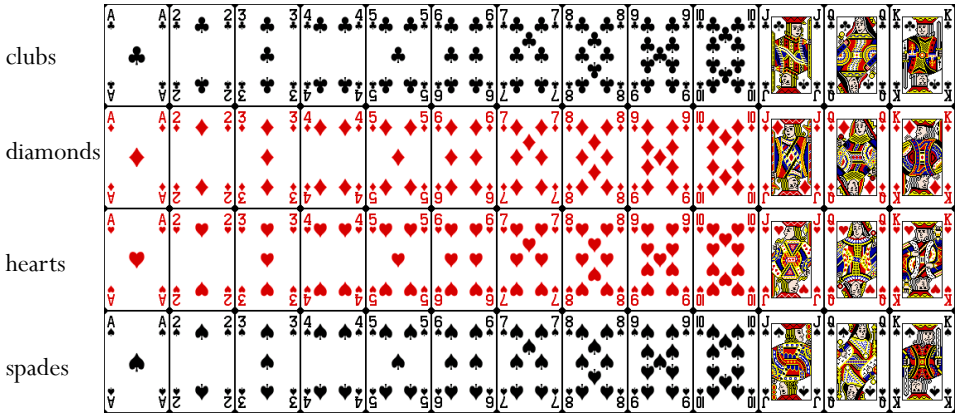


Figure 16: A Deck of Cards

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

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

- In particular, $\emptyset$ and Ω are independent of any events.
- One can also show that an event A is independent of itself if and only if $P(A)$ is 0 or 1.

6.37. Now that we have 6.36, we can now extend the "practival definition" from 6.34 to include events with zero probabilities:

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.

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

1) $A \perp\!\!\!\perp B$

2) $P(A \cap B) = P(A)P(B)$

3) $P(A|B) = P(A)$

4) $P(B|A) = P(B)$

6.39. 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 6.40. If $P(A|B) = 0.4$, $P(B) = 0.8$, and $P(A) = 0.5$, are the events A and B independent? [15]

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

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

- To check whether A and B are disjoint, we only need to look at the sets themselves and see whether they have shared outcome(s). This can be answered without knowing probabilities.

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

- Addition vs. multiplication:

  (a) If events A and B are disjoint, we calculate the probability of their union $A \cup B$ by adding the probabilities of A and B.

  (b) For independent events A and B, we calculate the probability of their 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 6.42. 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 6.43 (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 l/850,000. Now, this is still a small number and might not have made the jurors judge differently. But what does the probability 1/850,000 have to do with Sallys guilt? Nothing! When her first child died, it was certified to have been from natural causes and there was no suspicion of foul play. The probability that it would happen again without foul play was 1/100, and if that number had been presented to the jury, Sally would not have had to spend three years in jail before the verdict was finally overturned and the expert witness (certainly no expert in probability) found guilty of "serious professional misconduct."

You may still ask the question what the probability 1/100 has to do with Sallys guilt. Is this the probability that she is innocent? Not at all. That would mean that 99% of all mothers who experience two cot deaths are murderers! The number 1/100 is simply the probability of a second cot death, which only means that among all families who experience one cot death, about 1% will suffer through another. If probability arguments are used in court cases, it is very important that all involved parties understand some basic probability. In Sallys case, nobody did.

References: [14, 118–119] and [17, 22–23].

Definition 6.44. 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 6.45. The events $A_1, A_2, \ldots, A_n$ are ***independent*** if and only if for any subcollection $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$,

$$P\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) = P\left(A_{i_1}\right) \times P\left(A_{i_2}\right) \times \cdots \times P\left(A_{i_n}\right).$$

- Note that part of the requirement is that

  $$P\left(A_1 \cap A_2 \cap \cdots \cap A_n\right) = P\left(A_1\right) \times P\left(A_2\right) \times \cdots \times P\left(A_n\right).$$

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

  $$P\left(A_1 \cap A_2 \cap \cdots \cap A_n\right) = P\left(A_1\right) \times P\left(A_2\right) \times \cdots \times P\left(A_n\right).$$

- Equivalently, this is the same as the requirement that

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

- 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.

6.46. 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:

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

Example 6.47. Suppose five events A, B, C, D, E are independent with

$$P(A) = P(B) = P(C) = P(D) = P(E) = \frac{1}{3}.$$

(a) Can they be (mutually) disjoint?

(b) Find $P(A \cup B)$

(c) Find $P\left((A \cup B) \cap C\right)$

(d) Find $P\left(A \cap C \cap D^c\right)$

(e) Find $P(A \cap B|C)$

6.3 Bernoulli Trials

Example 6.48. 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 6.49. Each of the random experiments in Example 6.48 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 6.50. 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 $1 - p$.

- Only one important parameter:

$$p = \text{success probability (probability of "success")}$$

- 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 6.51. 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?".

Definition 6.52. (Independent) ***Bernoulli Trials*** = a Bernoulli trial is repeated many times.

(a) It is usually[28] 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 6.53. Toss ***unfair coin*** n times.

- The overall sample space is $\Omega = \{H, T\}^n$.
  - There are 2^n elements. Each has the form $(\omega_1, \omega_2, \ldots, \omega_n)$ where ω_i = H or T.
- The n tosses are independent. Therefore,

$$P(\{\text{HHHTT}\}) =$$

[28] Unless stated otherwise or having enough evidence against, assume the trials are independent.

Example 6.54. What is the probability of two failures and three successes in five Bernoulli trials with success probability p.

Let's represent success and failure by 1 and 0, respectively. The outcomes with three successes in five trials are listed below:

Outcome	Corresponding probability
11100	
11010	
11001	
10110	
10101	
10011	
01110	
01101	
01011	
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 $10(1-p)^2 p^3$

6.55. The probability of exactly k successes in n bernoulli trials is

$$\binom{n}{k}(1-p)^{n-k}p^k.$$

Example 6.56. Consider a particular disease with prevalence $P(D) = 10^{-4}$: when a person is selected randomly from the general population, the probability that (s)he has this disease is 10^{-4} or 1-in-n where $n = 10^4$.

Suppose we randomly select $n = 10^4$ people from the general population. What is the chance that we find at least one person with this disease?

Example 6.57. At least one occurrence of a 1-in-n-chance event in n repeated trials:

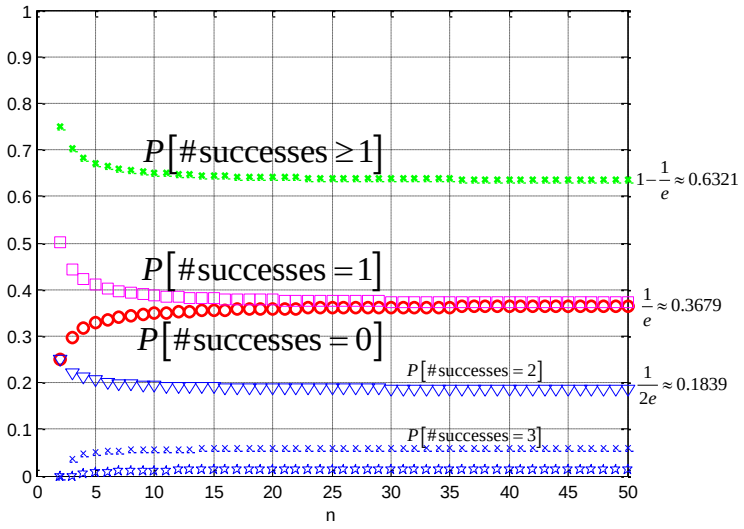
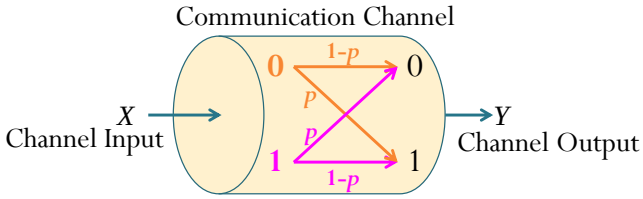


Figure 17: Number of occurrences of 1-in-n-chance event in n repeated Bernoulli trials

Example 6.58. ***Digital communication over unreliable channels***: Consider a digital communication system through the **binary symmetric channel** (BSC) discussed in Example 6.18. We repeat its compact description here.



Again this channel can be described as a channel that introduces random bit errors with probability p. This p is called the crossover probability.

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:

- At the transmitter, the "encoder" box performs the following task:
  - To send a 1, it will send 11111 through the channel.
  - To send a 0, it will send 00000 through the channel.
- 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.
- 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,
  - if $\#1 > \#0$, then it claims that "1" was transmitted,
  - if $\#0 > \#1$, then it claims that "0" was transmitted.

Two ways to calculate the probability of error:

(a) (transmission) error occurs if and only if the number of bits in error are ≥ 3.

(b) (transmission) error occurs if and only if the number of bits *not* in error are ≤ 2.

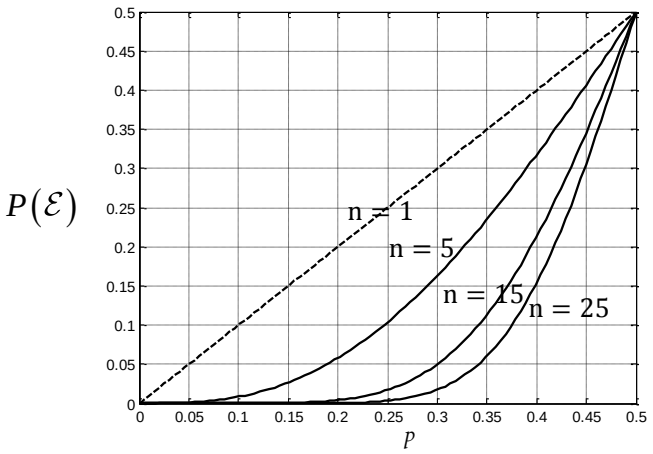


Figure 18: Overall bit error probability for a simple system that uses repetition code at the transmitter (repeat each bit n times) and majority vote at the receiver. The channel is assumed to be binary symmetric with bit error probability p.

Exercise 6.59 (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.

Example 6.60. A stream of bits is transmitted over a binary symmetric channel with crossover probability p.

(a) Consider the first seven bits.

(i) What is the probability that exactly four bits are received in error?

(ii) What is the probability that at least one bit is received correctly?

(b) What is the probability that the *first* error occurs at the fifth bit?

(c) What is the probability that the *first* error occurs at the kth bit?

(d) What is the probability that the *first* error occurs before or at the kth bit?