ECS 315: Probability and Random Processes

Article · September 2010

Prapun Suksompong
Sirindhorn International Institute of Technology (SIIT)

Sirindhorn International Institute of Technology

Thammasat University

School of Information, Computer and Communication Technology

ECS 315: Probability and Random Processes

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

August 4, 2010

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

Contents

1 Motivation

1.1. Random phenomena arise because of [9]:

(a) our partial ignorance of the generating mechanism

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

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

1.2. Communication Theory [13]: The essence of communication is randomness.

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

- 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.

1.3. 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.

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

Definition 1.5. Important terms [9]:

(a) An experiment 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.

   - Tossing/flipping a coin, rolling a die, and drawing a card from a deck are some examples of random experiments. See Section **??** for more details.

(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 of a random experiment. Each outcome is an element, or sample point, of this space.

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

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

1.6. Although the outcome of a random experiment is unpredictable, there is a statistical regularity about the outcomes.

- For example, if a coin is tossed a large number of times, about half the times the outcome will be "heads," and the remaining half of the times it will be "tails."

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

$$f(A) = \lim_{n\to\infty} \frac{N(A,n)}{n}$$

is called the **relative frequency** of the event A.

2 Set Theory

2.1. Basic Set Identities:

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

- Commutativity (symmetry):

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

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

name	*rule*
Commutative laws	$A \cap B = B \cap A \qquad A \cup B = B \cup A$
Associative laws	$A \cap (B \cap C) = (A \cap B) \cap C$ $A \cup (B \cup C) = (A \cup B) \cup C$
Distributive laws	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
DeMorgan's laws	$\overline{A \cap B} = \overline{A} \cup \overline{B} \qquad \overline{A \cup B} = \overline{A} \cap \overline{B}$
Complement laws	$A \cap \overline{A} = \emptyset \qquad A \cup \overline{A} = U$
Double complement law	$\overline{\overline{A}} = A$
Idempotent laws	$A \cap A = A \qquad A \cup A = A$
Absorption laws	$A \cap (A \cup B) = A \qquad A \cup (A \cap B) = A$
Dominance laws	$A \cap \emptyset = \emptyset \qquad A \cup U = U$
Identity laws	$A \cup \emptyset = A \qquad A \cap U = A$

Figure 1: Set Identities [10]

2.2. Basic Terminology:

- $A \cap B$ is sometimes written simply as AB.
- The ***set difference*** operation is defined by $B \setminus A = B \cap A^c$.

- Sets A and B are said to be ***disjoint*** ($A \perp B$) if and only if $A \cap B = \emptyset$

- A collection of sets ($A_i : i \in I$) is said to be pair-wise disjoint or mutually exclusive [6, p. 9] if and only if $\mathrm{A}_i \cap \mathrm{A}_j = \emptyset$ when $\mathrm{i} \neq \mathrm{j}$.

- A collection $\Pi = (A_\alpha : \alpha \in I)$ of subsets of Ω (in this case, indexed or labeled by α taking values in an index or label set I) is said to be a ***partition*** of Ω if

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

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

  In which case, the collection $(B \cap A_\alpha : \alpha \in I)$ is a partition of B. In other words, any set B can be expressed as $B = \bigcup\limits_{\alpha} (B \cap A_i)$ where the union is a disjoint union.

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

- An infinite set A is said to be ***countable*** if the elements of A can be enumerated or listed in a sequence: $a_1, a_2, \dots$. Empty set and finite sets are also said to be countable.

- By a ***countably infinite*** set, we mean a countable set that is not finite. Examples of such sets include

  - The set $\mathbb{N} = \{1, 2, , 3, \dots\}$ 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
  - The set $\mathbb{Q}$ of all rational numbers
  - The set $\mathbb{Q}^+$ of positive rational numbers
  - The set of all finite-length sequences of natural numbers.
  - The set of all finite subsets of the natural numbers.

- A ***singleton*** is a set with exactly one element.

- $\mathbb{R} = (-\infty, \infty)$.

- 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 2.3. Probability theory renames some of the terminology in set theory. See Table 1 and Table 2.

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

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

Table 1: The terminology of set theory and probability theory

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

Table 2: Event Language

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. Classical probability remains of importance today and provides the most accessible introduction to the more general theory of probability.

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

$$P(A) = \frac{\|A\|}{\|\Omega\|} = \frac{\text{the number of cases favorable to the outcome of the event}}{\text{the total number of possible cases}}.$$

- In this section, we are more apt to refer to equipossible cases as ones selected at random. Probabilities can be evaluated for events whose elements are chosen at random by enumerating the number of elements in the event.
- The bases for identifying equipossibility were often
  - physical symmetry (e.g. a well-balanced die, made of homogeneous material in a cubical shape) or
  - a balance of information or knowledge concerning the various possible outcomes.
- Equipossibility is meaningful only for finite sample space, and, in this case, the evaluation of probability is accomplished through the definition of classical probability.

3.1. Basic properties of classical probability:

- $P(A) \geq 0$
- $P(\Omega) = 1$
- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ which comes directly from

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

- $A \perp B$ is equivalent to $P(A \cap B) = 0$.
  - ◦ In general probability theory, the above statement is not true. However, when we limit ourselves to classical probability, then the statement is true.
- $A \perp B \Rightarrow P(A \cup B) = P(A) + P(B)$
- Suppose $\Omega = \{\omega_1, \ldots, \omega_n\}$ and $P(\omega_i) = \frac{1}{n}$. Then $P(A) = \sum\limits_{\omega \in A} p(\omega)$.
  - ◦ The probability of an event is equal to the sum of the probabilities of its component outcomes because outcomes are mutually exclusive

3.2. Background on some frequently used examples

(a) 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.

(b) When a **coin** is tossed, it does not necessarily fall heads or tails; it can roll away or stand on its edge. Nevertheless, we shall agree to regard "**head**" (Problem Solutions Chapter 1) and "**tail**" (**T**) as the only possible outcomes of the experiment.

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

(c) A complete set of **cards** is called a pack or **deck**.

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

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

(iii) 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.

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

- Cards of the same face value are called of the same **kind**.

- "court" or face card: a king, queen, or jack of any suit.

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

3.3. *Chevalier de Mere's Scandal of Arithmetic*:

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

We have

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

and

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

Therefore, the first case is more probable.

Remark: Probability theory was originally inspired by gambling problems. In 1654, Chevalier de Mere invented a gambling system which bet even money on the second case above. However, when he began losing money, he asked his mathematician friend Blaise 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.

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. See Chapter 2 of the handbook [10] for more details.

4.2. Rules Of Sum, Product, And Quotient:

(a) ***Rule of sum***: When there are m cases such that the ith case has n_i options, for $i = 1, ..., m$, and no two of the cases have any options in common, the total number of options is $n_1 + n_2 + \cdots + n_m$.

(i) In set-theoretic terms, if sets $S_1, \ldots, S_m$ are finite and pairwise disjoint, then $|S_1 \cap S_2 \cap \cdots \cap S_m| = |S_1| + |S_2| + \cdots + |S_m|$.

(b) ***Rule of product***:

- If one task has n possible outcomes and a second task has m possible outcomes, then the joint occurrence of the two tasks has $n \times m$ possible outcomes.

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

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

4.3. Choosing objects from a collection is also called **sampling**, and the chosen objects are known as a **sample**. The four kinds of counting problems are:

(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 4.11 for "bars and stars" argument.

4.4. Given a set of n distinct items/objects, select a distinct **ordered**[1] sequence (word) of length r drawn from this set.

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

- Meaning
  - Ordered sampling of r out of n items with replacement.
    - ∗ An object can be chosen repeatedly.
- $\mu_{n,1} = n$
- $\mu_{1,r} = 1$
- Examples:
  - There are n^r different *sequences* of r cards drawn from a deck of n cards with replacement; i.e., we draw each card, make a note of it, put the card back in the deck and re-shuffle the deck before choosing the next card.
  - Suppose A is a finite set, then the cardinality of its power set is $\left|2^A\right| = 2^{|A|}$.
  - There are 2^r binary strings/sequences of length r.

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

(b) Sampling without replacement:

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

- Alternative notation: $C(n,r)$.
- Meaning
  - ○ Ordered sampling of $r \le n$ out of n items without replacement.
    - ∗ Once we choose an object, we remove that object from the collection and we cannot choose it again.
  - ○ "the number of possible r-**permutations** of n distinguishable objects"
- For integers r, n such that $r > n$, we have $(n)_r = 0$.
- Extended definition: The definition in product form

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

  can be extended to *any real number* n and a non-negative integer r. We define $(n)_0 = 1$. (This makes sense because we usually take the empty product to be 1.)
- $(n)_1 = n$
- $(n)_r = (n-(r-1))(n)_{r-1}$. For example, $(7)_5 = (7-4)(7)_4$.
- $(1)_r = \begin{cases} 1, & \text{if r = 1} \\ 0, & \text{if r > 1} \end{cases}$

- Ratio:

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

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

- For any integer n greater than 1, the symbol $n!$, pronounced "n factorial," is defined as the product of all positive integers less than or equal to n.

- $0! = 1! = 1$
- $n! = n(n-1)!$
- $n! = \int\limits_0^\infty e^{-t} t^n dt$
- Meaning: The number of ways that n distinct objects can be ordered.
- Computation:

  (a) `MATLAB`:

  - ○ Use `factorial(n)`. Since double precision numbers only have about 15 digits, the answer is only accurate for $n \leq 21$. For larger n, the answer will have the right magnitude, and is accurate for the first 15 digits.
  - ○ Use `perms(v)`, where v is a row vector of length n, to creates a matrix whose rows consist of all possible permutations of the n elements of v. (So the matrix will contain $n!$ rows and n columns.)

  (b) Google's web search box built-in calculator: `n!`

- Approximation: Stirling's Formula [4, p. 52]:

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

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

4.6. ***Binomial coefficient***:

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

(a) Read "n choose r".

(b) Alternative notation: $C(n,r)$.

(c) Meaning:

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

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

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

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

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

(d) 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.

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

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

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

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

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

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

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

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

4.7. ***Binomial theorem***:

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

(a) Let $x = y = 1$, then $\sum\limits_{r=0}^{n} \binom{n}{r} = 2^n$.

(b) Sum involving only the even terms (or only the odd terms):

$$\sum_{\substack{r=0\\ r\text{ even}}}^{n} \binom{n}{r} x^r y^{n-r} = \frac{1}{2}\left((x+y)^n + (y-x)^n\right), \text{ and}$$

$$\sum_{\substack{r=0\\ r\text{ odd}}}^{n} \binom{n}{r} x^r y^{n-r} = \frac{1}{2}\left((x+y)^n - (y-x)^n\right).$$

In particular, if $x + y = 1$, then

$$\sum_{\substack{r=0\\ r\text{ even}}}^{n} \binom{n}{r} x^r y^{n-r} = \frac{1}{2}\left(1 + (1-2x)^n\right), \text{ and} \tag{2a}$$

$$\sum_{\substack{r=0\\ r\text{ odd}}}^{n} \binom{n}{r} x^r y^{n-r} = \frac{1}{2}\left(1 - (1-2x)^n\right). \tag{2b}$$

(c) By repeated differentiating with respect to x followed by multiplication by x, we have

- $\sum_{r=0}^{n} r\binom{n}{r} x^r y^{n-r} = nx(x+y)^{n-1}$ and
- $\sum_{r=0}^{n} r^2\binom{n}{r} x^r y^{n-r} = nx\left(x(n-1)(x+y)^{n-2} + (x+y)^{n-1}\right)$.

For $x+y=1$, we have

- $\sum_{r=0}^{n} r\binom{n}{r} x^r (1-x)^{n-r} = nx$ and
- $\sum_{r=0}^{n} r^2\binom{n}{r} x^r (1-x)^{n-r} = nx\left(nx+1-x\right)$.

All identities above can be verified easily via `Mathcad`.

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

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

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

4.9. ***Multinomial Theorem***:

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

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

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

More specifically, we have

$$(x_1+\ldots+x_r)^n = \sum_{i_1=0}^{n}\sum_{i_2=0}^{n-i_1}\cdots\sum_{i_{r-1}=0}^{n-\sum\limits_{j<r-1} i_j} \frac{n!}{\left(n-\sum\limits_{k<n} i_k\right)!\prod\limits_{k<n} i_k!} x_r^{n-\sum\limits_{j<r} i_j} \prod_{k=1}^{r-1} x_k^{i_k}$$

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

4.10. The number of solutions to $x_1+x_2+\cdots+x_n = k$ when the x_i's are nonnegative integers is $\binom{k+n-1}{k} = \binom{k+n-1}{n-1}$.

(a) Suppose we further require that the x_i are strictly positive ($x_i \geq 1$), then there are $\binom{k-1}{n-1}$ 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 $\binom{k-(a_1+a_2+\cdots+a_n)+n-1}{n-1}$. Note that here we work with equivalent problem: $y_1+y_2+\cdots+y_n = k-\sum_{i=1}^{n} a_i$ where $y_i \geq 0$.

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

- 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.

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

- Suppose r letters are drawn with replacement from a set $\{a_1, a_2, \ldots, a_n\}$. Given a drawn sequence, let x_i be the number of a_i in the drawn sequence. Then, there are $\binom{n+r-1}{r}$ possible $x = x_1^n$.

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

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

The probability that at least one element appears twice is

$$p_u(n, r) = 1 - \prod_{i=1}^{r-1}\left(1 - \frac{i}{n}\right) \approx 1 - e^{-\frac{r(r-1)}{2n}}.$$

In fact, when $r - 1 < \frac{n}{2}$, (**??**) gives

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

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

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

Example 4.13. *Probability of coincidence birthday*: Probability that there is at least two people who have the same birthday[2] in a group of r persons

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

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

Example 4.14. ***Birthday Paradox***: In a group of 23 randomly selected people, the probability that at least two will share a birthday (assuming birthdays are equally likely to occur on any given day of the year) is about 0.5.

- At first glance it is surprising that the probability of 2 people having the same birthday is so large[3], 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. [2, p. 9]

Remark: The group size must be at least 253 people if you want a probability > 0.5 that someone will have the same birthday as you. [2, 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.

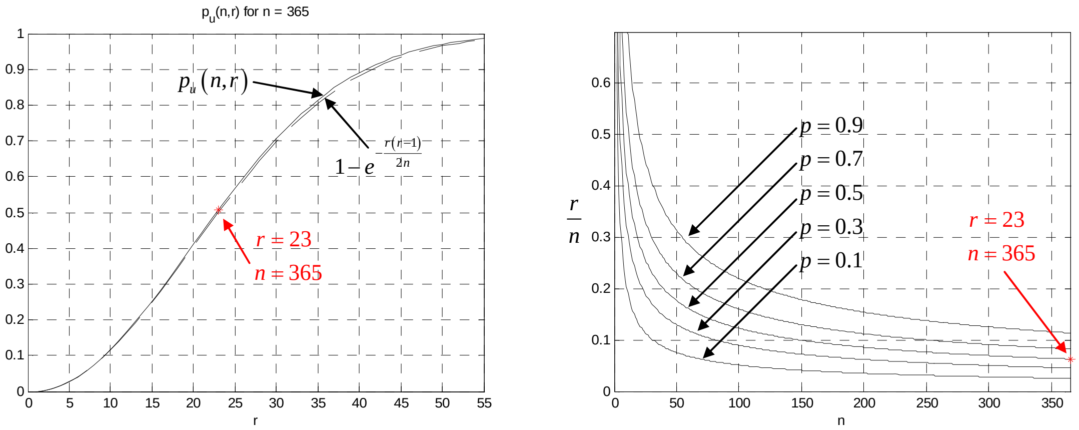

Figure 2: $p_u(n,r)$: The probability of the event that at least one element appears twice in random sample of size r with replacement is taken from a population of n elements.

Example 4.15. The Grand Duke of Tuscany "ordered" Galileo to explain a paradox arising in the experiment of tossing three dice [1]:

> "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:

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

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.

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

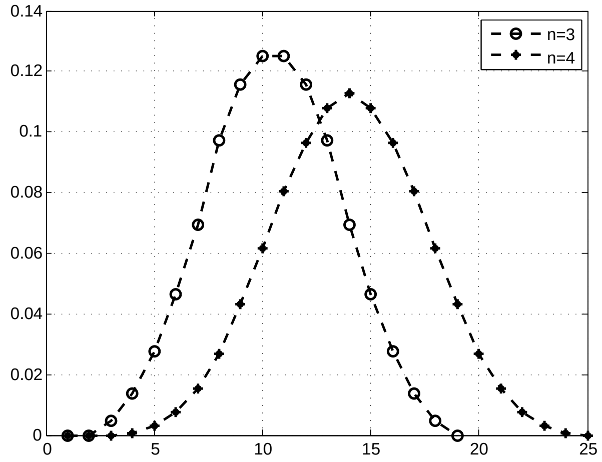

Figure 3: pmf of S_n for $n = 3$ and $n = 4$.

5 Probability Foundations

To study formal definition of probability, we start with the probability space $(\Omega, \mathcal{A}, P)$. Let Ω be an arbitrary space or set of points ω. 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 and typically denoted $\mathcal{A}$ (or $\mathcal{F}$) that contain all the events of interest[4] (regarding the random experiment $\mathcal{E}$) to us, and are such that we have knowledge of their likelihood of occurrence. The probability P itself is defined as a number in the range $[0, 1]$ associated with each event in $\mathcal{A}$.

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

Definition 5.1. Kolmogorov's Axioms for Probability [8]: A set function satisfying K0–K4 is called a ***probability measure***.

K0 Setup: The random experiment $\mathcal{E}$ is described by a probability space $(\Omega, \mathcal{A}, P)$ consisting of an event σ-algebra $\mathcal{A}$ and a real-valued function $P : \mathcal{A} \to \mathbb{R}$.

K1 Nonnegativity: $\forall A \in \mathcal{A},\ P(A) \geq 0$.

K2 Unit normalization: $P(\Omega) = 1$.

K3 Finite additivity: If A, B are disjoint, then $P(A \cup B) = P(A) + P(B)$.

K4 Monotone continuity: If $(\forall i > 1)\ A_{i+1} \subset A_i$ and $\cap_{i \in \mathbb{N}} A_i = \emptyset$ (a nested series of sets shrinking to the empty set), then

$$\lim_{i \to \infty} P(A_i) = 0.$$

K4′ Countable or σ-additivity: If (A_i) is a countable collection of pairwise disjoint (non-overlapping) events, then

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

- Note that there is never a problem with the convergence of the infinite sum; all partial sums of these non-negative summands are bounded above by 1.
- K4 is not a property of limits of sequences of relative frequencies nor meaningful in the finite sample space context of classical probability. It is offered by Kolmogorov to ensure a degree of mathematical closure under limiting operations [5, p. 111].
- K4 is an idealization that is rejected by some accounts (usually subjectivist) of probability.

[4] The class 2^{Ω} of all subsets can be too large[5] for us to define probability measures with consistency, across all member of the class.

- If P satisfies K0–K3, then it satisfies K4 if and only if it satisfies K4′ [5, Theorem 3.5.1 p. 111].
- K3 implies that K4′ holds for a finite number of events, but for infinitely many events this is a new assumption. Not everyone believes that K4′ (or its equivalent K4) should be used. However, without K4′ (or its equivalent K4) the theory of probability becomes much more difficult and less useful. In many cases the sample space is finite, so K4′ (or its equivalent K4) is not relevant anyway.

Equivalently, in stead of K0-K4, we can define probability measure using P0-P2 below.

Definition 5.2. A **probability measure** defined on a σ-algebra $\mathcal{A}$ of Ω is a (set) function

(P0) $P : \mathcal{A} \to [0, 1]$

that satisfies:

(P1,K2) $P(\Omega) = 1$

(P2,K4′) ***Countable additivity***: For every countable sequence $(A_n)_{n=1}^{\infty}$ of disjoint elements of $\mathcal{A}$, one has $P\left(\bigcup\limits_{n=1}^{\infty} A_n\right) = \sum\limits_{n=1}^{\infty} P(A_n)$

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

5.1 Properties of probability measures

5.3. Properties of probability measures:

(a) $P(\emptyset) = 0$

(b) $0 \le P \le 1$: For any $A \in \mathcal{A}$, $0 \le P(A) \le 1$

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

(d) Additivity: $A, B \in \mathcal{A}$, $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

(e) Monotonicity: $A, B \in \mathcal{A}$, $A \subset B \Rightarrow P(A) \le P(B)$ and $P(B \setminus A) = P(B) - P(A)$

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

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

(h) **Finite additivity**: If $A = \bigcup\limits_{j=1}^{n} A_j$ with $A_j \in \mathcal{A}$ disjoint, then $P(A) = \sum\limits_{j=1}^{n} P(A_j)$

- If A and B are disjoint sets in $\mathcal{A}$, then $P(A \cup B) = P(A) + P(B)$

(i) **Subadditivity or Boole's Inequality**: If $A_1, \ldots, A_n$ are events, not necessarily disjoint, then $P\left(\bigcup\limits_{i=1}^{n} A_i\right) \le \sum\limits_{i=1}^{n} P(A_i)$

(j) **σ-subadditivity**: If $A_1, A_2, \ldots$ is a sequence of measurable sets, not necessarily disjoint, then $P\left(\bigcup\limits_{i=1}^{\infty} A_i\right) \le \sum\limits_{i=1}^{\infty} P(A_i)$

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

5.4. $\mathcal{A}$ can not contain an uncountable, disjoint collection of sets of positive probability.

Definition 5.5. ***Discrete probability measure*** P is a discrete probability measure if $\exists$ finitely or countably many points ω_k and nonnegative masses m_k such that $\forall A \in \mathcal{A}$ $P(A) = \sum\limits_{k:\omega_k \in A} m_k = \sum\limits_{k} m_k I_A(\omega_k)$

If there is just one of these points, say ω_0, with mass $m_0 = 1$, then P is a ***unit mass*** at ω_0. In this case, $\forall A \in \mathcal{A}$, $P(A) = I_A(\omega_0)$.
Notation: $P = \delta_{\omega_0}$

- Here, Ω can be ***un***countable.

5.2 Countable Ω

A sample space Ω is countable if it is either finite or countably infinite. It is countably infinite if it has as many elements as there are integers. In either case, the element of Ω can be enumerated as, say, $\omega_1, \omega_2, \ldots$. If the event algebra $\mathcal{A}$ contains each singleton set $\{\omega_k\}$ (from which it follows that $\mathcal{A}$ is the power set of Ω), then we specify probabilities satisfying the Kolmogorov axioms through a restriction to the set $\mathcal{S} = \{\{\omega_k\}\}$ of singleton events.

Definition 5.6. When Ω is countable, a ***probability mass function*** (pmf) is any function $p : \Omega \to [0,1]$ such that

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

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

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

$$p(\omega) = P(\{\omega\}), \tag{3}$$

$$P(A) = \sum_{\omega \in A} p(\omega). \tag{4}$$

The convenience of a specification by pmf becomes clear when Ω is a finite set of, say, n elements. Specifying P requires specifying 2^n values, one for each event in $\mathcal{A}$, and doing so in a manner that is consistent with the Kolmogorov axioms. However, specifying p requires only providing n values, one for each element of Ω, satisfying the simple constraints of nonnegativity and addition to 1. The probability measure P satisfying (4) automatically satisfies the Kolmogorov axioms.

6 Event-based Independence and Conditional Probability

6.1 Event-based Conditional Probability

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

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{5}$$

- It is the updated probability of the event A given that we now know that B occurred.
- Read "the (conditional) probability of A given B".
- Sometimes, we refer to $P(A)$ as **a priori probability** , or the **prior probability** of A.
- $P(A|B) = P(A \cap B|B) \geq 0$
- For any A such that $B \subset A$, we have $P(A|B) = 1$. This implies

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

- If $A \perp C$, $P\left(A \cup C\left|B\right.\right) = P(A\left|B\right.) + P(C\left|B\right.)$
- $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.
  - **Bayes' Theorem**: $P(B|A) = P(A|B)\frac{P(B)}{P(A)}$
- $P(A \cap B) \leq P(A|B)$
- $P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B)$
- $P(A \cap B) = P(A) \times P(B|A)$
- $P(A \cap B \cap C) = P(A \cap B) \times P(C|A \cap B)$
- $P\left(A, B\left|C\right.\right) = P\left(A\left|C\right.\right) P\left(B\left|A, C\right.\right) = P\left(B\left|C\right.\right) P\left(A\left|B, C\right.\right)$

6.2. ***Total Probability and Bayes Theorem*** If $\{B_i, \ldots, B_n\}$ is a partition of Ω, then for any set A,

- *Total Probability Theorem*: $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$.
- *Bayes Theorem*: Suppose $P(A) > 0$, we have $P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$.

Example 6.3. Some seemingly surprising results in conditional probability: Someone has rolled a fair die twice. You know that one of the rolls turned up a face value of six. The probability that the other roll turned up a six as well is $\frac{1}{11}$ (not $\frac{1}{6}$). [11, Example 8.1, p. 244]

Example 6.4. *Monte Hall's Game*: Started with showing a contestant 3 closed doors behind of which was a prize. The contestant selected a door but before the door was opened, Monte Hall, who knew which door hid the prize, opened a remaining door. The contestant was then allowed to either stay with his original guess or change to the other closed door. Question: better to stay or to switch? Answer: Switch. Because after given that the contestant switched, then the probability that he won the prize is $\frac{2}{3}$.

6.2 Event-based Independence

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

Definition 6.6. Two events A, B are called ***independent*** if

$$P\left(A \cap B\right) = P\left(A\right) P\left(B\right)$$

- Notation: $A \perp\!\!\!\perp B$
- Read "A and B are independent" or "A is independent of B"
- In classical probability, this is equivalent to

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

6.7. Properties involving independence between two events:

(a) An event with probability 0 or 1 is independent of any event (including itself). In particular, $\emptyset$ and Ω are independent of any events.

(b) Two events A, B with positive probabilities are independent if and only if $P\left(B\left|A\right.\right) = P\left(B\right)$, which is equivalent to $P\left(A\left|B\right.\right) = P\left(A\right)$
When A and/or B has zero probability, A and B are automatically independent.

(c) An event A is independent of itself if and only if $P\left(A\right)$ is 0 or 1.

(d) If A an B are independent, then the two classes $\sigma\left(\{A\}\right) = \{\emptyset, A, A^c, \Omega\}$ and $\sigma\left(\{B\}\right) = \{\emptyset, B, B^c, \Omega\}$ are independent.

(i) If A and B are independent events, then so are A and B^c, A^c and B, and A^c and B^c. By interchanging the roles of A and A^c and/or B and B^c, it follows that if any one of the four pairs is independent, then so are the other three. [6, p.31]

(ii) These four conditions are equivalent:

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

(e) Suppose A and B are disjoint. A and B are independent if and only if $P(A) = 0$ or $P(B) = 0$.

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

Example 6.8. Experiment of flipping a fair coin twice. $\Omega = \{HH, HT, TH, TT\}$. Define event A to be the event that the first flip gives a H; that is $A = \{HH, HT\}$. Event B is the event that the second flip gives a H; that is $B = \{HH, TH\}$. $C = \{HH, TT\}$. Note also that even though the events A and B are not disjoint, they are independent.

Definition 6.9. Three events A_1, A_2, A_3 are independent if and only if

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

Remarks:

(a) The first equality alone is not enough for independence. See a counter example below. In fact, it is possible for the first equation to hold while the last three fail as shown in (6.10.b). It is also possible to construct events such that the last three equations hold (pairwise independence), but the first one does not as demonstrated in (6.10.a).

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

Example 6.10.

(a) Having three pairwise independent events does ***not*** imply that the three events are jointly independent. In other words,

$$A \perp\!\!\!\perp B, B \perp\!\!\!\perp C, A \perp\!\!\!\perp C \not\Rightarrow A \perp\!\!\!\perp B \perp\!\!\!\perp C.$$

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

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

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

(iii) A fair coin is tossed three times and the following events are considered: A = [toss 1 and toss 2 produce different outcomes], B = toss 2 and toss 3 produce different outcomes, C = toss 3 and toss 1 produce different outcomes.

(iv) Roll three dice. Let A = [the numbers on the first and second add to 7], B = [the numbers on the second and third add to 7], and C = [the numbers on the third and first add to 7]. [2, Ex. 1.10]

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

Definition 6.11. Independence between many events:

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

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

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

$\equiv \; P(B_1 \cap B_2 \cap \dots \cap B_n) = P(B_1)P(B_2)\cdots P(B_n)$ where $B_i = A_i$ or $B_i = \Omega$

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

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

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

Definition 6.12. A collection of events $\{A_\alpha\}$ is called ***pairwise independent*** if for every distinct events $A_{\alpha_1}, A_{\alpha_2}$, we have $P(A_{\alpha_1} \cap A_{\alpha_2}) = P(A_{\alpha_1})P(A_{\alpha_2})$

- If a collection of events $\{A_{:\alpha} : \alpha \in I\}$ is independent, then it is pairwise independent. The converse is false. See (a) in example (6.10).
- For $K \subset J$, $P\left(\bigcap\limits_{\alpha\in J} A_\alpha\right) = \prod\limits_{\alpha\in J} P(A)$ does not imply $P\left(\bigcap\limits_{\alpha\in K} A_\alpha\right) = \prod\limits_{\alpha\in K} P(A)$

Definition 6.13. Independence between many collections of events:

(a) Independence for finite collection $\{\mathcal{A}_1, \dots, \mathcal{A}_n\}$ of classes:

$\equiv$ the finite collection of sets $A_1, \dots, A_n$ is independent where $A_i \in \mathcal{A}_i$.

$\equiv \; P(B_1 \cap B_2 \cap \dots \cap B_n) = P(B_1)P(B_2)\cdots P(B_n)$ where $B_i \in \mathcal{A}_i$ or $B_i = \Omega$

$\equiv \; P(B_1 \cap B_2 \cap \dots \cap B_n) = P(B_1)P(B_2)\cdots P(B_n)$ where $B_i \in \mathcal{A}_i \cup \{\Omega\}$

$\equiv \forall i\ \forall \mathcal{B}_i \subset \mathcal{A}_i \quad \mathcal{B}_1, \dots, \mathcal{B}_n$ are independent.
$\equiv \mathcal{A}_1 \cup \{\Omega\}, \dots, \mathcal{A}_n \cup \{\Omega\}$ are independent.
$\equiv \mathcal{A}_1 \cup \{\emptyset\}, \dots, \mathcal{A}_n \cup \{\emptyset\}$ are independent.

(b) Independence for collection $\{\mathcal{A}_\theta : \theta \in \Theta\}$ of classes:

$\equiv$ Any collection $\{A_\theta : \theta \in \Theta\}$ of sets is independent where $A_\theta \in \mathcal{A}_\theta$

$\equiv$ Any finite subcollection of classes is independent.

$\equiv \forall$ finite $\Lambda \subset \Theta$, $P\left(\bigcap\limits_{\theta \in \Lambda} A_\theta\right) = \prod\limits_{\theta \in \Lambda} P(A_\theta)$

- By definition, a subcollection of independent events is also independent.
- The class $\{\emptyset, \Omega\}$ is independent from any class.

Example 6.14. The paradox of "almost sure" events: Consider two random events with probabilities of 99% and 99.99%, respectively. One could say that the two probabilities are nearly the same, both events are almost sure to occur. Nevertheless the difference may become significant in certain cases. Consider, for instance, independent events which may occur on any day of the year with probability $p = 99\%$; then the probability P that it will occur every day of the year is less than 3%, while if $p = 99.99\%$ then $P = 97\%$.

7 Random variables

Definition 7.1. A real-valued function $X(\omega)$ defined for points ω in a sample space Ω is called a ***random variable***.

- Random variables are important because they provide a compact way of referring to events via their numerical attributes.
- The abbreviation r.v. will be used for "real-valued random variables" [7, p. 1].
- Technically, a random variable must be *measurable*.

7.2. ***Law*** of X or ***Distribution*** of X: $P^X = \mu_X = PX^{-1} = \mathcal{L}(X) : \underset{(E,\mathcal{E})}{E} \to [0,1]$

$$\begin{aligned}\mu_X(A) = P^X(A) = P\left(X^{-1}(A)\right) &= P \circ X^{-1}(A) \\ &= P(\{\omega : X(\omega) \in A\}) = P([X \in A])\end{aligned}$$

7.3. 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" but it is really there in the background.

7.4. For $B \in \mathbb{R}$, we use the shorthand

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

- $P[X < x]$ is a shorthand for $P(\{\omega \in \Omega : X(\omega) < x\})$.

Theorem 7.5. A random variable can have at most countably many point x such that$P\left[X = x\right] > 0$.

8 Discrete Random Variables

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

$$\sum_k P[X = x_k] = 1.$$

$\equiv$ $\exists$ countable set $\{x_1, x_2, \ldots\}$ such that $P^X\left(\{x_1, x_2, \ldots\}\right) = 1$

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

8.2. When X is a discrete random variable,

- X is completely determined by the values $P\left[X = x_1\right], P\left[X = x_2\right], \ldots$.
- We may define its ***probability mass function*** (pmf) by
$$p_X(x) = P[X = x].$$
  (a) We can use stem plot to visualize p_X.
  (b) 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 in stead of $p_X(x)$.
- $P[X \in B] = \sum_{x_k \in B} P[X = x_k]$.

Definition 8.3. An integer-valued random variable is a discrete random variable whose distinct values are $x_k = k$.

For integer-valued random variables,

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

(b) $F_X(x) = F_X(\lfloor x \rfloor)$.

8.4. Properties of pmf

- $p : \Omega \to [0, 1]$.
- $0 \leq p_X \leq 1$.
- $\sum_k p_X(x_k) = 1$.

8.5. Point masses probability measures / Direc measures, usually written $\varepsilon_\alpha, \delta_\alpha$, is used to denote point mass of size one at the point α. In this case,

- $P^X\left\{\alpha\right\} = 1$
- $P^X\left(\{\alpha\}^c\right) = 0$
- $F_X\left(x\right) = 1_{[\alpha,\infty)}\left(x\right)$

9 CDF: Cumulative Distribution Function

9.1. The (***cumulative***) ***distribution function*** (***cdf***) of the random variable X is the function $F_X(x) = P^X((-\infty, x]) = P[X \leq x]$.

- The distribution P^X can be obtained from the distribution function by setting $P^X(-\infty, x] = F_X(x)$; that is F_X uniquely determines P^X
- $0 \leq F_X \leq 1$

C1 F_X is non-decreasing

C2 F_X is right continuous:

$$\forall x \; F_X\left(x^+\right) \equiv \lim_{\substack{y \to x \\ y > x}} F_X(y) \equiv \lim_{y \searrow x} F_X(y) = F_X(x) = P[X \leq x]$$

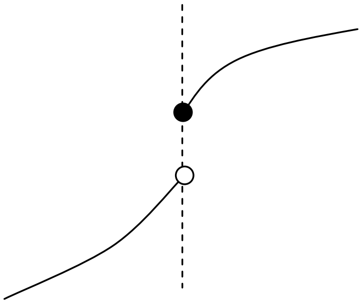

Figure 4: Right-continuous function at jump point

- ○ The function $g(x) = P[X < x]$ is left-continuous in x.

C3 $\lim\limits_{x \to -\infty} F_X(x) = 0$ and $\lim\limits_{x \to \infty} F_X(x) = 1$.

- $\forall x \; F_X(x^-) \equiv \lim\limits_{\substack{y \to x \\ y < x}} F_X(y) \equiv \lim\limits_{y \nearrow x} F_X(y) = P^X(-\infty, x) = P[X < x]$
- $P[X = x] = P^X(\{x\}) = F(x) - F(x^-) =$ the jump or saltus in F at x
- $\forall \; x < y$

$$\begin{aligned} P((x,y]) &= F(y) - F(x) \\ P([x,y]) &= F(y) - F\left(x^-\right) \\ P([x,y)) &= F\left(y^-\right) - F\left(x^-\right) \\ P((x,y)) &= F\left(y^-\right) - F(x) \\ P(\{x\}) &= F(x) - F\left(x^-\right) \end{aligned}$$

- A function F is the distribution function of some random variable if and only if one has (C1), (C2), and (C3).
- F_X is continuous if and only if $P\left[X = x\right] = 0$ for all x.
- F_X has at most countably many points of discontinuity.

Definition 9.2. It is traditional to write $X \sim F$ to indicate that "X has distribution F" [12, p. 25].

9.3. If F is non-decreasing, right continuous, with $\lim\limits_{x \to -\infty} F\left(x\right) = 0$ and $\lim\limits_{x \to \infty} F\left(x\right) = 1$, then F is the CDF of some random variable. See also 18.2.

Definition 9.4 (Discrete CDF)**.** A cdf which can be written in the form $F_d(x) = \sum_k p_i U(x - x_k)$ is called a discrete cdf [5, Def. 5.4.1, p. 163]. Here, U is the unit step function, $\{x_k\}$ is an arbitrary countable set of real numbers, and $\{p_i\}$ is a countable set of positive numbers that sum to 1.

9.1 CDF for Discrete Random Variables

9.5. For discrete random variable X whose support is $\{x_1, x_2, \ldots\}$

$$F_X(x) = \sum_{x_k} p_X(x_k) U(x - x_k).$$

10 Families of Discrete Random Variables

The following pmf will be defined on its support S. For Ω larger than S, we will simply put the pmf to be 0.

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

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

10.1 Random/Uniform

Definition 10.1. We say that X is uniformly distributed on a finite set S if

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

- We write $X \sim \mathcal{U}(S)$ or $X \sim \text{Uniform}(S)$.
- Read "X is uniform on S".

Example 10.2. This pmf is used when the random variable can take finite number of "equally likely" or "totally random" values.

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

10.2 Bernoulli and Binary distributions

Definition 10.3. ***Bernoulli***:

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

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

Definition 10.4. ***Binary***:

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

- X takes only two values: a and b
- X can be expressed as $X = (b-a)\,I + a$, where I is a Bernoulli random variable with $P\left[I = 1\right] = p$.

10.3 Binomial: $\mathcal{B}(n,p)$

Definition 10.5. Binomial pmf with size $n \in \mathbb{N}$ and parameter $p \in (0,1)$

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

- Use `binopdf(x,n,p)` in MATLAB.
- Interpretation: X is the number of successes in n independent Bernoulli trials and hence the sum of n independent, identically distributed Bernoulli r.v.
- Maximum probability value happens at $k_{max} = \operatorname{mode} X = \lfloor (n+1)\,p \rfloor \approx np$
  - When $(n+1)p$ is an integer, then the maximum is achieved at k_{max} and $k_{max}-1$.
- By (2),

$$P[X \text{ is even}] = \frac{1}{2}\left(1 + (1-2p)^n\right), \text{ and}$$
$$P[X \text{ is odd}] = \frac{1}{2}\left(1 - (1-2p)^n\right).$$

10.6. Approximation: When n is large, binomial distribution becomes difficult to compute directly because of the need to calculate factorial terms.

(a) When p is not close to either 0 or 1 so that the variance is also large, we can use

$$P[X=k] \simeq \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}, \tag{6}$$

which comes from approximating X by Gaussian Y with the same mean and variance and the relation

$$\begin{aligned} P[X=k] &\simeq P[X \le k] - P[X \le k-1] \\ &\simeq P[Y \le k] - P[Y \le k-1] \simeq f_Y(k). \end{aligned}$$

See also (**??**).

(b) When p is small, then the binomial can be approximated by $\mathcal{P}(np)$. In particular, suppose X_n has a binomial distribution with parameters n and p_n. If $p_n \to 0$ and $np_n \to \lambda$ as $n \to \infty$, then

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

See also 10.29.

- If p is very close to 1, then $n - X$ will behave approximately Poisson.

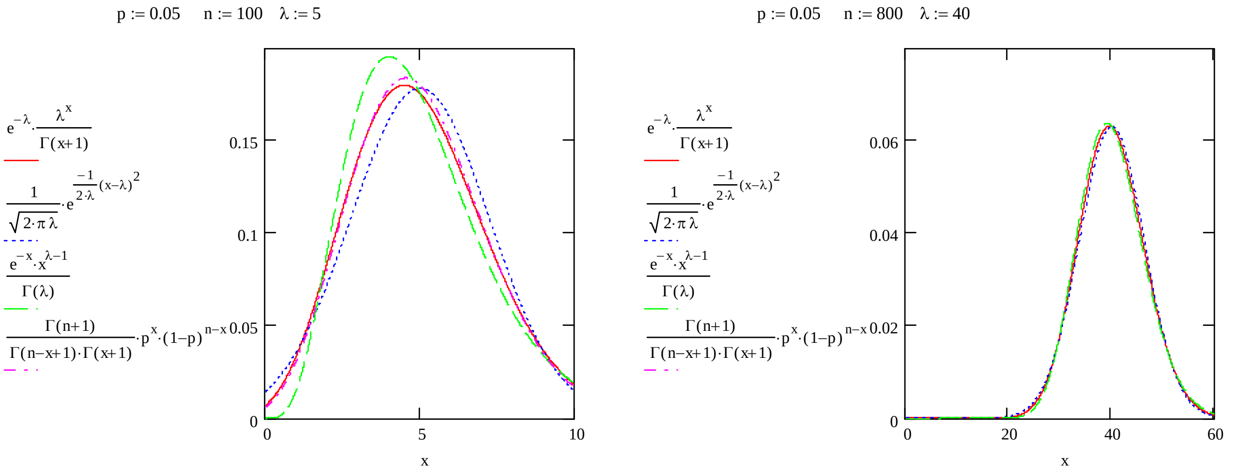

Figure 5: Gaussian approximation to Binomial, Poisson distribution, and Gamma distribution.

10.4 Geometric: $\mathcal{G}(\beta)$

A geometric distribution is defined by the fact that for some $\beta \in [0,1)$, $p_{k+1} = \beta p_k$ for all $k \in S$ where S can be either $\mathbb{N}$ or $\mathbb{N} \cup \{0\}$.

- When its support is $\mathbb{N}$, $p_k = (1-\beta)\,\beta^{k-1}$. This is referred to as $\mathcal{G}_1(\beta)$ or geometric$_1(\beta)$. In MATLAB, use `geopdf(k-1,1-`β`)`.
- When its support is $\mathbb{N} \cup \{0\}$, $p_k = (1-\beta)\,\beta^k$. This is referred to as $\mathcal{G}_0(\beta)$ or geometric$_0(\beta)$. In MATLAB, use `geopdf(k,1-`β`)`.

10.7. Consider $X \sim \mathcal{G}_0(\beta)$.

- $p_i = (1-\beta)\,\beta^i$, for $S = \mathbb{N} \cup \{0\}, 0 \le \beta < 1$.
- $F_x(x) = \begin{cases} 1-\beta^{\lfloor x \rfloor + 1}, & x \ge 0 \\ 0, & x < 0. \end{cases}$.
- $\beta = \frac{m}{m+1}$ where m = average waiting time/ lifetime
- $P[X=k] = P[k$ failures followed by a success$] = (P[\text{failures}])^k\,P[\text{success}]$
  $P[X \ge k] = \beta^k$ = the probability of having at least k initial failure = the probability of having to perform at least k+1 trials.
  $P[X > k] = \beta^{k+1}$ = the probability of having at least k+1 initial failure.
- Memoryless property:
  - $P[X \ge k+c \,|\, X \ge k] = P[X \ge c]$, $k, c > 0$.
  - $P[X > k+c \,|\, X \ge k] = P[X > c]$, $k, c > 0$.
  - If a success has not occurred in the first k trials (already fails for k times), then the probability of having to perform at least j more trials is the same the probability of initially having to perform at least $j$trials.

  - Each time a failure occurs, the system “forgets” and begins anew as if it were performing the first trial.
  - Geometric r.v. is the only discrete r.v. that satisfies the memoryless property.

- Ex.

  - lifetimes of components, measured in discrete time units, when the fail catastrophically (without degradation due to aging)
  - waiting times
    - for next customer in a queue
    - between radioactive disintegrations
    - between photon emission
  - number of repeated, unlinked random experiments that must be performed prior to the first occurrence of a given event A
    - number of coin tosses prior to the first appearance of a ‘head’
      number of trials required to observe the first success

- The sum of independent $\mathcal{G}_0(p)$ and $\mathcal{G}_0(q)$ has pmf

$$\begin{cases} (1-p)\,(1-q)\,\frac{q^{k+1}-p^{k+1}}{q-p}, & p \neq q \\ (k+1)\,(1-p)^2\,p^k, & p = q \end{cases}$$

  for $k \in \mathbb{N} \cup \{0\}$.

10.8. Consider $X \sim \mathcal{G}_1(\beta)$.

- $P[X > k] = \beta^k$
- Suppose independent $X_i \sim \mathcal{G}_1(\beta_i)$. $\min(X_1, X_2, \dots, X_n) \sim \mathcal{G}_1\left(\prod_{i=1}^{n} \beta_i\right)$.

10.5 Poisson Distribution: $\mathcal{P}(\lambda)$

10.9. Characterized by

- $p_X(k) = P[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}$; or equivalently,
- $\varphi_X(u) = e^{\lambda\left(e^{iu}-1\right)}$,

where $\lambda \in (0, \infty)$ is called the ***parameter*** or ***intensity parameter*** of the distribution. In `MATLAB`, use `poisspdf(k,lambda)`.

10.10. Denoted by $\mathcal{P}(\lambda)$.

10.11. In stead of X, Poisson random variable is usually denoted by Λ.

10.12. $\mathbb{E}X = \operatorname{Var} X = \lambda$.

	Most probable value ($i_{\max}$)	Associated max probability
$0 < \lambda < 1$	0	$e^{-\lambda}$
$\lambda \in \mathbb{N}$	$\lambda - 1, \lambda$	$\frac{\lambda^\lambda}{\lambda!} e^{-\lambda}$
$\lambda \geq 1, \lambda \notin \mathbb{N}$	$\lfloor \lambda \rfloor$	$\frac{\lambda^{\lfloor \lambda \rfloor}}{\lfloor \lambda \rfloor !} e^{-\lambda}$

10.13. Successive probabilities are connected via the relation $kp_X(k) = \lambda p_X(k-1)$.

10.14. $\operatorname{mode} X = \lfloor \lambda \rfloor$.

- Note that when $\lambda \in \mathbb{N}$, there are two maximums at $\lambda - 1$ and λ.
- When $\lambda \gg 1$, $p_X\left(\lfloor \lambda \rfloor\right) \approx \frac{1}{\sqrt{2\pi\lambda}}$ via the Stirling's formula (4.5).

10.15. $P\left[X \geq 2\right] = 1 - e^{-\lambda} - \lambda e^{-\lambda} = O\left(\lambda^2\right)$.
The cumulative probabilities can be found by

$$P\left[X \leq k\right] \overset{(*)}{=} P\left[\sum_{i=1}^{k+1} X_i > 1\right] = \frac{1}{\Gamma\left(k+1\right)} \int\limits_{\lambda}^{\infty} e^{-t} t^x dt,$$

$$P\left[X > k\right] = P\left[X \geq k+1\right] \overset{(*)}{=} P\left[\sum_{i=1}^{k+1} X_i \leq 1\right] = \frac{1}{\Gamma\left(k+1\right)} \int\limits_{0}^{\lambda} e^{-t} t^x dt,$$

where the X_i's are i.i.d. $\mathcal{E}(\lambda)$. The equalities given by (*) are easily obtained via counting the number of events from rate-λ Poisson process on interval $[0,1]$.

10.16. Fano factor (index of dispersion): $\frac{\operatorname{Var} X}{\mathbb{E}X} = 1$

An important property of the Poisson and Compound Poisson laws is that their classes are close under convolution (independent summation). In particular, we have divisibility properties (10.21) and (**??**) which are straightforward to prove from their characteristic functions.

10.17 (Recursion equations)**.** Suppose $X \sim \mathcal{P}(\lambda)$. Let $m_k(\lambda) = \mathbb{E}\left[X^k\right]$ and $\mu_k(\lambda) = \mathbb{E}\left[(X - \mathbb{E}X)^k\right]$.

$$m_{k+1}(\lambda) = \lambda\left(m_k(\lambda) + m_k'(\lambda)\right) \tag{7}$$

$$\mu_{k+1}(\lambda) = \lambda\left(k\mu_{k-1}(\lambda) + \mu_k'(\lambda)\right) \tag{8}$$

[**?**, p 112]. Starting with $m_1 = \lambda = \mu_2$ and $\mu_1 = 0$, the above equations lead to recursive determination of the moments m_k and μ_k.

10.18. $\mathbb{E}\left[\frac{1}{X+1}\right] = \frac{1}{\lambda}\left(1 - e^{-\lambda}\right)$. This result can be used to find the expectation of $\frac{\text{polynomial in } X}{X+1}$. To see this, note that for $d \in \mathbb{N}$, $Y = \frac{1}{X+1}\left(\sum_{n=0}^{d} a_n X^n\right)$ can be expressed as $\left(\sum_{n=0}^{d-1} b_n X^n\right) + \frac{c}{X+1}$, the value of $\mathbb{E}Y$ is easy to find if we know $\mathbb{E}X^n$.

10.19. Mixed Poisson distribution: Let X be Poisson with mean λ. Suppose, that the mean λ is chosen in accord with a probability distribution whose characteristic function is φ_Λ. Then,

$$\varphi_X(u) = \mathbb{E}\left[\mathbb{E}\left[e^{iuX}\left|\Lambda\right.\right]\right] = \mathbb{E}\left[e^{\Lambda\left(e^{iu}-1\right)}\right] = \mathbb{E}\left[e^{i\left(-i\left(e^{iu}-1\right)\right)\Lambda}\right] = \varphi_\Lambda\left(-i\left(e^{iu}-1\right)\right).$$

- $\mathbb{E}X = \mathbb{E}\Lambda$.
- $\operatorname{Var} X = \operatorname{Var}\Lambda + \mathbb{E}\Lambda$.
- $\mathbb{E}\left[X^2\right] = \mathbb{E}\left[\Lambda^2\right] + \mathbb{E}\Lambda$.
- $\operatorname{Var}[X|\Lambda] = \mathbb{E}\left[X|\Lambda\right] = \Lambda$.
- When Λ is a nonnegative integer-valued random variable, we have $G_X(z) = G_\Lambda\left(e^{z-1}\right)$ and $P[X=0] = G_\Lambda\left(\frac{1}{z}\right)$.
- $\mathbb{E}\left[X\Lambda\right] = \mathbb{E}\left[\Lambda^2\right]$
- $\operatorname{Cov}\left[X,\Lambda\right] = \operatorname{Var}\Lambda$

10.20. Thinned Poisson: Suppose we have $X \to \boxed{s} \to Y$ where $X \sim \mathcal{P}(\lambda)$. The box $\boxed{s}$ is a binomial channel with success probability s. (Each 1 in the X get through the channel with success probability s.)

- Note that Y is in fact a random sum $\sum_{i=1}^{X} I_i$ where i.i.d. I_i has Bernoulli distribution with parameter s.
- $Y \sim \mathcal{P}(s\lambda)$;
- $p\left(x\left|y\right.\right) = e^{-\lambda(1-s)}\frac{(\lambda(1-s))^{x-y}}{(x-y)!}$; $x \ge y$ (shifted Poisson);

[Levy and Baxter, 2002]

10.21. Finite additivity: Suppose we have independent $\Lambda_i \sim \mathcal{P}(\lambda_i)$, then $\sum_{i=1}^{n}\Lambda_i \sim \mathcal{P}\left(\sum_{i=1}^{n}\lambda_i\right)$.

10.22. Raikov's theorem: independent random variables can have their sum Poisson-distributed only if every component of the sum is Poisson-distributed.

10.23. Countable Additivity Theorem [?, p 5]: Let $(X_j : j = \in \mathbb{N})$ be independent random variables, and assume that X_j has the distribution $\mathcal{P}(\mu_j)$ for each j. If

$$\sum_{j=1}^{\infty}\mu_j \tag{9}$$

converges to μ, then $S = \sum\limits_{j=1}^{\infty} X_j$ converges with probability 1, and S has distribution $\mathcal{P}(\mu)$.

If on the other hand (9) diverges, then S diverges with probability 1.

10.24. Let $X_1, X_2, \dots, X_n$ be independent, and let X_j have distribution $\mathcal{P}(\mu_j)$ for all j. Then $S_n = \sum_{j=1}^n X_j$ has distribution $\mathcal{P}(\mu)$, with $\mu = \sum_{j=1}^n \mu_j$; and so, whenever $\sum_{j=1}^n r_j = s$,

$$P\left[X_j = r_j \ \forall j | S_n = s\right] = \frac{s!}{r_1! r_2! \cdots r_n!} \prod_{j=1}^{n} \left(\frac{\mu_j}{\mu}\right)^{r_j}$$

which follows the multinomial distribution [**?**, p 6–7].

- If X and Y are independent Poisson random variables with respective parameters λ and μ, then (1) $Z = X + Y$ is $\mathcal{P}(\lambda + \mu)$ and (2) conditioned on $Z = z$, X is $\mathcal{B}\left(z, \frac{\lambda}{\lambda+\mu}\right)$. So, $\mathbb{E}\left[X|Z\right] = \frac{\lambda}{\lambda+\mu} Z$, $\text{Var}[X|Z] = Z\frac{\lambda\mu}{(\lambda+\mu)^2}$, and $\mathbb{E}\left[\text{Var}[X|Z]\right] = \frac{\lambda\mu}{\lambda+\mu}$.

10.25. One of the reasons why Poisson distribution is important is because many natural phenomenons can be modeled by Poisson processes. For example, if we consider the number of occurrences Λ during a time interval of length τ in a rate-λ homogeneous Poisson process, then $\Lambda \sim \mathcal{P}(\lambda\tau)$.

Example 10.26.

- The first use of the Poisson model is said to have been by a Prussian (German) physician, von Bortkiewicz, who found that the annual number of late-19th-century Prussian (German) soldiers kicked to death by horses fitted a Poisson distribution [5, p 150],[2, Ex 2.23][6].
- #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
- number of occurrences of rare events in time τ
- #soldiers kicked to death by horses
- Counts of defects in a semiconductor chip.

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

10.27. Normal Approximation to Poisson Distribution with large λ: Let $X \sim \mathcal{P}(\lambda)$. X can be though of as a sum of i.i.d.$X_i \sim \mathcal{P}(\lambda_n)$, i.e., $X = \sum_{i=1}^{n} X_i$, where $n\lambda_n = \lambda$. Hence X is approximately normal $\mathcal{N}(\lambda, \lambda)$ for λ large.

Some says that the normal approximation is good when $\lambda > 5$.

10.28. Poisson distribution can be obtained as a limit from negative binomial distributions. Thus, the negative binomial distribution with parameters r and p can be approximated by the Poisson distribution with parameter $\lambda = \frac{rq}{p}$ (mean-matching), provided that p is "sufficiently" close to 1 and r is "sufficiently" large.

10.29. ***Convergence of sum of bernoulli random variables to the Poisson Law*** Suppose that for each $n \in \mathbb{N}$

$$X_{n,1}, X_{n,2}, \ldots, X_{n,r_n}$$

are independent; the probability space for the sequence may change with n. Such a collection is called a ***triangular array*** [?] or **double sequence** [?] which captures the nature of the collection when it is arranged as

$$\left.\begin{array}{cccc} X_{1,1}, & X_{1,2}, & \ldots, & X_{1,r_1}, \\ X_{2,1}, & X_{2,2}, & \ldots, & X_{2,r_2}, \\ \vdots & \vdots & \cdots & \vdots \\ X_{n,1}, & X_{n,2}, & \ldots, & X_{n,r_n}, \\ \vdots & \vdots & \cdots & \vdots \end{array}\right\}$$

where the random variables in each row are independent. Let $S_n = X_{n,1} + X_{n,2} + \cdots + X_{n,r_n}$ be the sum of the random variables in the n^{th} row.

Consider a triangular array of bernoulli random variables $X_{n,k}$ with $P[X_{n,k} = 1] = p_{n,k}$. If $\max\limits_{1 \le k \le r_n} p_{n,k} \to 0$ and $\sum_{k=1}^{r_n} p_{n,k} \to \lambda$ as $n \to \infty$, then the sums S_n converges in distribution to the Poisson law. In other words, Poisson distribution is a rare-event limit of the binomial (large n, small p).

As a simple special case, consider a triangular array of bernoulli random variables $X_{n,k}$ with $P[X_{n,k} = 1] = p_n$. If $np_n \to \lambda$ as $n \to \infty$, then the sums S_n converges in distribution to the Poisson law.

To show this special case directly, we bound the first i terms of $n!$ to get $\frac{(n-i)^i}{i!} \le \binom{n}{i} \le \frac{n^i}{i!}$. Using the upper bound,

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

The lower bound gives the same limit because $(n-i)^i = \left(\frac{n-i}{n}\right)^i n^i$ where the first term $\to 1$.

10.30. General Poisson approximation result: Consider independent events A_i, $i = 1, 2, \ldots, n$, with probabilities $p_i = P(A_i)$. Let N be the number of events that occur, let

$\lambda = p_1 + \ldots + p_n$, and let Λ have a Poisson distribution with parameter λ. Then, for any set of integers B,

$$|P\left[N \in B\right] - P\left[\Lambda \in B\right]| \le \sum_{i=1}^{n} p_i^2 \tag{10}$$

[2, Theorem 2.5, p. 52] We can simplify the RHS of (10) by noting

$$\sum_{i=1}^{n} p_i^2 \le \max_i p_i \sum_{i=1}^{n} p_i = \lambda \max_i p_i.$$

This says that if all the p_i are small then the distribution of Λ is close to a Poisson with parameter λ. Taking $B = \{k\}$, we see that the individual probabilities $P\left[N = k\right]$ are close to $P\left[\Lambda = k\right]$, but this result says more. The probabilities of events such as $P\left[3 \le N \le 8\right]$ are close to $P\left[3 \le \Lambda \le 8\right]$ and we have an explicit bound on the error.

11 Multiple Random Variables

11.1. If X and Y are random variables, we use the shorthand

- $[X \in B, Y \in C] = \{\omega \in \Omega : X(\omega) \in B \text{ and } Y(\omega) \in C\} = [X \in B] \cap [Y \in C]$.
- $P[X \in B, Y \in C] = P\left([X \in B] \cap [Y \in C]\right)$.

11.2. When X and Y take finitely many values, say $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, respectively, we can arrange the probabilities $p_{X,Y}(x_i, y_j)$ in the $m \times n$ matrix

$$\begin{bmatrix} p_{X,Y}(x_1, y_1) & p_{X,Y}(x_1, y_2) & \ldots & p_{X,Y}(x_1, y_n) \\ p_{X,Y}(x_2, y_1) & p_{X,Y}(x_2, y_2) & \ldots & p_{X,Y}(x_2, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ p_{X,Y}(x_m, y_1) & p_{X,Y}(x_m, y_2) & \ldots & p_{X,Y}(x_m, y_n) \end{bmatrix}.$$

- The sum of the entries in the ith row is $P_X(x_i)$, and the sum of the entries in the jth column is $P_Y(y_j)$.
- The sum of all the entries in the matrix is one.

Definition 11.3. A family of random variables $\{X_i : i \in I\}$ is ***independent*** if $\forall$ finite $J \subset I$, the family of random variables $\{X_i : i \in J\}$ is independent. In words, "an infinite collection of random elements is by definition independent if each finite subcollection is." Hence, we only need to know how to test independence for finite collection.

Definition 11.4. A ***pairwise independent*** collection of random variables is a set of random variables any two of which are independent.

(a) Any collection of (mutually) independent random variables is pairwise independent

(b) Some pairwise independent collections are not independent. See Example (11.5).

Example 11.5. Let suppose X, Y, and Z have the following joint probability distribution: $p_{X,Y,Z}(x,y,z) = \frac{1}{4}$ for $(x,y,z) \in \{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}$. This, for example, can be constructed by starting with independent X and Y that are Bernoulli-$\frac{1}{2}$. Then set $Z = X \oplus Y = X + Y \mod 2$.

(a) X, Y, Z are pairwise independent.

(b) The combination of $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$ does not imply $(X,Y) \perp\!\!\!\perp Z$.

12 Bernoulli Trial

13 Expectation of Discrete Random Variable

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

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

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

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

Definition 13.1. Expectation of a discrete random variable: Suppose x is a discrete random variable, we define the **expectation** of X by

$$\mathbb{E}X = \sum_x xP[X = x] = \sum_x x\ p_X(x). \tag{11}$$

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

- Some references use m_X or μ_X to represent $\mathbb{E}X$.
- For conciseness, we simply write x under the summation symbol in (11); this means that the sum runs over all possible x values. However, there will be many points x which has $p_X(x) = 0$. (By definition of begin discrete random variable, there are only countably many points x that can have $p_X(x) > 0$.) For those zero-probability points, their contributions, which is$x \times p_X(x)$ will be zero anyway and hence, we don't have to care about them.
- Definition (11) is only meaningful if the sum is well defined.
  - ○ The sum of infinitely many nonnegative terms is always well-defined, with $+\infty$ as a possible value for the sum.

- Some care is necessary when computing expectations of signed random variables that take more than finitely many values. The sum over countably infinite many terms is not always well defined when both positive and negative terms are involved.
  - ○ For example, the infinite series $1-1+1-1+\dots$ has the sum 0 when you sum the terms according to $(1-1)+(1-1)+\cdots$, whereas you get the sum 1 when you sum the terms according to $1+(-1+1)+(-1+1)+(-1+1)+\cdots$. Such abnormalities cannot happen when all terms in the infinite summation are nonnegative.

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

$$\mathbb{E}X = \sum_{i:x_i\geq 0} x_i p_X(x_i) + \sum_{i:x_i<0} x_i p_X(x_i),$$

  assuming that at least one of these sums is finite.

  - ○ If the first sum is $+\infty$ and the second one is $-\infty$, then no value is assigned to $\mathbb{E}X$, and we say that $\mathbb{E}X$ is **undefined**.

Example 13.2. When $X \sim \text{Bernoulli}(p)$ with $p \in (0,1)$, $\mathbb{E}X = 0\times(1-p)+1\times p = p$. Note that, since X takes only the values 0 and 1, its "typical" value p is never seen.

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

Example 13.4. $p_X(x) = \begin{cases} 1/4, & x=0 \\ 3/4, & x=2 \\ 0, & \text{otherwise} \end{cases}$

Example 13.5. For $X \sim \mathcal{P}(\alpha)$,

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

14 Function of Discrete Random Variables

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

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

Example 14.2. Let $p_X(x) = \begin{cases} \frac{1}{c}x^2, & x = \pm 1, \pm 2 \\ 0, & \text{otherwise} \end{cases}$ and $Y = X^4$. Find $p_Y(y)$ and then calculate $\mathbb{E}Y$.

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

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

If we want to compute $\mathbb{E}Y$, it might seem that we first have to find the pmf of Y. Typically, this requires a detailed analysis of g. However, we can compute $\mathbb{E}Y = \mathbb{E}[g(X)]$ without actually finding the pmf of Y.

Example 14.4. When $Z = X + Y$,

$$p_Z(z) = \sum_{(x,y):x+y=z} p_{X,Y}(x,y) = \sum_y p_{X,Y}(z-y,y) = \sum_x p_{X,Y}(x,z-x).$$

Furthermore, if X and Y are independent,

$$p_Z(z) = \sum_{(x,y):x+y=z} p_X(x)\, p_Y(y) = \sum_y p_X(z-y)\, p_Y(y) = \sum_x p_X(x)\, p_Y(z-x).$$

Example 14.5. Set $S = \{0, 1, 2, \dots, M\}$, then the sum of two i.i.d. uniform discrete random variables on S has pmf

$$p(k) = \frac{(M+1) - |k - M|}{(M+1)^2}$$

for $k = 0, 1, \dots, 2M$. Note its triangular shape with maximum value at $p(M) = \frac{1}{M+1}$. To visualize the pmf in MATLAB, try

```
k = 0:2*M;
P = (1/((M+1)^2))*ones(1,M+1);
P = conv(P,P); stem(k,P)
```

Example 14.6. Suppose $\Lambda_1 \sim \mathcal{P}(\lambda_1)$ and $\Lambda_2 \sim \mathcal{P}(\lambda_2)$ are independent. Find the pmf of $\Lambda = \Lambda_1 + \Lambda_2$.

First, note that $p_\Lambda(x)$ would be positive only on nonnegative integers because a sum of nonnegative integers is still a nonnegative integer. So, the support of Λ is the same as the support for Λ_1 and Λ_2. Now, we know that

$$P\left[\Lambda = k\right] = P\left[\Lambda_1 + \Lambda_2 = k\right] = \sum_i P\left[\Lambda_1 = i\right] P\left[\Lambda_2 = k - i\right]$$

Of course, we are interested in k that is a nonnegative integer. The summation runs over $i = 0, 1, 2, \dots$. Other values of i would make $P[\Lambda_1 = i] = 0$. Note also that if $i > k$, then $k - i < 0$ and $P\left[\Lambda_2 = k - i\right] = 0$. Hence, we conclude that the index i can only be integers from 0 to k:

$$\begin{aligned} P\left[\Lambda = k\right] &= \sum_{i=0}^{k} e^{-\lambda_1} \frac{\lambda_1^i}{i!} e^{-\lambda_2} \frac{\lambda_2^{k-i}}{(k-i)!} = e^{-(\lambda_1+\lambda_2)} \frac{\lambda_2^k}{k!} \sum_{i=0}^{k} \frac{k!}{i!\,(k-i)!} \left(\frac{\lambda_1}{\lambda_2}\right)^i \\ &= e^{-(\lambda_1+\lambda_2)} \frac{\lambda_2^k}{k!} \sum_{i=0}^{k} \binom{k}{i} \left(\frac{\lambda_1}{\lambda_2}\right)^i = e^{-(\lambda_1+\lambda_2)} \frac{\lambda_2^k}{k!} \left(1 + \frac{\lambda_1}{\lambda_2}\right)^k \\ &= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^k}{k!} \end{aligned}$$

Hence, the sum of two independent Poisson random variables is still Poisson!

15 Expectation of function of discrete random variables

15.1. Suppose x is a discrete random variable.

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

Similarly,

$$\mathbb{E}\left[g(X,Y)\right] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y).$$

These are called the **law/rule of the lazy statistician** (LOTUS) [12, Thm 3.6 p 48],[6, p. 149] because it is so much easier to use the above formula than to first find the pmf of Y. It is also called **substitution rule** [11, p 271].

Example 15.2. Back to Example 14.2.

Example 15.3. Let $p_X(x) = \begin{cases} \frac{1}{c}x^2, & x = \pm 1, \pm 2 \\ 0, & \text{otherwise} \end{cases}$ as in Example 14.2. Find

(a) $\mathbb{E}[X^2]$

(b) $\mathbb{E}[|X|]$

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

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

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

(b) An exception is the case of a linear function $g(x) = ax + b$. See also (15.8).

Example 15.5. For $X \sim$ Bernoulli(p),

(a) $\mathbb{E}X = p$

(b) $\mathbb{E}X^2 = 0^2 \times (1-p) + 1^2 \times p = p \neq (\mathbb{E}X)^2$.

Example 15.6. Continue from Example 13.4. Suppose $X \sim \mathcal{P}(\alpha)$.

$$\mathbb{E}\left[X^2\right] = \sum_{i=0}^{\infty} i^2 e^{-\alpha}\frac{\alpha^i}{i!} = e^{-\alpha}\alpha \sum_{i=0}^{\infty} i \frac{\alpha^{i-1}}{(i-1)!} \tag{12}$$

We can evaluate the infinite sum in (12) by rewriting i as $i-1+1$:

$$\begin{aligned}\sum_{i=1}^{\infty} i\frac{\alpha^{i-1}}{(i-1)!} &= \sum_{i=1}^{\infty}(i-1+1)\frac{\alpha^{i-1}}{(i-1)!} = \sum_{i=1}^{\infty}(i-1)\frac{\alpha^{i-1}}{(i-1)!} + \sum_{i=1}^{\infty}\frac{\alpha^{i-1}}{(i-1)!} \\ &= \alpha\sum_{i=2}^{\infty}\frac{\alpha^{i-2}}{(i-2)!} + \sum_{i=1}^{\infty}\frac{\alpha^{i-1}}{(i-1)!} = \alpha e^{\alpha} + e^{\alpha} = e^{\alpha}(\alpha+1).\end{aligned}$$

Plugging this back into (12), we get

$$\mathbb{E}\left[X^2\right] = \alpha\left(\alpha+1\right) = \alpha^2+\alpha.$$

15.7. For discrete random variables X and Y, the pmf of a derived random variable

$$Z = g(X,Y)$$

is given by

$$p_Z(z) = \sum_{(x,y):g(X,Y)=z} p_{X,Y}(x,y).$$

15.8. Basic Properties of Expectations

(a) For $c \in \mathbb{R}$, $\mathbb{E}\left[c\right] = c$

(b) For $c \in \mathbb{R}$, $\mathbb{E}\left[X+c\right] = \mathbb{E}X + c$ and $\mathbb{E}\left[cX\right] = c\mathbb{E}X$

(c) For constants a, b, we have $\mathbb{E}\left[aX+b\right] = a\mathbb{E}X + b$.

(d) $\mathbb{E}\left[\cdot\right]$ is a **linear** operator: $\mathbb{E}\left[aX+bY\right] = a\mathbb{E}X + b\mathbb{E}Y$.

- (i) Homogeneous: $\mathbb{E}\left[cX\right] = c\mathbb{E}X$
- (ii) Additive: $\mathbb{E}\left[X+Y\right] = \mathbb{E}X + \mathbb{E}Y$
- (iii) Extension: $\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}X_i$.

(e) $\mathbb{E}\left[X - \mathbb{E}X\right] = 0$.

Example 15.9. A binary communication link has bit-error probability p. What is the expected number of bit errors in a transmission of n bits.

Recall that when iid $X_i \sim$ Bernoulli(p), $Y = X_1 + X_2 + \cdots X_n$ is Binomial(n,p). Also, from Example 13.2, we have $\mathbb{E}X_i = p$. Hence,

$$\mathbb{E}Y = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \sum_{i=1}^{n} p = np.$$

Therefore, the expectation of a binomial random variable with parameters n and p is np.

Definition 15.10. Some definitions involving expectation of a function of a random variable:

(a) ***Absolute moment***: $\mathbb{E}\left[|X|^k\right]$, where we define $\mathbb{E}\left[|X|^0\right] = 1$

(b) ***Moment***: $m_k = \mathbb{E}\left[X^k\right]$ = the k^{th} moment of X, $\quad k \in \mathbb{N}$.

- The first moment of X is its expectation $\mathbb{E}X$.
- The second moment of X is $\mathbb{E}\left[X^2\right]$.

Recall that an average (expectation) can be regarded as one number that summarizes an entire probability model. After finding an average, someone who wants to look further into the probability model might ask, "How typical is the average?" or, "What are the chances of observing an event far from the average?" A measure of dispersion is an answer to these questions wrapped up in a single number. If this measure is small, observations are likely to be near the average. A high measure of dispersion suggests that it is not unusual to observe events that are far from the average.

Example 15.11. Consider your score on the midterm exam. After you find out your score is 7 points above average, you are likely to ask, "How good is that? Is it near the top of the class or somewhere near the middle?".

Definition 15.12. The most important ***measures of dispersion*** are the standard deviation and its close relative, the variance.

(a) ***Variance***:

$$\operatorname{Var} X = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]. \tag{13}$$

- Read "the variance of X"
- *Notation*: D_X, or $\sigma^2(X)$, or σ_X^2, or $\mathbb{V}X$ [12, p. 51]
- In some references, to avoid confusion from the two expectation symbols, they first define $m = \mathbb{E}X$ and then define the variance of X by

$$\operatorname{Var} X = \mathbb{E}\left[(X - m)^2\right].$$

- We can also calculate the variance via another identity:

$$\operatorname{Var} X = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2$$

- The units of the variance are squares of the units of the random variable.
- $\operatorname{Var} X \geq 0$.

- $\operatorname{Var} X \leq \mathbb{E}\left[X^2\right]$.
- $\operatorname{Var}[cX] = c^2 \operatorname{Var} X$.
- $\operatorname{Var}[X + c] = \operatorname{Var} X$.
- $\operatorname{Var} X = \mathbb{E}\left[X(X - \mathbb{E}X)\right]$

(b) ***Standard Deviation***: $\sigma_X = \sqrt{\operatorname{Var}[X]}$.

- One uses the standard deviation, which is defined as the square root of the variance to measure of the *spread* of the possible values of X.
- It is useful to work with the standard deviation since it has the same units as $\mathbb{E}X$.
- $\sigma_{cX} = |c|\,\sigma_X$.
- Informally we think of outcomes within $\pm\sigma_X$ of $\mathbb{E}X$ as being in the center of the distribution. Some references would informally interpret sample values within $\pm\sigma_X$ of the expected value, $x \in [\mathbb{E}X - \sigma_X, \mathbb{E}X + \sigma_X]$, as "typical" values of X and other values as "unusual".

Example 15.13. Continue from Example 15.11. If the standard deviation of exam scores is 12 points, the student with a score of +7 with respect to the mean can think of herself in the middle of the class. If the standard deviation is 3 points, she is likely to be near the top.

Example 15.14. Suppose $X \sim \text{Bernoulli}(p)$.

(a) $\mathbb{E}X^2 = 0^2 \times (1 - p) + 1^2 \times p = p$.

(b) $\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2 = p(1 - p)$.
Alternatively, if we directly use (13), we have

$$\begin{aligned}\operatorname{Var} X = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] &= (0 - p)^2 \times (1 - p) + (1 - p)^2 \times p \\ &= p(1 - p)(p + (1 - p)) = p(1 - p).\end{aligned}$$

Example 15.15. Continue from Example 13.4 and Example 15.6. Suppose $X \sim \mathcal{P}(\alpha)$. We have

$$\operatorname{Var} X = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2 = \alpha^2 + \alpha - \alpha^2 = \alpha.$$

Therefore, for Poisson random variable, the expected value is the same as the variance.

Example 15.16. For X uniform on [N:M] (the set of integers from N to M), we have

$$\mathbb{E}X = \frac{M+N}{2}$$

and

$$\operatorname{Var} X = \frac{1}{12}(M-N)(M-N-2) = \frac{1}{12}(n^2-1),$$

where $n = M - N + 1$.

- For X uniform on [-M:M], we have $\mathbb{E}X = 0$ and $\operatorname{Var} X = \frac{M(M+1)}{3}$.

Example 15.17. Suppose X is a binary random variable with

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

(a) $\operatorname{Var} X = (b-a)^2 \operatorname{Var} I = (b-a)^2 p\,(1-p)\,.$

(b) Suppose $a = -b$. Then, $X = 2I + a = 2I - b$. In which case, $\operatorname{Var} X = 2b^2 p\,(1-p)$.

15.18. ***Markov's Inequality***: $P\left[|X| \geq a\right] \leq \frac{1}{a}\mathbb{E}\,|X|$, $a > 0$.

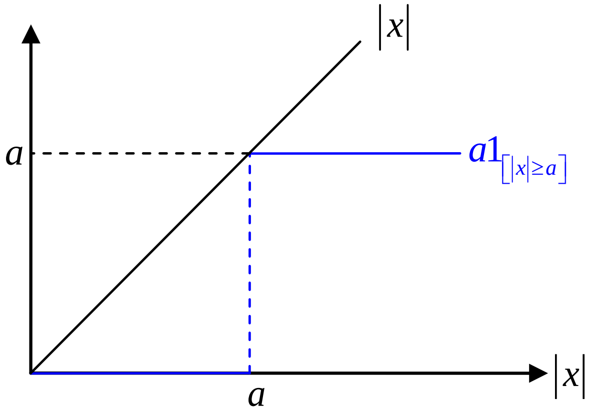

Figure 6: Proof of Markov's Inequality

(a) Useless when $a \leq \mathbb{E}\,|X|$. Hence, good for bounding the "tails" of a distribution.

(b) Remark: $P\left[|X| > a\right] \leq P\left[|X| \geq a\right]$

(c) $P\left[|X| \geq a\mathbb{E}\,|X|\right] \leq \frac{1}{a}, a > 0.$

(d) ***Chebyshev's Inequality***: $P[|X| > a] \le P[|X| \ge a] \le \frac{1}{a^2}\mathbb{E}X^2$, $a > 0$.

  (i) $P[|X - \mathbb{E}X| \ge \alpha] \le \frac{\sigma_X^2}{\alpha^2}$; that is $P[|X - \mathbb{E}X| \ge n\sigma_X] \le \frac{1}{n^2}$
  - Useful only when $\alpha > \sigma_X$

Definition 15.19. More definitions involving expectation of a function of a random variable:

(a) ***Coefficient of Variation***: $CV_X = \frac{\sigma_X}{\mathbb{E}X}$.
  - It is the standard deviation of the "normalized" random variable $\frac{X}{\mathbb{E}X}$.
  - 1 for exponential.

(b) ***Fano Factor*** (index of dispersion): $\frac{\operatorname{Var} X}{\mathbb{E}X}$.
  - 1 for Poisson.

(c) ***Central Moments***: A generalization of the variance is the nth central moment which is defined to be $\mu_n = \mathbb{E}[(X - \mathbb{E}X)^n]$.

  (i) $\mu_1 = \mathbb{E}[X - \mathbb{E}X] = 0$.

  (ii) $\mu_2 = \sigma_X^2 = \operatorname{Var} X$: the second central moment is the variance.

Example 15.20. If X has mean m and variance σ^2, it is sometimes convenient to introduce the normalized random variable

$$Y = \frac{X - m}{\sigma}.$$

Theorem 15.21 (Expectation and Independence)**.** Two random variables X and Y are independent if and only if

$$\mathbb{E}[h(X)g(Y)] = \mathbb{E}[h(X)]\,\mathbb{E}[g(Y)]$$

for all functions h and g.

- In other words, X and Y are independent if and only if for every pair of functions h and g, the expectation of the product $h(X)g(Y)$ is equal to the product of the individual expectations.

- One special case is that
$$\mathbb{E}[XY] = (\mathbb{E}X)(\mathbb{E}Y). \tag{14}$$
However, independence means more than this property. It is possible that (14) is true while X and Y are not independent. See Example 15.23

Definition 15.22. Some definitions involving expectation of a function of two random variables:

- **Correlation** between X and Y: $\mathbb{E}[XY]$.
- **Covariance** between X and Y:
$$\begin{aligned}\operatorname{Cov}[X,Y] &= \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y \\ &= \mathbb{E}[X(Y-\mathbb{E}Y)] = \mathbb{E}[Y(X-\mathbb{E}X)].\end{aligned}$$

- Note that $\operatorname{Var} X = \operatorname{Cov}[X,X]$.
- X and Y are said to be ***uncorrelated*** if and only if $\operatorname{Cov}[X,Y]=0$.

  $\equiv \mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y$

  - If $X \perp\!\!\!\perp Y$, then $\operatorname{Cov}[X,Y]=0$. The converse is not true.
- X and Y are said to be ***orthogonal*** if $\mathbb{E}[XY]=0$.
  - When $\mathbb{E}X=0$ or $\mathbb{E}Y=0$, orthogonality is equivalent to uncorrelatedness.

Example 15.23. Being uncorrelated does not imply independence:

(a) Suppose two fair dice are tossed. Denote by the random variable V_1 the number appearing on the first die and by the random variable V_2 the number appearing on the second die. Let $X = V_1 + V_2$ and $Y = V_1 - V_2$. It is readily seen that the random variables X and Y are not independent. You may verify that $\mathbb{E}X = 7$, $\mathbb{E}Y = 0$, and $\mathbb{E}[XY] = 0$ and so $\mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y$.

(b) Let X be uniform on $\{\pm 1, \pm 2\}$ and $Y = |X|$. Consider the point $x_0 = 1$.

Remark: This example can be generalized as followed. Suppose p_X is an even function with $p_X(0) = 0$. Let $Y = g(X)$ where g is also an even function. Then, $\mathbb{E}[XY] = \mathbb{E}[X] = \mathbb{E}[X]\mathbb{E}[Y] = \operatorname{Cov}[X,Y] = 0$. Consider a point x_0 such that $p_X(x_0) > 0$. Then,
$$p_{X,Y}(x_0, g(x_0)) = p_X(x_0).$$
We only need to show that $p_Y(g(x_0)) \neq 1$ to show that X and Y are not independent.

Definition 15.24. ***Correlation coefficient, autocorrelation, normalized covariance***:

$$\rho_{XY} = \frac{\mathrm{Cov}\left[X,Y\right]}{\sigma_X \sigma_Y} = \mathbb{E}\left[\left(\frac{X-\mathbb{E}X}{\sigma_X}\right)\left(\frac{Y-\mathbb{E}Y}{\sigma_Y}\right)\right] = \frac{\mathbb{E}\left[XY\right]-\mathbb{E}X\mathbb{E}Y}{\sigma_X \sigma_Y}.$$

- $\rho_{X,X} = 1$
- $\rho_{X,Y} = 0$ if and only if X and Y are uncorrelated.

15.25. Linear Dependence and Caychy-Schwartz Inequality

(a) If $Y = aX + b$, then $\rho_{X,Y} = 1$.

(b) Cauchy-Schwartz Inequality:

$$\left(\mathrm{Cov}\left[X,Y\right]\right)^2 \leq \sigma_X^2 \sigma_Y^2$$

(c) This implies $|\rho_{X,Y}| \leq 1$. In other words, $\rho_{XY} \in [-1,1]$.

(d) When $\sigma_Y, \sigma_X > 0$, equality occurs if and only if the following conditions holds

$$\begin{aligned} &\equiv \exists a \neq 0 \text{ such that } (X-\mathbb{E}X) = a(Y-\mathbb{E}Y) \\ &\equiv \exists c \neq 0 \text{ and } b \in \mathbb{R} \text{ such that } Y = cX + b \\ &\equiv \exists a \neq 0 \text{ and } b \in \mathbb{R} \text{ such that } X = aY + b \\ &\equiv |\rho_{XY}| = 1 \end{aligned}$$

In which case, $|a| = \frac{\sigma_X}{\sigma_Y}$ and $\rho_{XY} = \frac{a}{|a|} = \operatorname{sgn} a$. Hence, ρ_{XY} is used to quantify **linear dependence** between X and Y. The closer $|\rho_{XY}|$ to 1, the higher degree of linear dependence between X and Y.

15.26. Variance and covaraince for linear combination of random variables:

(a) For finite index set I,

$$\mathrm{Var}\left[\sum_{i\in I} a_i X_i\right] = \sum_{i\in I} a_i^2 \,\mathrm{Var}\, X_i + 2 \sum_{\substack{(i,j)\in I\times I \\ i\neq j}} a_j a_j \mathrm{Cov}\left[X_i, X_j\right].$$

In particular

$$\mathrm{Var}\left[X+Y\right] = \mathrm{Var}\, X + \mathrm{Var}\, Y + 2\mathrm{Cov}\left[X,Y\right]$$

and

$$\mathrm{Var}\left[X-Y\right] = \mathrm{Var}\, X + \mathrm{Var}\, Y - 2\mathrm{Cov}\left[X,Y\right].$$

(b) For finite index set I and J,

$$\operatorname{Cov}\left[\sum_{i\in I} a_i X_i, \sum_{j\in J} b_j Y_j\right] = \sum_{i\in I}\sum_{j\in J} a_i b_j \operatorname{Cov}\left[X_i, Y_j\right].$$

16 Continuous Random Variables and pdf

At this level, we will not distinguish between the continuous random variable and absolutely continuous random variable.

Definition 16.1. A random variable X is said to be a ***continuous random variable*** if and only if any one of the following equivalent conditions holds.

$\equiv$ $\forall x$, $P\left[X = x\right] = 0$

$\equiv$ F_X is continuous

Definition 16.2. f is the (*probability*) ***density*** *function* f of a random variable X (or the distribution P^X)

$\equiv$ f is a nonnegative function on $\mathbb{R}$ such that

$$\forall B \quad P_X\left(B\right) = \int_B f(x)dx.$$

$\equiv$ X is ***absolutely continuous***

$\equiv$ X (or F^X) comes from the density f

$\equiv$ $\forall x \in \mathbb{R}\ F_X\left(x\right) = \int\limits_{-\infty}^{x} f\left(t\right)dt$

$\equiv$ $\forall a, b\ F_X\left(b\right) - F_X\left(a\right) = \int\limits_{a}^{b} f\left(x\right)dx$

16.3. If F does differentiate to f and f is continuous, it follows by the fundamental theorem of calculus that f is indeed a density for F. That is, if F has a continuous derivative, this derivative can serve as the density f.

16.4. Suppose a random variable X has a density f.

- F need not differentiate to f everywhere.
  - When $X \sim \mathcal{U}(a, b)$, F_X is not differentiable at a nor b.
- $\int\limits_{\Omega} f\left(x\right) dx = 1$
- f is determined only Lebesgue-a.e. That is, If $g = f$ Lebesgue-a.e., then g can also serve as a density for X and P_X

- f is nonnegative a.e. [6, stated on p. 138]
- X is a continuous random variable
- f at its continuity points must be the derivative of F
- $P\left[X \in [a,b]\right] = P\left[X \in [a,b)\right] = P\left[X \in (a,b]\right] = P\left[X \in (a,b)\right]$ because the corresponding integrals over an interval are not affected by whether or not the endpoints are included or excluded. In other words, $P[X = a] = P[X = b] = 0$.

16.5. $f_X\left(x\right) = \mathbb{E}\left[\delta\left(X - x\right)\right]$

Definition 16.6 (Absolutely Continuous CDF)**.** An absolutely continuous cdf F_{ac} can be written in the form

$$F_{ac}(x) = \int_{-\infty}^{x} f(z)dz,$$

where the integrand,

$$f(x) = \frac{d}{dx}F_{ac}(x),$$

is defined a.e., and is a nonnegative, integrable function (possibly having discontinuities) satisfying

$$\int f(x)dx = 1.$$

16.7. Any nonnegative function that integrates to one is a ***probability density function*** (pdf) [6, p. 139].

16.8. Remarks: Some useful intuitions

(a) Approximately, for a small Δx, $P\left[X \in [x, x+\Delta x]\right] = \int_x^{x+\Delta x} f_X(t)dt \approx f - X(x)\Delta x$. This is why we call f_X the density function.

(b) In fact, $f_X(x) = \lim\limits_{\Delta x \to 0} \frac{P[x < X \le x + \Delta x]}{\Delta x}$

16.9. Let T be an absolutely continuous nonnegative random variable with cumulative distribution function F and density f on the interval $[0, \infty)$. The following terms are often used when T denotes the lieftime of a device or system.

(a) Its survival-, survivor-, or reliability-function is:

$$R\left(t\right) = P\left[T > t\right] = \int\limits_t^\infty f\left(x\right)dx = 1 - F\left(t\right).$$

- $R(0) = P[T > 0] = P[T \ge 0] = 1$.

(b) The mean time of failure (MTTF) $= \mathbb{E}\left[T\right] = \int_0^\infty R(t)dt$.

(c) The (age-specific) failure rate or hazard function os a device or system with lifetime T is

$$r(t) = \lim_{\delta \to 0} \frac{P[T \le t + \delta | T > t]}{\delta} = -\frac{R'(t)}{R(t)} = \frac{f(t)}{R(t)} = \frac{d}{dt} \ln R(t).$$

(i) $r(t)\delta \approx P[T \in (t, t+\delta] | T > t]$

(ii) $R(t) = e^{-\int_0^t r(\tau) d\tau}$.

(iii) $f(t) = r(t) e^{-\int_0^t r(\tau) d\tau}$

- For $T \sim \mathcal{E}(\lambda)$, $r(t) = \lambda$.

See also [6, section 5.7].

Definition 16.10. A random variable whose cdf is continuous but whose derivative is the zero function is said to be ***singular***.

- See Cantor-type distribution in [3, p. 35–36].
- It has no density. (Otherwise, the cdf is the zero function.) So, $\exists$ continuous random variable X with no density. Hence, $\exists$ random variable X with no density.
- Even when we allow the use of delta function for the density as in the case of mixed r.v., it still has no density because there is no jump in the cdf.
- There exists singular r.v. whose cdf is strictly increasing.

Definition 16.11. $f_{X,A}(x) = \frac{d}{dx} F_{X,A}(x)$. See also definition **??**.

Definition 16.12. Sometimes, it is convenient to work with the "pdf" of a discrete r.v. Given that X is a discrete random variable which is defined as in definition **??**. Then, the "pdf" of X is

$$f_X(x) = \sum_{x_k} p_X(x_k) \delta(x - x_k), \quad x \in \mathbb{R}. \tag{15}$$

Although the delta function is not a well-defined function[7], this technique does allow easy manipulation of mixed distribution. The definition of quantities involving discrete random variables and the corresponding properties can then be derived from the pdf and hence there is no need to talk about pmf at all!

17 Function of Continuous Random Variables

18 Generation of Random Variable

18.1. Left-continuous inverse: $g^{-1}(y) = \inf \{x \in \mathbb{R} : g(x) \ge y\}$, $y \in (0,1)$

- *Trick*: Just flip the graph along the line $x = y$, then make the graph left-continuous.

[7] Rigorously, it is a unit measure at 0.

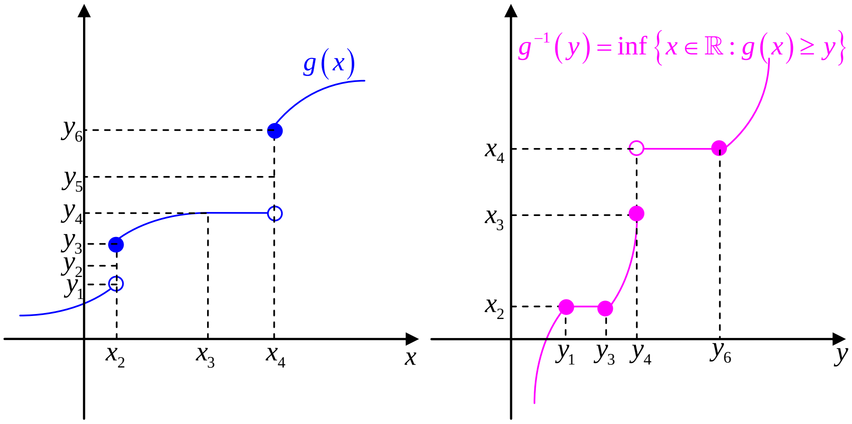

Figure 7: Left-continuous inverse on (0,1)

- If g is a cdf, then only consider $y \in (0,1)$. It is called the inverse CDF [5, Def 8.4.1, p. 238] or quantile function.
    - ○ In [12, Def 2.16, p. 25], the inverse CDF is defined using strict inequality "$>$" rather than "$\geq$".
- See Table 4 for examples.

Distribution	F	F^{-1}
Exponential	$1 - e^{-\lambda x}$	$-\frac{1}{\lambda}\ln(u)$
Extreme value	$1 - e^{-e^{\frac{x-a}{b}}}$	$a + b\ln\ln u$
Geometric	$1 - (1-p)^i$	$\left\lceil \frac{\ln u}{\ln(1-p)} \right\rceil$
Logistic	$1 - \frac{1}{1+e^{\frac{x-\mu}{b}}}$	$\mu - b\ln\left(\frac{1}{u} - 1\right)$
Pareto	$1 - x^{-a}$	$u^{-\frac{1}{a}}$
Weibull	$1 - e^{\left(\frac{x}{a}\right)^b}$	$a(\ln u)^{\frac{1}{b}}$

Table 4: Left-continuous inverse

18.2. Inverse-Transform Method: To generate a random variable X with CDF F, set $X = F^{-1}(U)$ where U is uniform on $(0,1)$. Here, F^{-1} is the left-continuous inverse of F.

Example 18.3. For example, to generate $X \sim \mathcal{E}(\lambda)$, set $X = -\frac{1}{\lambda}\ln(U)$

References

[1] F. N. David. *Games, Gods and Gambling: A History of Probability and Statistical Ideas*. Dover Publications, unabridged edition, February 1998. 4.15

[2] Rick Durrett. *Elementary Probability for Applications.* Cambridge University Press, 1 edition, July 2009. 4.14, 1d, 10.26, 10.30

[3] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. John Wiley & Sons, 1971. 16.10

[4] William Feller. *An Introduction to Probability Theory and Its Applications, Volume 1.* Wiley, 3 edition, 1968. 4.5

[5] Terrence L. Fine. *Probability and Probabilistic Reasoning for Electrical Engineering.* Prentice Hall, 2005. 5.1, 9.4, 10.26, 18.1

[6] John A. Gubner. *Probability and Random Processes for Electrical and Computer Engineers.* Cambridge University Press, 2006. 2.2, 4a, 15.1, 16.4, 16.7, 16.9

[7] Samuel Karlin and Howard E. Taylor. *A First Course in Stochastic Processes.* Academic Press, 1975. 7.1

[8] A.N. Kolmogorov. *The Foundations of Probability.* 1933. 5.1

[9] B. P. Lathi. *Modern Digital and Analog Communication Systems.* Oxford University Press, 1998. 1.1, 1.5

[10] Kenneth H. Rosen, editor. *Handbook of Discrete and Combinatorial Mathematics.* CRC, 1999. 1, 4.1

[11] Henk Tijms. *Understanding Probability: Chance Rules in Everyday Life.* Cambridge University Press, 2 edition, August 2007. 6.3, 15.1

[12] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer, 2004. 9.2, 15.1, 1, 18.1

[13] John M. Wozencraft and Irwin Mark Jacobs. *Principles of Communication Engineering.* Waveland Press, June 1990. 1.2