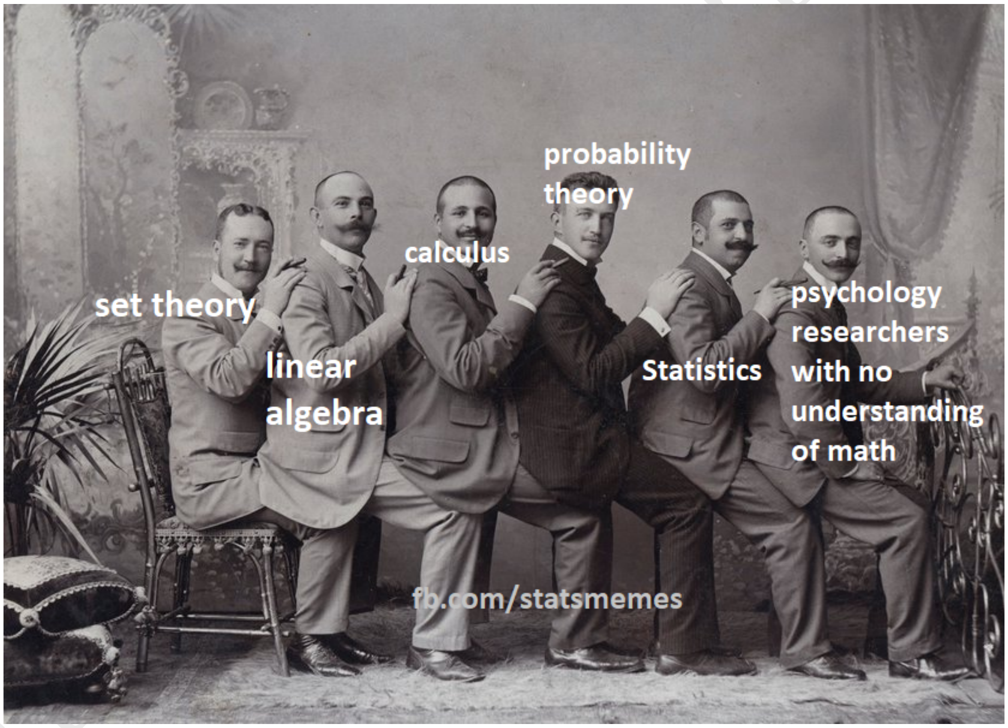
Fantuan's Academia

Fantuan's Math Notes series

Notes on Classical Probability Theory

Author: Jingxuan Xu



July 3, 2024

Contents

All the Sections with * are hard sections and can be skipped without losing coherence.

This note is referenced on **Introduction to Probability** by Joseph K. Blitzstein and Jessica Hwang [1].

Chapter 1

Basics in Classical Probability Theory

Definition 1.0.1: Sample Space

The **sample space** S of an experiment is the set of all possible outcomes of the experiment.

Definition 1.0.2: Event

An **event** A is a subset of the sample space S, i.e., a collection of outcomes of an experiment.

Definition 1.0.3: Disjoint/Partition

- Two events are **disjoint** (or **mutually exclusive**) if $A \cap B = \emptyset$.
- Events $A_1, A_2, \cdots, A_n$ are pairwise disjoint if $A_i \cap A_j = \emptyset$ for all $i \neq j$.
- If $A_1, A_2, \cdots, A_n$ are pairwise disjoint, and $\bigcup_{i=1}^n A_i = S$, they form a **partition** of S.

1.1 Naive Definition of Probability, Counting

Definition 1.1.1: Naive Definition of Probability

Let A be an event for an experiment with finite sample space S. The **naive probability** of A is

$$P_{\text{naive}}(A) = \frac{|A|}{|S|}$$

where $|\cdot|$ denotes the cardinality.

Two types of counting:

- **Sampling with replacement**: n objects and k choices from them. Then, there are n^k possible outcomes (where order matters).
- **Sampling without replacement**: n objects and k choices from them. Then, there are $n(n-1)(n-2)\cdots(n-k+1)$ possible outcomes for $1 \leqslant k \leqslant n$ (where order matters).

Example 1.1.2: Famous Example: Birthday Problem

There are k people in a room. Assume each person's birthday is equally likely to be any of the 365 days of a year (Exclude Feb 29), and people's birthdays are independent. What is the probability that at least one pair of people in the group have the same birthday?

Proof.
Sampling with replacement, there are 365^k ways to assign birthdays to people in the room.
Sampling without replacement, there are $365 \times 364 \times 363 \times \cdots (365 - k + 1)$ ways to assign birthdays to k people such that no two people share a birthday ($k \leqslant 365$).

Therefore, the probability of at least one birthday match is

$$P(\text{at least 1 birthday match}) = \begin{cases} 1 - \dfrac{365 \times 364 \times 363 \times \cdots (365 - k + 1)}{365^k}, & k \leqslant 365 \\ 1, & k > 365 \end{cases}$$

The surprising and famous point of this example is, for only $k = 23$ people, the probability of a match exceeds 0.5, where at $k = 57$ it already exceeds 0.99, which is extremely counter-intuitive.

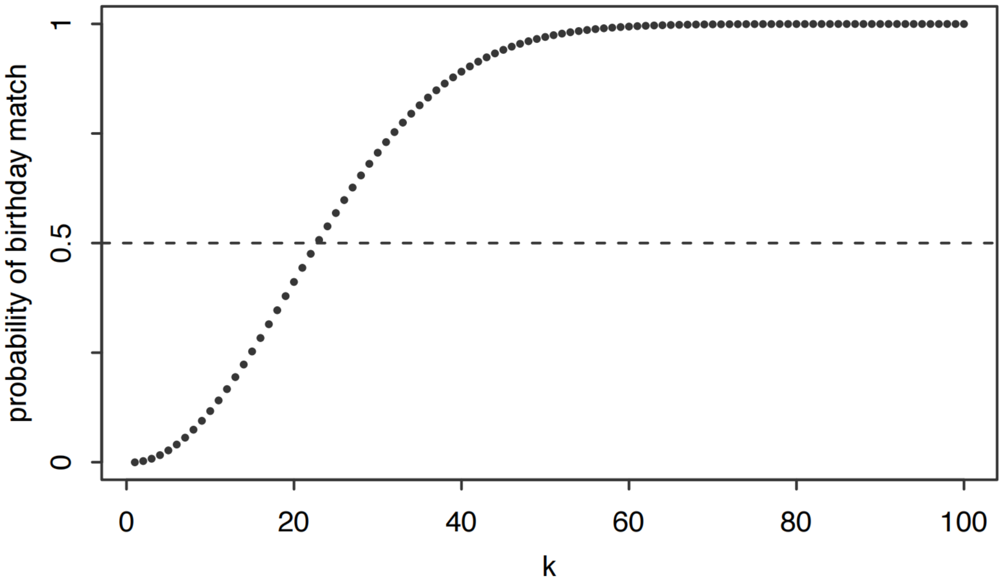


Figure 1.1: Probability of birthday match at different k

□

If we do not consider the order, there is another counting method.

Definition 1.1.3: Combination

For any nonnegative integers $k \leqslant n$, the **binomial coefficient**, or **combination**, $\binom{n}{k}$, is the number of subsets of size k for a set of size n. The value is

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

To choose k objects from n, there are $n(n-1)(n-2)\cdots(n-k+1)$ ways if order matters. This overcounts each subset of interest by a factor $k!$, since we don't care the order. So we can adjust for the overcounting by dividing k such that

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

This explains the formula.

Example 1.1.4: Permutation of a Word

1. How many ways to permute the letters in the word 'LALALAAA'?
2. How many ways to permute the letters in the word 'STATISTICS'?

Proof.

1. We just need to choose where the 5 A's go, so

$$\binom{8}{5} = \binom{8}{3} = \frac{8 \times 7 \times 6}{3!} = 56$$

2. Two ways to do this. We could choose where to put S's, and then T's, and then I's, and then A, with C determined. Thus,

$$\binom{10}{3}\binom{7}{3}\binom{4}{2}\binom{2}{1} = 50400$$

Or we start with 10! and adjust for overcounting of each letter. Thus,

$$\frac{10!}{3!3!2!} = 50400$$

□

Example 1.1.5: Binomial Theorem

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

Proof.
There are n factors

$$\underbrace{(x+y)(x+y)\cdots(x+y)}_{n \text{ factors}}$$

we can choose either x or y from each factor. There are $\binom{n}{k}$ ways to choose exactly k of x's, and each such choice yields the power $x^k y^{n-k}$. □

Example 1.1.6: Full House Poker

A 5-card hand is dealt from a standard, well-shuffled 52-card deck. The hand is called a *full house* if it consists 3 cards of some rank and 2 cards of another rank (e.g. three 7's and two 10's in any order). What is a probability of a full house?

Proof.
We first choose which rank should appear three times (13 possibilities), and choose three suits of it. Then, we choose which rank should appear two times (12 possibilities since we have chosen one for the three), and choose two suits of it. Thus,

$$P(\text{full house}) = \frac{13\binom{4}{3}12\binom{4}{2}}{\binom{52}{5}} \approx 0.00144$$

□

Example 1.1.7: Newton-Pepys Problem

Which of the following events has the highest probability?

- A: At least one 6 appears when 6 fair dice are rolled.
- B: At least two 6's appears when 12 fair dice are rolled.
- C: At least three 6's appears when 18 fair dice are rolled.

Proof.

- A: We count the ways to obtain no 6. This means to sample 1 to 5 with replacement six times. Thus,

$$P(A) = 1 - \frac{5^6}{6^6} \approx 0.67$$

- B: We count the ways to obtain no 6 or one 6. The previous is the same, and the second aims to choose one 6 from 12 dice, and sample 1 to 5 with replacement for the other 11 dice. Thus,

$$P(B) = 1 - \frac{5^{12} + \binom{12}{1}5^{11}}{6^{12}} \approx 0.62$$

- C: Similar as before, we have

$$P(C) = 1 - \frac{5^{18} + \binom{18}{1}5^{17} + \binom{18}{2}5^{16}}{6^{18}} \approx 0.60$$

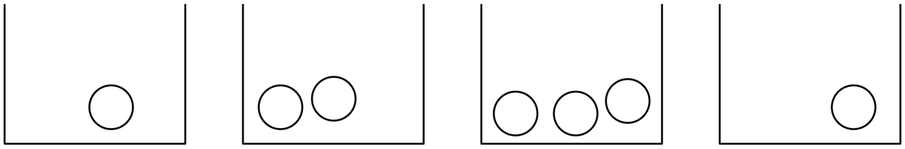
Hence A has the highest probability. □

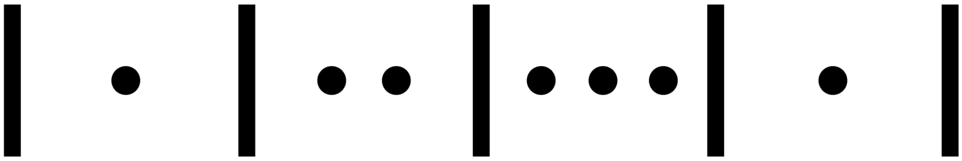
Example 1.1.8: Bose-Einstein Problem

How many ways are there to choose k times from a set of n objects with replacement, if order does not matter?

Proof.

We study its **isomorphic problem**. Let us find the number of ways to put k indistinguishable particles into n distinguishable boxes, i.e., swapping particles in any way is not considered as a separate possibility. This can be stated as a ordering of a sequence of walls and particles, as shown below.





A sequence must start and end with two walls. So, there leaves us $(n-1)$ walls and k particles in between to order. We only need to choose k positions in this $(n+k-1)$-long sequence. So the answer is

$$\binom{n+k-1}{k}$$

Why this problem is isomorphic to our original concern? We can let each box correspond to one of the n objects and use the particles as 'check marks' to tell how many times each object is selected. The particles are indistinguishable since we do not consider the order.

Another isomorphic problem is to count the number of solutions $(x_1, x_2, \cdots, x_n)$ to the equation

$$x_1 + x_2 + \cdots + x_n = k, \quad x_i \in \mathbb{N}$$

□

Note: This result cannot be used in the naive definition of probability, since the valid samples are not equally likely. For example, we choose 2 times from a set of 3 objects with replacement. The total number of ways is

$$\binom{2+3-1}{2} = 6$$

However, these 6 situations correspond to events:

- $\{1,2\}$ or $\{2,1\}$ is chosen.
- $\{1,3\}$ or $\{1,3\}$ is chosen.
- $\{2,3\}$ or $\{3,2\}$ is chosen.
- $\{1,1\}$ is chosen.

- $\{2,2\}$ is chosen.
- $\{3,3\}$ is chosen.

We see that the number of outcomes in each case is not equal.

1.2 Probability Space

Now we see the general definition of a probability.

Definition 1.2.1: Axiom of Probability

A **probability space** is a double (S, P) where S is a sample space and P is a **probability function** which takes an event $A \subseteq S$ to a real number between 0 and 1, with

- $P(\emptyset) = 0$ and $P(S) = 1$
- If $A_1, A_2, \cdots$ are disjoint events, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$$

Some desired properties of probability can be derived from these axioms.

Property 1.2.2: Properties of Probability

1. $P(A^c) = 1 - P(A)$.
2. If $A \subseteq B$, then $P(A) \leqslant P(B)$.

Proof.

1. Since A and A^c are disjoint, and $P(S) = 1$, we have

$$P(S) = P(A \cup A^c) = P(A^c) + P(A) = 1$$

2. Since A and $B \cap A^c$ is disjoint, and $A \subseteq B$, we have

$$P(A \cup B) = P(B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c) \geqslant P(A)$$

□

Continuing the proof 2, for arbitrary events A and B, we can write $A \cup B$ by

$$P(A \cup B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c)$$

Moreover, since $A \cap B$ and $A^c \cap B$ are disjoint, and their union is B, we have

$$P(A \cap B) + P(A^c \cap B) = P(B)$$

So combining the two equations, we have the formula,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.1}$$

Similarly, for three events, we can prove the following formula,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \tag{1.2}$$

Geometrically using Venn diagram, Equation 1.1 can be explained as adding the two area of A and B, and substract the overlapping region which is count twice. Similarly, for Equation 1.2 can be explained as adding the three area, and substract the three overlapping region which is count twice. Finally, the central region has been added three times and substracted three times, we need to add that region back again.

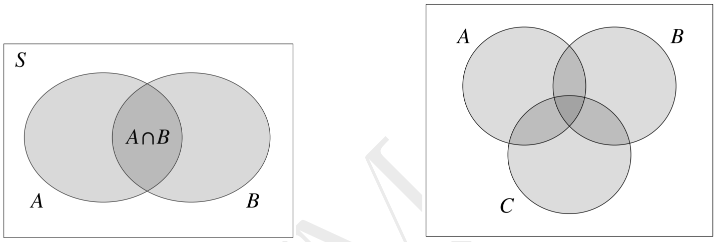


Figure 1.2: Venn Diagram of 2 and 3 events cases

Inductively, we can prove the **inclusion-exclusion criteria**.

Theorem 1.2.3: Inclusion-Exclusion Formula

For any events $A_1, A_2, \cdots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n)$$

Example 1.2.4: de Montmort's Matching Problem

Consider a well-shuffled deck of n cards, labeled through 1 to n. You flip over the cards one by one. You win the game if, at some point, the kth flipped card is the card labeled k. What is the probability of winning?

Proof. Let A_i be the event that ith card flipped has the number i on it. We are interested in the probability of the union $A_1 \cup A_2 \cup \cdots \cup A_n$, since as long as at least one of the cards has a number matching its position in the deck, you win the game. First note that

$$P(A_i) = \frac{1}{n}, \quad \forall\, i \in \{1, 2, \cdots, n\}$$

This can be seen in two ways:

- There are $n!$ possible orderings, and $(n-1)!$ of them are in A_i.
- **Symmetry:** the card number i is equally likely to be in any of the n positions.

Second,

$$P(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$$

since There are $n!$ possible orderings, and $(n-2)!$ of them are in $A_i \cap A_j$ (fixing two positions). Similarly, we can obtain

$$P(A_i \cap A_j \cap A_k) = \frac{1}{n(n-1)(n-2)}$$

and the pattern continues. In the inclusion-exclusion formula, there are n terms onvolving one event, $\binom{n}{2}$ terms involving two events, $\binom{n}{3}$ terms involving three events, and so forth. By **symmetry**, all n terms of the form $P(A_i)$ are equal, all $\binom{n}{2}$ terms of the form $P(A_i \cap A_j)$ are equal, and so forth. Therefore, using inclusion-exclusion formula:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= \frac{n}{n} - \binom{n}{2}\frac{1}{n(n-1)} + \binom{n}{3}\frac{1}{n(n-1)(n-2)} - \cdots + (-1)^{n+1}\frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1}\frac{1}{n!} \end{aligned}$$

Using Taylor Series, we see that as n goes to infinity,

$$\lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots = 1 - \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots\right) = 1 - e^{-1} \approx 0.63$$

Therefore, as the number of cards go to infinity, the probability of winning approaches 0.63. □

We have note that **symmetry** has a very important role in this kind of combinatorial situation. We will use symmetry many times later.

Chapter 2

Conditional Probability

2.1 Definition

Definition 2.1.1: Conditional Probability

If A and B are events with $P(B) > 0$, then the **conditional probability** of A given B, denoted by $P(A|B)$, is defined as

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

To explain why it is defined as this formula, consider a sample space with 9 outcomes as shown below. Events A and B are the two subsets of this space. We condition on B is happened. Therefore, we are restricting our vision from the whole space to the sub-sample space B. Within this subspace, if A happens, then the event is $A \cap B$. To **renormalize** the space so that $P(B) = 1$ (to make this B truly a sample space satisfying the axiom of probability), we divide by the propotion of B in the whole space S.

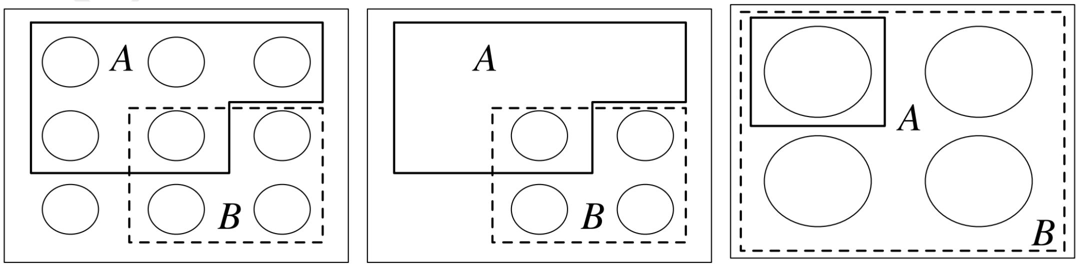


Figure 2.1: explanation of conditional probability

2.2 Independence

Definition 2.2.1: Independence

Events A and B are **independent** if

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

If $P(A) > 0$ and $P(B) > 0$, this is equivalent to

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B)$$

Note that this is a **symmetric relation:** If A is independent of B, then B is independent of A.

The definition just said that, condition on B, the probability of A does not change, which intuitively explain the terminology 'independence'. Also intuitively, if an event is independent from the other, then its complement is also independent from the other one.

Proposition 2.2.2

If A and B are independent, then A and B^c, A^c and B, A^c and B^c are all independent.

Proof. If $P(A) = 0$, then A is independent of every event. So assume $P(A) \neq 0$. Then,

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

which shows that A and B^c are independent. Swap the role of A and B, we will get A^c and B are independent. With A^c playing the role of A, we can get A^c and B^c are independent. □

For more than two sets, the matter is getting more difficult.

Definition 2.2.3: Independence for more than two sets

For n events $A_1, A_2, \cdots, A_n$,

- They are **pairwise independent** if $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j$.
- They are **jointly independent** if $P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n)$
- They are **independent** if $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j$, $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$ for all i, j, k distinct, and similarly for all quadruplets, quintuplets, and so on.
- For infinitely many events, they are independent if every finite subset of events is independent.

Note:

independence ⇒ **pairwise independence** and **joint independence**

pairwise independence ⇏ **joint independence**

joint independence $\not\Rightarrow$ pairwise independence

For the second and third comment, consider two fair, independent coin tosses. Let A be the event that the first is head. Let B be the event that the second is head. Let C be the event that both tosses have the same result. Then, A,B and C are pairwise independent, since

$$P(A\cap B)=\frac{1}{4}=P(A)P(B),\ P(A\cap C)=\frac{1}{4}=P(A)P(C),\ P(B\cap C)=\frac{1}{4}=P(B)P(C)$$

However, they are not jointly independent since

$$P(A\cap B\cap C)=\frac{1}{4}\neq P(A)P(B)P(C)=\frac{1}{8}$$

On the other hand, joint independence does not imply pairwise independence. This is easy to see when $P(A)=0$, which the joint dependence tells nothing about B and C.

We can also define conditional independence.

Definition 2.2.4: Conditional Independence

Events A and B are **conditionally independent** if

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

Note: Conditional Independence has no business to do with independence!

- **Two events can be conditionally independent given E, but not independent given E^c.** Suppose there are two types of class: good and bad. In good class, if you work hard, you are very likely to get an A. In bad class, the professor randomly grade students regardless their effort. Let G be the event that a class is good. Let W be the event that you work hard. Let A be the event that you receive an A. Then, W,A are conditionally independent given G^c, but conditionally dependent given G.

- **Two events can be conditionally independent given E, but not independent.** There is a fair coin and a biased coin with head probability $3/4$. We randomly select one, and toss a number of times. Conditional on choosing the fair coin, the coin tosses are independent. Similarly, conditinal on choosing the biased one, tosses are independent. However, the coin tosses are not unconditionally independent.

- **Two events can be independent, but not conditionally independent given** E. Let A,B be two independent random variables. Let $C=A+B$. Then, A and B are not conditionally independent given C, since as long as A is given, the value of B is known.

2.3 Bayes' Rule, Law of Total Probability (LOTP)

Rewriting the formula for conditional probability, we have

$$P(A\cap B)=P(B)P(A|B)=P(A)P(B|A) \tag{2.1}$$

This can be generalized to n events, by inductively using definition of conditional probability. Below we use comma to denote intersection, we have

$$P(A_1, A_2, \cdots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\cdots, P(A_n|A_1, A_2, \cdots, A_{n-1}) \tag{2.2}$$

Order does not matter. Indeed, we can have, for example

$$P(A_1, A_2, A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) = P(A_2)P(A_3|A_2)P(A_1|A_2, A_3)$$

From Equation 2.1, we get our famous **Bayes' Rule**, which relates two symmetric conditional probabilities.

Theorem 2.3.1: Bayes' Rule

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

The **Law of total probability**, instead, relates conditional probability to unconditional probability.

Theorem 2.3.2: Law of Total Probability

Let $A_1, A_2, \cdots, A_n$ be a partition of the sample space S with $P(A_i) > 0$ for all i. Then,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Proof. Since A_i is a partition of S, we can decompose B as

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

These terms of decomposition are disjoint, thus,

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n)$$

This can be seen as we slice the event B into n pieces, as the example graph 2.2 showed below where $n = 6$. Apply Equation 2.1 we will have

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n)$$

which is the formula showed above. □

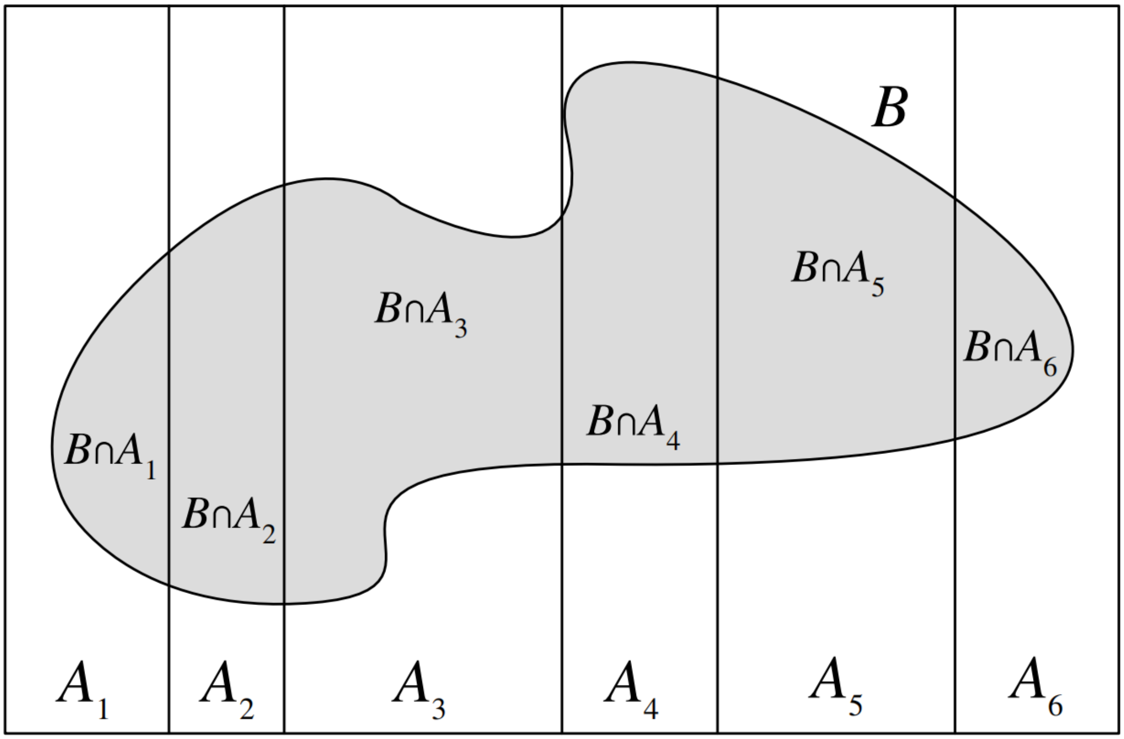


Figure 2.2: Law of Total Probability with $n = 6$

Example 2.3.3: Testing Rare Disease

A patient is tested for a rare disease with prevalence only 1% of the population. The test result is positive, i.e., the result claims that the patient has this disease. Suppose that the test is 95% accurate, meaning that there is a 95% probability that the test gives the correct claim. What is the probability that the patient truly has the disease?

Proof. Let D be the event that the patient has the disease. Let T be the event that the patient is tested positive. We know that $P(T|D) = 0.95$ and $P(T^c|D^c) = 0.95$, we want to find $P(D|T)$. Then, by **Bayes' Rule** and **Law of Total Probability**, we can derive

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} \approx 0.16 \end{aligned}$$

Therefore, there is only 16% chance that patient indeed has the disease, despite the test has a high accuracy. □

This weird and surprising results mainly results from the rareness of the disease, as shown below in Figure 2.3. There are two factors at play: the evidence from the test, and our prior information about the prevalence of the disease.

However, if the test happens twice, the situation would be significantly different.

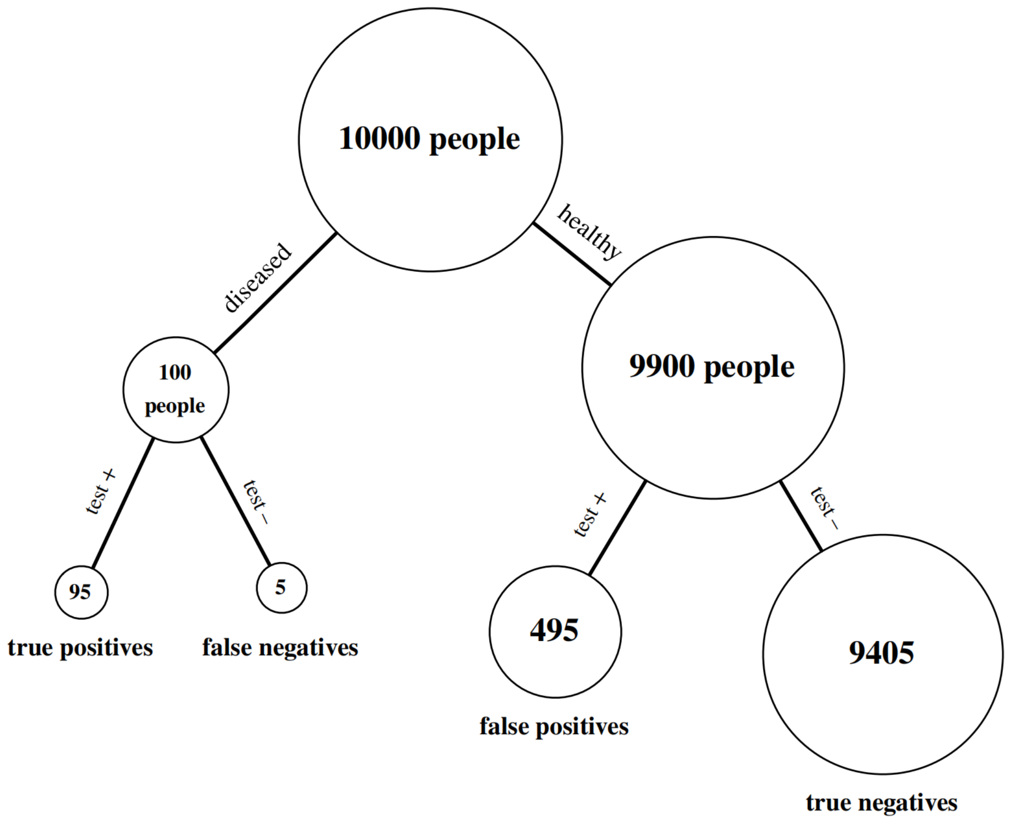


Figure 2.3: Test of a rare disease

Example 2.3.4: Testing Rare Disease: Continued

With the same setting as Example 2.3.3, what is the probability of the patient indeed has the disease if he is tested positive twice (Suppose the two tests are independent).

Proof. Let D denote that the patient has the disease, T_1 be that first test result is positive, T_2 be that the second test result is positive. The two events T_1 and T_2 are independent conditional on D. We have

$$\begin{aligned}P(D|T_1 \cap T_2) &= \frac{P(T_1 \cap T_2|D)P(D)}{P(T_1 \cap T_2)} = \frac{P(T_1 \cap T_2|D)P(D)}{P(T_1 \cap T_2|D)P(D) + P(T_1 \cap T_2|D^c)P(D^c)} \\ &= \frac{0.95^2 \times 0.01}{0.95^2 \times 0.01 + 0.05^2 \times 0.99} \approx 0.78\end{aligned}$$

which is significantly larger than 0.16. This shows that for a rare disease, a second test is necessary for diagnosing. □

To end this section, we notice that Bayes' Rule and Law of Total Probabilities can be stated with extra conditioning.

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

and

$$P(B|E) = \sum_{i=1}^{n} P(B|A_i,E)P(A_i|E)$$

2.4 Example: Monty Hall Problem

Example 2.4.1: Monty Hall Problem

A contestant chooses one of the three doors, two of which have a goat behind and one of which has a car. The host Monty, who knows where the car is, then open one of the remaining two that was not chosen behind which there is a goat. If the remaining two are both goats, he picks a door at random with equal probabilities. Monty then offers the contestant the option of switching to the other unopened door. If the contestant's goal is to get the car, should he/she switch doors?

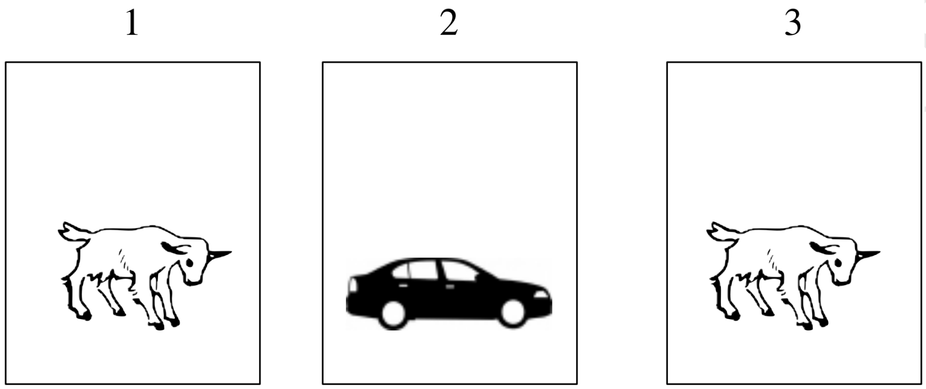


Figure 2.4: The Monty Hall Problem

Proof.
Let's label the door 1 through 3. Without loss of generality, we can assume the contestant picked door 1 (if he/she didn't, we can just relabel the doors). Then Monty reveals a door with goat. Let C_i be the event that the car is behind door i, for $i = 1, 2, 3$. By the law of total probability,

$$P(\text{get car}) = P(\text{get car}|C_1) \times \frac{1}{3} + P(\text{get car}|C_2) \times \frac{1}{3} + P(\text{get car}|C_3) \times \frac{1}{3}$$

Suppose the contestant use the switching strategy. If car is behind 1, he/she fails. So $P(\text{get car}|C_1) = 0$. If the car is behind 2 or 3, then because Monty always reveal a goat, the remaining one must contain a car, so $P(\text{get car}|C_2) = P(\text{get car}|C_3) = 1$. This leads to

$$P(\text{get car}) = 0 \times \frac{1}{3} + 1 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{2}{3}$$

The contest should switch to the other door. □

Many people, upon seeing this problem for the first time, argue that there is no advantage to switching: "There are two doors remaining, and one of them has the car, so the chances are 50-50."

To build correct intuition, let's consider an extreme case. Suppose that there are a million doors, 999,999 of which contain goats and 1 of which has a car. After the contestant's initial pick, Monty opens 999,998 doors with goats behind them and offers the choice to switch. In this extreme case, it becomes clear that the probabilities are not 50-50 for the two unopened doors; very few people would stubbornly stick with their original choice. The same is true for the three-door case.

2.5 Example: First-Step Analysis

In problems with recursive structure, *first-step analysis* is always a very powerful tool to use.

Example 2.5.1: Branching Process

A single amoeba lives in a pond. After one minute it will either die, split into two, or stay the same with equal probability. In subsequent minutes, all living amoebas will behave in the same way, independently. What is the probability that the amoeba population will eventually die out?

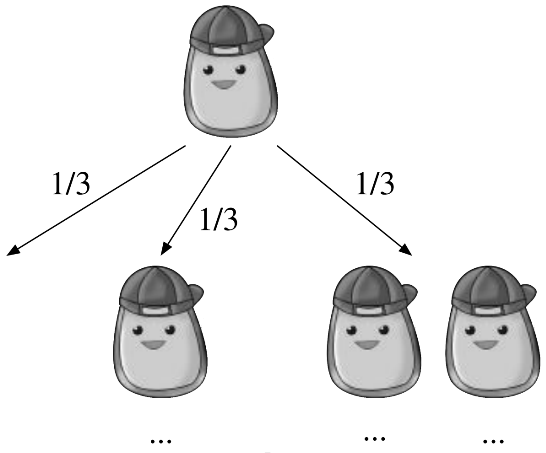


Figure 2.5: The Branching Process of amoeba

Proof.
Let D be the probability that the population will eventually die out. We proceed by conditioning on the outcome at the first step: Let B_i be the event that amoeba will turn into i amoebas after the first minute, for $i = 0, 1, 2$. We know that $P(D|B_0) = 1$ and $P(D|B_1) = P(D)$. If it splits in two, then we just have two independent versions of our original problem. We need both to die out, so $P(D|B_2) = P(D)^2$. Therefore,

$$P(D) = P(D|B_0) \times \frac{1}{3} + P(D|B_1) \times \frac{1}{3} + P(D|B_2) \times \frac{1}{3} = \frac{1}{3} + \frac{1}{3}P(D) + \frac{1}{3}P(D)^2$$

Solve the equation, we have $P(D) = 1$. □

Example 2.5.2: Gambler's Ruin

Two gamblers A and B make a sequence of \$1 bets. In each bet, A has probability p of winning, and B has probability $q = 1 - p$ of winning. A starts with i dollars and gambler B starts with $N - i$ dollars. The game ends when A or B is ruined. What is the probability that A wins all the money?

Proof.
We can visualize this as a *random walk* on the integers between 0 and N. It starts from i, at each step it has probability

p to move forward to N, and probability $1-p$ to move backwards to 0. The game ends when it reaches 0 or N. A wins if it reaches N.

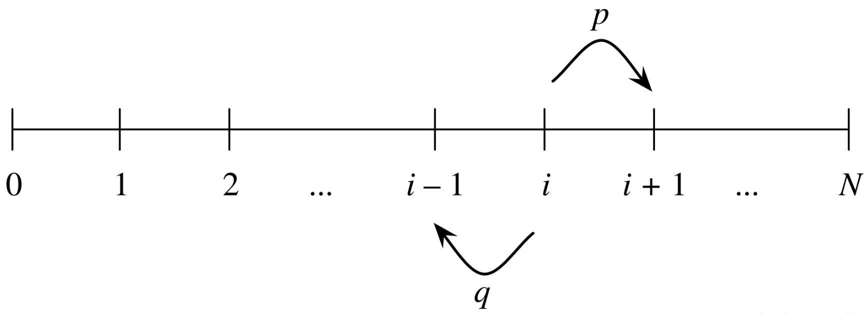


Figure 2.6: Random walk of Gambler's ruin

Let p_i be the probability that A wins the game, given that it starts with i dollars. Let W be the event that A wins the game. Conditioning on the first step, we have

$$\begin{aligned} p_i &= P(W|\text{A starts at } i, \text{ wins round 1}) \times p + P(W|\text{A starts at } i, \text{ loses round 1}) \times q \\ &= P(W|\text{A starts at } i+1) \times p + P(W|\text{A starts at } i-1) \times q \\ &= p_{i+1} \times p + p_{i-1} \times q \end{aligned}$$

The *characteristic function* of this *difference equation* is

$$px^2 - x + q = 0$$

which has root 1 and q/p. If $p \neq 1/2$, these roots are distinct, and the general solution is

$$p_i = a + b\left(\frac{q}{p}\right)^i$$

Using the boundary conditions $p_0 = 0$ and $p_N = 1$, we get

$$a = -b = \frac{1}{1 - \left(\frac{q}{p}\right)^N}$$

If $p = 1/2$, the roots are overlapped, the general solution is then

$$p_i = a + bi$$

The boundary conditions then gives $a = 0$ and $b = 1/N$.

In summary, the probability of A winning with a starting point of i is

$$p_i = \begin{cases} \dfrac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^N}, & \text{if } p \neq 1/2 \\ \dfrac{i}{N}, & \text{if } p = 1/2 \end{cases}$$

The $p = 1/2$ case is consistent with the $p \neq 1/2$ case, in the sense that

$$\lim_{p \to 1/2} \frac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^N} = \frac{i}{N}$$

By symmetry, we can also get the winning probability of B,

$$P(\text{B wins}|\text{B starts at } N-i) = \begin{cases} \dfrac{1-\left(\frac{q}{p}\right)^{N-i}}{1-\left(\frac{q}{p}\right)^N}, & \text{if } p \neq 1/2 \\ \dfrac{N-i}{N}, & \text{if } p = 1/2 \end{cases}$$

□

The surprising part of this problem is, having $p < 1/2$ will make A's chance of winning very low, even if p is only a little bit less than 1/2 and the players start out with the same amount of money. For example, if $p = 0.49$ and each player starts out with \$100, then A has only about a 1.8% chance of winning the game.

So, don't gamble!

Chapter 3

Random Variable

Definition 3.0.1: Random Variable

Given an experiment with sample space S, a **random variable (r.v.)** is a function $X : S \to \mathbb{R}$.

There are generally three types of random variables, i.e., *discrete, continuous, and mixed.* We will mainly consider discrete and continuous one here.

3.1 Cumulative Distribution Function (CDF)

Cumulative density function (CDF) is a function defined for all kinds of r.v.s.

Definition 3.1.1: Cumulative Distribution Function (CDF)

The **cumulative distribution function (CDF)** of a r.v. X is the function F_X given by

$$F_X(x) = P(X \leqslant x)$$

Valid CDF satisfy the following criteria.

Proposition 3.1.2: Valid CDF Properties

Any CDF F has the following properties.

- **Increasing:** If $x_1 \leqslant x_2$, then $F(x_1) \leqslant F(x_2)$
- **Right-continuous:** For any a, $F(a) = \lim_{x \to a^+} F(x)$.
- **Normalization:** $\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1$

Conversely, given any function F satisfying these criteria, we can construct a random variable whose CDF is F.

3.2 Discrete Random Variable and Probability Mass Function (PMF)

Definition 3.2.1: Discrete Random Variable

A random variable X is **discrete** if there is a countable list (finite or countably infinite) of values $a_1, a_2, \cdots$ such that $P(X = a_j \text{ for some } j) = 1$. The set of values x such that $P(X = x) > 0$ is called the **support** of X.

Probability mass function (PMF) is the special way to describe discrete r.v.s.

Definition 3.2.2: Probability Mass Function (PMF)

The **probability mass function (PMF)** of a discrete r.v. X is the function p_X given by $p_X(x) = P(X = x)$.

Valid PMF satisfies the following criteria.

Proposition 3.2.3: Valid PMF Properties

Let X be an discrete r.v. with support $\{x_j\}_{j=1}^{\infty}$. The PMF p_X of X has the following properties:

- **Nonnegative:** $p_X(x) > 0$ if $x = x_j$ for some j, and $p_X(x) = 0$ otherwise.
- **Normalization:** $\sum_{j=1}^{\infty} p_X(x_j) = 1$.

Conversely, if distinct values $\{x_j\}$ are specified and we have a function p satisfying all these criteria, we can construct an r.v. with PMF p.

3.3 Continuous Random Variable and Probability Density Function (PDF)

Definition 3.3.1: Continuous Random Variable

An r.v. is **continuous** if its CDF is continuous everywhere and differentiable except for possibly finitely many points.

Probability density function (PDF) is the special way to describe continuous r.v.s.

Definition 3.3.2: Probability Density Function

For a continuous r.v. X with CDF F, its **probability density function (PDF)** f is given by $f(x) = F'(x)$. The **support** of X is the set of x where $f(x) > 0$.

By Fundamental Theorem of Calculus, the following proposition is obvious.

Proposition 3.3.3: PDF to CDF

Let X be a continuous r.v. with PDF f. Then the CDF of X is given by

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

or for an arbitrary region $A \subseteq \mathbb{R}$,

$$P(X \in A) = \int_A f(x)\,\mathrm{d}x$$

Note: An important way that continuous r.v.s differ from discrete ones is that for continuous r.v. X, $P(X = x) = 0$ for all x. This means that the endpoints does not matter

$$P(a < X < b) = P(a \leqslant X < b) = P(a < X \leqslant b) = P(a \leqslant X \leqslant b)$$

Valid PDF satisfies the following criteria.

Proposition 3.3.4: Valid PDF Properties

The PDF f of a continuous r.v. has the following properties.

- **Nonnegative:** $f(x) \geqslant 0$.
- **Normalization:** $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

Conversely, any such function f is the PDF of some continuous r.v..

3.4 Expectation

Definition 3.4.1: Expectation

- The **expectation (mean)** of a discrete r.v. X with support $\{x_j\}$ is

$$E(X) = \sum_{j=1}^{\infty} x_j P(X = x_j)$$

- The expectation oof a continuous r.v. X with PDF f is

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

The most important property for expectation is *linearity*.

Theorem 3.4.2: Linearity of Expectation

For any r.v.s X, Y and any constant c,

1. $E(X + Y) = E(X) + E(Y)$
2. $E(cX) = cE(X)$
3. If X and Y are independent, $E(XY) = E(X)E(Y)$

Another important property is the *law of the unconscious statistician (LOTUS)*.

Theorem 3.4.3: Law of the Unconscious Statistician (LOTUS)

- If X is a discrete r.v. and $g : \mathbb{R} \to \mathbb{R}$ is a function. Then,

$$E(g(X)) = \sum_x g(x)P(X = x)$$

  where the sum is taken over all possible values of X.

- If X is a continuous r.v. with PDF f and $g : \mathbb{R} \to \mathbb{R}$ is a function. Then,

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x$$

There are some other properties for *conditional expectations*.

- **Law of total Expectation**: $A_1, A_2, \cdots, A_n$ be a partition of a sample space, with $P(A_i) > 0$ for all i, then

$$E(Y) = \sum_{i=1}^{n} E(Y|A_i)P(A_i)$$

- **Taking out what is known:** For any function h,

$$E(h(X)Y|X) = h(X)E(Y|X)$$

- **Tower Rule:**

$$E(E(Y|X)) = E(Y)$$

- **Projection:** The r.v. $Y - E(Y|X)$ is uncorrelated with $h(X)$ for any function h.

3.5 Variance

Definition 3.5.1: Variance

The variance of an r.v. X with expectation $\mu = E(X)$ is

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 p_X(x), & \text{if X is discrete with PMF} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x), & \text{if X is continuous with PDF} \end{cases}$$

Proposition 3.5.2: Equivalent formulation of variance

$$\mathrm{Var}(X) = E(X^2) - E(X)^2$$

Proof. Let $\mu = E(X)$. Then,

$$E[(X-\mu)^2] = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - E(X)^2$$

□

Several properties of variance are listed here.

Theorem 3.5.3: Properties of Variance

- $\mathrm{Var}(X+c) = \mathrm{Var}(X)$
- $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$
- If X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
- $\mathrm{Var}(X) \geqslant 0$, with equality if and only if $P(X=a)=1$ for some constant a.

The property for *conditional variance*

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

- **Law of Total Variance:**
$$\mathrm{Var}(Y) = E(\mathrm{Var}(Y|X)) + \mathrm{Var}(E(Y|X))$$

Note: This section is lack of proofs, since many of them are possible after introducing new concepts in later chapters. Proofs of some theorems are in the appendix.

Chapter 4

Examples of Discrete Random Variables

4.1 Discrete Uniform

Definition 4.1.1: Discrete Uniform Distribution

Let C be a finite, nonempty set of numbers. $X \sim \text{DUnif}(C)$ if

$$P(X = x) = \frac{1}{|C|}, \quad x \in C$$

4.2 Bernoulli and Binomial

Definition 4.2.1: Bernoulli Distribution

An experiment that can result in either a ‘success’ or a ‘failure’ is called a **Bernoulli trial**. a **Bernoulli distribution** with success probability p, denoted as $X \sim \text{Ber}(p)$ is

$$\mathbb{P}(X = k) = \begin{cases} p, & \text{if } k = 1 \\ 1 - p, & \text{if } k = 0 \end{cases}$$

Expectation: $E(X) = p$. Variance: $\text{Var}(X) = p(1 - p)$.

Having more than one trial leads to the *Binomial Distribution*.

Definition 4.2.2: Binomial Distribution

Suppose that n *independent* Bernoulli trials are performed, each with the same success probability p. Let X be the number of successes. The distribution of X is called the **binomial distribution** with parameter n and p. We write $X \sim \text{Bin}(n, p)$, the PMF is

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \cdots, n$$

Expectation: $E(X) = np$. Variance: $\mathrm{Var}(X) = np(1-p)$.

There are two ways to calculate expectation of binomial distribution:

- **Direct computation:** Using the combinatorics equation $k\binom{n}{k} = n\binom{n-1}{k-1}$, let $q = 1-p$, we have

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k q^{n-k} \\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}q^{n-k} = np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j q^{n-1-j} = np \end{aligned}$$

  The sum of the last formula equals 1 since it is the sum of PMF of $\mathrm{Bin}(n-1, p)$.

- **Sum of Independent Bernoulli Trials:** X is the sum of n independent $\mathrm{Ber}(p)$ r.v.s:

$$X = I_1 + I_2 + \cdots + I_n$$

  where each I_j has expectation $E(I_j) = p$. By linearity of expectation,

$$E(X) = E(I_1) + E(I_2) + \cdots + E(I_n) = np$$

Similarly, we can use the later thought to calculate the variance:
Each I_j has variance

$$\mathrm{Var}(I_j) = p(1-p)$$

Since I_j are independent, using the property of variance, we have

$$\mathrm{Var}(X) = \mathrm{Var}(I_1) + \mathrm{Var}(I_2) + \cdots + \mathrm{Var}(I_n) = np(1-p)$$

Below are some symmetric properties of binomial distribution.

Corollary 4.2.3: Symmetry of Binomial Distribution

Let $X = \mathrm{Bin}(n, p)$, and $q = 1-p$.

- $n - X \sim \mathrm{Bin}(n, q)$.
- If $p = 1/2$ and n is even, then the distribution is symmetric about $n/2$, in the sense that $P(X = n/2 + j) = P(X = n/2 - j)$ for all nonnegative integers j.

Proof. • Let $Y = n - X$. Then,

$$P(Y = k) = P(X = n-k) = \binom{n}{n-k}p^{n-k}q^k = \binom{n}{k}q^k p^{n-k}$$

- By the last property, $n - X \sim \text{Bin}(n, 1/2)$, so

$$P(X = k) = P(n - X = k) = P(X = n - k)$$

for all nonnegative integer k. Let $k = n/2 + j$, the desired result follows.

□

An important property of binomial distribution is that, the sum of two binomial r.v.s with same success probability is still a binomial r.v..

Theorem 4.2.4: Sum of Binomial Distributions

If $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, and X is independent of Y, then $X + Y \sim \text{Bin}(n + m, p)$.

Proof.
By law of total probability, we have

$$P(X+Y=k) = \sum_{j=0}^{k} P(X+Y=k|X=j)P(X=j) = \sum_{j=0}^{k} P(Y=k-j|X=j)P(X=j)$$

Since X and Y are independent, we have

$$\begin{aligned} P(X+Y=k) &= \sum_{j=0}^{k} P(Y=k-j)P(X=j) = \sum_{j=0}^{k} \binom{m}{k-j} p^{k-j} q^{m-k+j} \binom{n}{j} p^j q^{n-j} \\ &= p^k q^{n+m-k} \sum_{j=0}^{k} \binom{m}{k-j}\binom{n}{j} = \binom{n+m}{k} p^k q^{n+m-k} \end{aligned}$$

where the fact $\sum_{j=0}^{k} \binom{m}{k-j}\binom{n}{j} = \binom{n+m}{k}$ (**Vandermonde's Identity**) is used in the last step. □

Explanation: Let $X = X_1 + X_2 + \cdots + X_n$ and $Y = Y_1 + Y_2 + \cdots Y_m$ be the sum of i.i.d. Bernoulli r.v.s. Then, $X + Y$ is the sum of $n + m$ i.i.d. Bernoulli r.v.s., so its distribution is naturally binomial with parameters $n + m$ and p.

4.2.1 Example: Random Walk

Example 4.2.5: Random Walk

A particle moves n steps on a number line. The particle starts at 0, and at each step it moves 1 unit to the right or to the left, with equal probabilities. Assume all steps are independent. Let Y be the particle's position after n steps.

1. Find the PMF of Y.
2. Let D be the particle's distance from the origin after n steps. Assume that n is even. Find the PMF of D.

Proof.

1. Consider each step to be a Bernoulli trial, where right is considered a success and left is considered as a failure. Then the number of steps to the right is $\text{Bin}(n, 1/2)$ random variable, denote this by X. If $X = j$, then the particle has taken j steps to the right and $n - j$ to the left, given a final position $j - (n - j) = 2j - n$. Therefore, $Y = 2X - n$. Since X takes values in $\{0, 1, 2, \cdots, n\}$, Y takes values in $\{-n, 2 - n, 4 - n, \cdots, n\}$. The PMF of Y is
$$P(Y = k) = P(2X - n = k) = P(X = (n + k)/2) = \binom{n}{\frac{n+k}{2}}\left(\frac{1}{2}\right)^n$$
if k is an integer between $-n$ and n such that $n + k$ is an even number.

2. $D = |Y|$. The event $D = 0$ is same as $Y = 0$. For $k = 2, 4, \cdots, n$, the event $D = k$ is the same as the event $\{Y = k\} \cup \{Y = -k\}$. So the PMF of D is
$$P(D = 0) = \binom{n}{\frac{n}{2}}\left(\frac{1}{2}\right)^n$$
and
$$P(D = k) = P(Y = k) + P(Y = -k) = 2\binom{n}{\frac{n+k}{2}}\left(\frac{1}{2}\right)^n$$
for $k = 2, 4, \cdots, n$. Note that by symmetry, $P(Y = k) = P(Y = -k)$. □

4.3 Hypergeometric

If we have an urn filled with w white and b black balls, then drawing n balls out of the urn *with replacement* yields a $\text{Bin}(n, w/(w + b))$ distribution for the number of white balls obtained from the n trials. If we instead sample *without replacement*, then the number of white balls follows a *hypergeometric distribution*.

Definition 4.3.1: Hypergeometric Distribution

The hypergeometric distribution $X \sim \text{HGeom}(w, b, n)$ has PMF
$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}, \quad 0 \leqslant k \leqslant w,\ 0 \leqslant n - k \leqslant b$$
Expectation: $E(X) = nw/(w + b)$. Variance: $\text{Var}(X) = \frac{N-n}{N-1}npq$, where $N = w + b$, $p = w/N$, $q = 1 - p$.

The PMF is derived such that, the numerator counts the number of ways to collect k white balls and $n - k$ black balls, and the denominator counts the number of ways to collect n balls from $w + b$ balls, whatever black or white. Note that the Bernoulli trial here in hypergeometric is *dependent* since it is sampled without replacement.

Expectation:

We can write $X \sim \text{HGeom}(w, b, n)$ as a sum of Bernoulli random variables,
$$X = I_1 + I_2 + \cdots + I_n$$
where I_j equals 1 if the jth ball in the sample is white and 0 otherwise. By symmetry, $I_j \sim \text{Ber}(p)$ with $p = w/(w + b)$, since unconditionally the jth ball drawn is equally likely to be any of these balls. Though I_j are dependent, we can

still use linearity of expectation to see that

$$E(X) = \frac{nw}{w+b}$$

Variance:
Instead of finding $\text{Var}(X)$, we first find $E\binom{X}{2}$. $\binom{X}{2}$ is the number of pairs of draws such that both balls are white. Define $I_{ij}, i<j$ the *indicator variable* such that if the ith draw and jth draw are both white, $I_{ij}=1$, otherwise $I_{ij}=0$. By symmetry, $I_{ij} \sim \text{Ber}(p)$ with $p = \frac{w}{w+b}\frac{w-1}{w+b-1}$, since unconditionally the ith draw is equally likely to be any of these balls, and jth draw is equally likely to be any of the balls except for the drawn one in the ith draw. Therefore,

$$E\binom{X}{2} = \sum_{i<j} E(I_{ij}) = \binom{n}{2}\frac{w}{w+b}\frac{w-1}{w+b-1}$$

Therefore,

$$\frac{n!}{2!(n-2)!}\frac{w}{w+b}\frac{w-1}{w+b-1} = \frac{n(n-1)}{2}\frac{w}{w+b}\frac{w-1}{w+b-1} = E\left(\frac{X(X-1)}{2}\right) = \frac{1}{2}\left(E(X^2)-E(X)\right)$$

This leads to

$$E(X^2) = E(X) + n(n-1)p\frac{w-1}{N-1} = np + n(n-1)p\frac{w-1}{N-1}$$

Finally, the variance can be derived as

$$\begin{aligned}
\text{Var}(X) = E(X^2) - E(X)^2 &= np + n(n-1)p\frac{w-1}{N-1} - n^2p^2 \\
&= np\left((n-1)\frac{w-1}{N-1} + 1 - np\right) = np\left(\frac{(n-1)(w-1)+N-1}{N-1} - \frac{nw}{N}\right) \\
&= np\left(\frac{Nnw - nN - Nw + N^2 - Nnw + nw}{N(N-1)}\right) = np\left(\frac{(N-n)(N-w)}{N(N-1)}\right) \\
&= np\frac{N-n}{N-1}\frac{b}{N} = \frac{N-n}{N-1}npq
\end{aligned}$$

The next proposition shows the symmetry of hypergeometric distribution.

Proposition 4.3.2: Symmetry of Hypergeometric Distribution

The $\text{HGeom}(w,b,n)$ and $\text{HGom}(n, w+b-n, w)$ distributions are identical.

Proof. Directly checking this using computation is boring. So we make a story. Imagine an urn with w white balls, b black balls and a sample of size n without replacement. Think $\text{HGom}(n, w+b-n, w)$ as the number of sampled balls among the white balls. Both are counting the number of white sampled balls, so they have the same distribution. □

4.3.1 Connection Between Binomial and Hypergeometric

We can get from the Binomial to the Hypergeometric by *conditioning*, and we can get from the Hypergeometric to the binomial by *taking a limit*.

Theorem 4.3.3: Binomial to Hypergeometric

If $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$, and X is independent of Y. Then,

$$X|X + Y = r \sim \text{HGeom}(n, m, r)$$

Proof. By Bayes' rule, we have

$$P(X = x|X + Y = r) = \frac{P(X + Y = r|X = x)P(X = x)}{P(X + Y = r)} = \frac{P(Y = r - X)P(X = x)}{P(X + Y = r)}$$

where the last equation is justified by the independence of X and Y. where

$$P(X + Y = r|X = x) = P(Y = r - x|X = x) = P(Y = r - x)$$

Therefore,

$$P(X = x|X + Y = r) = \frac{\binom{m}{r-x}p^{r-x}(1-p)^{m-r+x}\binom{n}{x}p^x(1-p)^{n-x}}{\binom{n+m}{r}p^r(1-p)^{n+m-r}} = \frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}}$$

which is the PMF of Hypergeometric distribution. □

Explanation: Suppose X is the number of sampled white ball and Y is the number of sampled black ball. Conditioning on $X+Y = r$, we fix that we sample r balls from the boxes, whatever white or black. This is analog to the hypergeometric story.

Theorem 4.3.4: Hypergeometric to Binomial

If $X \sim \text{HGeom}(w, b, n)$ and $N = w+b \to \infty$ such that $p = w/(w+b)$ remain fixed, then the PMF of X converges to the PMF of distribution $\text{Bin}(n, p)$.

Proof. Rewrite the PMF of hypergeometric distribution by Proposition 4.3.2,

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} = \binom{n}{k}\frac{\binom{w+b-n}{w-k}}{\binom{w+b}{w}}$$

Expanding those binomial coefficients, we have

$$\begin{aligned}
P(X = k) &= \binom{n}{k}\frac{w!}{(w-k)!}\frac{b!}{(b-n+k)!}\frac{(w+b-n)!}{(w+b)!} \\
&= \binom{n}{k}\frac{w(w-1)\cdots(w-k+1)b(b-1)\cdots(b-n+k+1)}{(w+b)(w+b-1)\cdots(w+b-n+1)} \\
&= \binom{n}{k}\frac{p(p-\frac{1}{N})\cdots(p-\frac{k-1}{N})q(q-\frac{1}{N})\cdots(q-\frac{n-k-1}{N})}{(1-\frac{1}{N})(1-\frac{2}{N})\cdots(1-\frac{n-1}{N})}
\end{aligned}$$

as $N \to \infty$, we see that the denominator goes to 1, and the numerator goes to $p^k q^{n-k}$. Thus,

$$P(X = k) \to \binom{n}{k} p^k q^{n-k}$$

which is the $\text{Bin}(n, p)$ PMF. □

Explanation: As the number of balls in the urn grows to very large relative to the number of balls drawn, i.e., N is extremely large with respect to n, sampling with replacement and sampling without replacement become essentially equivalent.

4.4 Geometric

Definition 4.4.1: Geometric Distribution

Consider a sequence of independent Bernoulli trials, each with the same success probability $p \in (0, 1)$, with trial performed untill a success occurs. Let X be the *number of failures* before the first success. Then, X has the **Geometric distribution** with parameter p. Denoted $X \sim \text{Geom}(p)$, the PMF is

$$P(X = k) = (1 - p)^k p, \quad k = 0, 1, 2, \cdots$$

Expectation: $E(X) = (1 - p)/p$. Variance: $\text{Var}(X) = (1 - p)/p^2$.

Note: Some people use the *total number of trials* to define geometric distribution.
Expectation: There are two ways to compute the expectation

- *Direct Computation:* By definition,

$$E(X) = \sum_{k=0}^{\infty} k(1 - p)^k p$$

It reminds the derivative of geometric series. Therefore, consider

$$\sum_{k=0}^{\infty} (1 - p)^k = \frac{1}{1 - (1 - p)} = \frac{1}{p}$$

This series converges since $0 < 1 - p < 1$. Differentiate both sides,

$$-\sum_{k=0}^{\infty} k(1 - p)^{k-1} = -\frac{1}{p^2}$$

Therefore, we can compute our expectation as

$$E(X) = \sum_{k=0}^{\infty} k(1 - p)^k p = p(1 - p) \sum_{k=0}^{\infty} k(1 - p)^{k-1} = \frac{1 - p}{p}$$

- *First-step Analysis:* We condition on the outcome of first toss: if it is success, we are done; if it is failure, we waste one chance and back to where we started. Therefore,

$$E(X) = E(X|\text{first is success}) \times p + E(X|\text{first is failure}) \times (1-p) = 0 \times p + (1 + E(X)) \times (1-p)$$

Solve the equation and we will get the same answer.

Variance:
We already know that $E(X) = (1-p)/p$. By LOTUS,

$$E(X^2) = \sum_{k=0}^{\infty} k^2 (1-p)^k p$$

Again, consider the geometric series mentioned abouve, we have calculated that the derivative

$$\sum_{k=0}^{\infty} k(1-p)^{k-1} = \frac{1}{p^2}$$

If we multiply $(1-p)$ to both sides,

$$\sum_{k=0}^{\infty} k(1-p)^{k} = \frac{1-p}{p^2}$$

Differentiate again, we have

$$-\sum_{k=0}^{\infty} k^2(1-p)^{k-1} = \frac{-p^2 - (1-p)2p}{p^4} = \frac{p-2}{p^3}$$

Therefore, multiply $-p(1-p)$ on both sides, we have

$$E(X^2) = \sum_{k=0}^{\infty} k^2(1-p)^k p = \frac{(2-p)(1-p)}{p^2} = \frac{p^2 - 3p + 2}{p^2}$$

The variance is then

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{p^2 - 3p + 2}{p^2} - \frac{p^2 - 2p + 1}{p^2} = \frac{1-p}{p^2}$$

The next example will illustrate the usage of the different definition of geometric series using total number of trials.

Example 4.4.2: Coupon Collector

Suppose there are n types of toys, which you are collecting one by one, with the goal of getting a complete set. When collecting toys, the toy types are random (as is sometimes the case, for example, with toys included in cereal boxes or included with kids' meals from a fast food restaurant). Assume that each time you collect a toy, it is equally likely to be any of the n types. What is the expected number of toys needed until you have a complete set?

Proof. Let N be the number of toys needed; we want to find $E(N)$. Our strategy will be to break up N into a sum of simpler r.v.s so that we can apply linearity. So write

$$N = N_1 + N_2 + \cdots + N_n$$

where N_1 is the number of toys until the first toy type you haven't seen before (which is always 1, as the first toy is always a new type), N_2 is the additional number of toys until the second toy type you haven't seen before, and so forth.

Then, $N_1 = 1$, $N_2 \sim \text{Geom}((n-1)/n) + 1$: after collecting the first toy type, there's a $1/n$ chance of getting the same toy you already had (failure) and an $(n-1)/n$ chance you'll get something new (success). The $+1$ in the formula is because now we consider the *total number of trials* instead of total number of failures. In general,

$$N_j \sim \text{Geom}\left(\frac{n-j+1}{n}\right) + 1$$

Therefore, by linearity,

$$E(N) + E(N_1) + E(N_2) + \cdots + E(N_n) = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + n = n\sum_{j=1}^{n}\frac{1}{j}$$

Since for each N_i, its expectation is $\frac{j-1}{n-j+1} + 1 = \frac{n}{n-j+1}$. □

4.4.1 Memoryless Property

Geometric distribution has a famous *memoryless property.*

Theorem 4.4.3: Memoryless Property

For Geometric random variable X, we have

$$\mathbb{P}(X > j+k|X > j) = \mathbb{P}(X > k)$$

Moreover, geometric r.v. is the only discrete r.v. that has this property.

Proof.

- Using Bayes' Rule,

$$P(X > j+k|X > j) = \frac{P(X > j+k \cap X > j)}{P(X > j)} = \frac{P(X > j+k)}{P(X > j)}$$

Using the tail probability of geometric random variable,

$$P(X \geqslant k+1) = P(X > k) = \sum_{n=k+1}^{\infty}(1-p)^{n-1}p = p\sum_{n=k}^{\infty}(1-p)^n = p\frac{(1-p)^k}{1-(1-p)} = (1-p)^k$$

We can substitute this into the previous equation:

$$P(X > j+k|X > j) = \frac{P(X > j+k)}{P(X > j)} = \frac{(1-p)^{j+k}}{(1-p)^j} = (1-p)^k = P(X > k)$$

- Conversely, if we have memoryless property for all j, we must have it for $j = 1$. If we denote $p = 1 - P(X >$

1), we have

$$P(X > k+1) = P(X > k+1|X > 1)P(X > 1) = P(X > k)P(X > 1) = (1-p)P(X > k)$$

where the second last equality comes from our assumption of memoryless property. We can do this inductively such that

$$P(X > k+1) = (1-p)P(X > k) = (1-p)^2 P(X > k-1) = \cdots = (1-p)^{k+1}$$

which gives

$$P(X = k) = P(X > k-1) - P(X > k) = p(1-p)^{k-1}$$

which is the PMF of geometric distribution.

□

4.5 Negative Binomial

Negative binomial distribution generalizes geometric distribution. Instead of waiting for just one success, we can wait for any predetermined number r of successes.

Definition 4.5.1: Negative Binomial Distribution

In a sequence of independent Bernoulli trials with success probability p, if X is the *number of failures before the* r*th success*, then X is said to have the **Negative Binomial distribution** with parameters r and p, denoted by $X \sim \text{NegBin}(r, p)$. The PMF is

$$P(X = n) = \binom{n+r-1}{r-1} p^r (1-p)^n, \quad n = 0, 1, 2, \cdots$$

Expectation: $E(X) = r(1-p)/p$. Variance: $\text{Var}(X) = r(1-p)/p^2$.

Explanation of PMF: Imagine a string of 0's and 1's, with 1's representing successes. The probability of any specific string of n 0's and r 1's is $p^r(1-p)^n$. How many such strings are there? Because we stop as soon as we hit the rth success, the string must terminate in a 1. Among the other $n+r-1$ positions, we choose $r-1$ places for the remaining 1's to go.

4.5.1 Negative Binomial as Sum of Geometric R.v.s

Just as a Binomial r.v. can be represented as a sum of i.i.d. Bernoullis, a Negative Binomial r.v. can be represented as a sum of i.i.d. geometrics.

Theorem 4.5.2: Negative Binomial as Sum of Geometric R.v.s

Let $X \sim \text{NegBin}(r, p)$. Then we can write

$$X = X_1 + X_2 + \cdots + X_r$$

where X_i's are Geom(p) r.v.s and are independent.

Proof. Let X_1 be the number of failures until the first success, X_2 be the number of failures between the first success and the second success, and in general, X_i be the number of failures between the $(i-1)$st success and the ith success.

Then $X_1 \sim \text{Geom}(p)$ by the story of the Geometric distribution. After the first success, the number of additional failures until the next success is still Geometric! So $X_2 \sim \text{Geom}(p)$, and similarly for all the X_i. Furthermore, the X_i are independent because the trials are all independent of each other. Adding the X_i, we get the total number of failures before the rth success, which is X. □

For the proof in rigorous calculation, please see Appendix.

Expectation:

By linearity of expectation, we can easily get

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_r) = \frac{r(1-p)}{p}$$

Variance: Since $X_i's$ are independent, we can easily get

$$\text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_r) = \frac{r(1-p)}{p^2}$$

4.6 Indicator Random Variable

A very strong tool exists in probability theory, which is *indicator variable.*

Definition 4.6.1: Indicator Random Variable

The **indicator random variable** of an event A is the r.v. which equals 1 if A occurs and 0 otherwise. Denote as I_A. Note that $I_A \sim \text{Ber}(p)$ with $p = P(A)$.

This leads to its most important property:

Theorem 4.6.2: Fundamental Bridge between probability and Expectation

$$P(A) = E(I_A)$$

This can be easily seen from the definition of Bernoulli distribution. We use few examples to show its power.

Example 4.6.3: de Montmort's Matching Problem: Continued

Consider a well-shuffled deck of n cards, labeled through 1 to n. You flip over the cards one by one. You win the game if, at some point, the kth flipped card is the card labeled k. Let X be the number of matches. Find $E(X)$.

Proof.

In Chapter 1 we have shown that the probability of winning, as n goes to infinity, is about $1 - e^{-1}$. Now let's check $E(X)$. We can write X as sum of indicator r.v.s $X = I_1 + I_2 + \cdots + I_n$ such that

$$I_j = \begin{cases} 1, & \text{if the } j\text{th card in the deck is a match} \\ 0, & \text{otherwise} \end{cases}$$

Denote A_j as the event that jth card is a match. We know that for each I_j, by symmetry

$$E(I_j) = P(A_j) = \frac{1}{n}$$

So by linearity of expectation, we have

$$E(X) = E(I_1) + E(I_2) + \cdots + E(I_n) = n \times \frac{1}{n} = 1$$

□

Example 4.6.4: Distinct Birthday

In a group of n people, under the usual assumptions about birthdays, what is the expected number of distinct birthdays among the n people, i.e., the expected number of days on which at least one of the people was born? What is the expected number of birthday matches, i.e., pairs of people with the same birthday?

Proof.

- Let X be the number of distinct birthdays, and write $X = X_1 + \cdots + X_{365}$, where

  $$I_j = \begin{cases} 1, & \text{if the } j\text{th day is represented} \\ 0, & \text{otherwise} \end{cases}$$

  By the fundamental bridge,

  $$E(I_j) = P(j\text{th day is represented}) = 1 - P(\text{no one born on day } j) = 1 - \left(\frac{364}{365}\right)^n$$

  By linearity,

  $$E(X) = 365\left[1 - \left(\frac{364}{365}\right)^n\right]$$

- Let Y be the number of birthday matches. Label the people as $1, 2, \cdots, n$, and order the $\binom{n}{2}$ pairs of people in some definite way. Then, we can write

$$Y = J_1 + J_2 + \cdots + J_{\binom{n}{2}}$$

where J_i is the indicator of ith pair of people having the same birthday. The probability of any two people having the same birth day is (there are totally 365 days to choose, two birthday are independent, so each person has 1/365 probability to be born on that day):

$$E(J_i) = 365\frac{1}{365^2} = \frac{1}{365}$$

Therefore, by linearity,

$$E(Y) = \frac{1}{365}\binom{n}{2}$$

□

Example 4.6.5: Putnam Problem

A permutation $a_1, a_2, \cdots, a_n$ of $1, 2, \cdots, n$ has a local maximum at j for $2 \leqslant j \leqslant n-1$ if $a_j > a_{j-1}$ and $a_j > a_{j+1}$. For $j = 1$, a local maximum means $a_1 > a_2$ while for $j = n$, it means $a_n > a_{n-1}$. For $n \geqslant 2$, what is the average number of local maxima of a random permutation of $1, 2, \cdots, n$, with all $n!$ permutations equally likely?

Proof. Let $I_1, I_2, \cdots, I_n$ be indicator r.v.s. where I_j is 1 if there is a local maximum at position j, and 0 otherwise. Now, for $2 \leqslant j \leqslant n-1$,

$$E(I_j) = \frac{1}{3}$$

since having a local maximum at j is equivalent to a_j being the maximum of a_{j-1}, a_j and a_{j+1}, which has probability 1/3 by symmetry (orders are equally likely). For $j = 1$ and n,

$$E(I_j) = \frac{1}{2}$$

for the same reason. Thus, by linearity,

$$E\left(\sum_{j=1}^{n} I_j\right) = 2 \times \frac{1}{2} + (n-2) \times \frac{1}{3} = \frac{n+1}{3}$$

□

4.7 Negative Hypergeometric

Definition 4.7.1: Negative Hypergeometric Distribution

An urn contains w white balls and b black balls, which are randomly drawn one by one *without replacement*, until r white balls have been obtained. The number of black balls drawn before drawing the rth white ball has a **Negative Hypergeometric distribution** with parameters w, b, r. We denote this distribution by

NHGeom(w, b, r). The PMF is

$$P(X=k)=\frac{\binom{w}{r-1}\binom{b}{k}}{\binom{w+b}{r+k-1}}\frac{w-r+1}{w+b-r-k+1}=\frac{\binom{r+k-1}{r-1}\binom{w+b-r-k}{w-r}}{\binom{w+b}{w}},\quad k=0,1,2,\cdots,b$$

Expectation: $E(X)=rb/(w+1)$. Formula for variance is messy.

Explanation if PMF: In the urn context, $X=k$ means that the $(r+k)$th ball chosen is white and exactly $r-1$ of the first $r+k-1$ balls chosen are white. This gives the first equation.

Alternatively, we can imagine that we continue drawing balls until the urn has been emptied out; this is valid since whether or not we continue to draw balls after obtaining the rth white ball has no effect on X. Think of the $w+b$ balls as lined up in a random order, the order in which they will be drawn. Then $X=k$ means that among the first $r+k-1$ balls there are exactly $r-1$ white balls, then there is a white ball, and then among the last $w+b-r-k$ balls there are exactly $w-r$ white balls. All $\binom{w+b}{w}$ possibilities for the locations of the white balls in the line are equally likely. This gives the second equation.

We conclude four related trials below.

	With replacement	**Without replacement**
Fixed number of trials	Binomial	Hypergeometric
Fixed number of successes	Negative Binomial	Negative Hypergeometric

Figure 4.1: Four distributions and their relations

Expectation:

Assume we continue drawing balls until the urn is empty. First consider the case $r=1$. Label the black balls as $1,2,\cdots,b$, and let I_j be the indicator of black ball j being drawn before any white balls have been drawn. Then $P(I_j=1)=1/(w+1)$ since, listing out the order in which black ball j and the white balls are drawn (ignoring the other balls), all orders are equally likely by symmetry, and $I_j=1$ is equivalent to black ball j being first in this list. So by linearity,

$$E\left(\sum_{j=1}^{b} I_j\right)=\sum_{j=1}^{b}E(I_j)=\frac{b}{w+1}$$

For general r, write $X=X_1+X_2+\cdots+X_r$, where X_1 is the number of black balls before the first white ball, X_2 is the number of black balls after the first white ball but before the second white ball, etc. By essentially the same argument we used to handle the $r=1$ case, we have $E(X_j)=b/(w+1)$ for each j. So by linearity,

$$E\left(\sum_{j=1}^{r} X_j\right)=\sum_{j=1}^{r}E(X_j)=\frac{rb}{w+1}$$

4.8 Poisson

Definition 4.8.1: Poisson Distribution

An r.v. X has the **Poisson Distribution** with parameter λ, where $\lambda > 0$, if the PMF of X is

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \cdots$$

Denote as $X \sim \text{Pois}(\lambda)$.
Expectation: $E(X) = \lambda$. Variance: $\text{Var}(X) = \lambda$.

Expectation:

$$E(X) = e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty} k\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$$

where the second last equality holds since the sum is Taylor series for e^{λ}.

Variance:
Differentiate the Taylor series for e^{λ} with respect to λ,

$$\sum_{k=1}^{\infty} k\frac{\lambda^{k-1}}{k!} = e^{\lambda} \quad \Longrightarrow \quad \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!} = \lambda e^{\lambda}$$

Differentiate again:

$$\sum_{k=1}^{\infty} k^2\frac{\lambda^{k-1}}{k!} = (\lambda+1)e^{\lambda} \quad \Longrightarrow \quad \sum_{k=1}^{\infty} k^2\frac{\lambda^k}{k!} = \lambda(\lambda+1)e^{\lambda}$$

Therefore,

$$E(X^2) = e^{-\lambda}\sum_{k=0}^{\infty} k^2\frac{\lambda^k}{k!} = e^{-\lambda}\lambda(\lambda+1)e^{\lambda} = \lambda(\lambda+1)$$

Finally, the variance

$$\text{Var}(X) = E(X^2) - E(x)^2 = \lambda(\lambda+1) - \lambda^2 = \lambda$$

The sum of two independent Poisson r.v.s is still Poisson.

Proposition 4.8.2: Property of sum of Poisson

If $X \sim \text{Pois}(\lambda_1), Y \sim \text{Pois}(\lambda_2)$, and X is independent of Y, then

$$X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$$

Proof. Conditioning on X and using the law of total probability:

$$\begin{aligned} P(X+Y=k) &= \sum_{j=0}^{k} P(X+Y=k|X=j)P(X=j) = \sum_{j=0}^{k} P(Y=k-j)P(X=j) \\ &= \sum_{j=0}^{k} \frac{e^{-\lambda_2}\lambda_2^{k-j}}{(k-j)!}\frac{e^{-\lambda_1}\lambda_1^{j}}{j!} = \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{j=0}^{k}\binom{k}{j}\lambda_1^{j}\lambda_2^{k-j} = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^k}{k!} \end{aligned}$$

which is the PMF of Pois$(\lambda_1+\lambda_2)$. □

4.8.1 Connection between Binomial and Poisson

As the relation between binomial and hypergeometric, we can get binomial by conditioning on Poisson, and get Poisson by taking limit of binomial.

Theorem 4.8.3: Poisson to Binomial

If $X \sim \text{Pois}(\lambda_1), Y \sim \text{Pois}(\lambda_2)$, and X is independent of Y, then the conditional distribution

$$X|X+Y=n \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$$

Proof. Use Bayes' Rule,

$$\begin{aligned} P(X=k|X+Y=n) &= \frac{P(X+Y=n|X=k)P(X=k)}{P(X+Y=n)} = \frac{P(Y=n-k)P(X=k)}{P(X+Y=n)} \\ &= \frac{\left(\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}\right)\left(\frac{e^{-\lambda_1}\lambda_1^{k}}{k!}\right)}{\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}} = \binom{n}{k}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} = \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \end{aligned}$$

which is the PMF of Bin $\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$. □

Theorem 4.8.4: Binomial to Poisson

If $X \sim \text{Bin}(n,p)$ and we let $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remain fixed, then the PMF of X converges to that of Pois(λ). More generally, this holds if $n \to \infty$ and $p \to 0$ such that np converges to a constant λ.

Proof. We can use the equation

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

To get that,

$$\begin{aligned}\mathbb{P}(X=k) &= \binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\ &= \frac{1}{k!}\frac{n(n-1)......(n-k+1)}{n^k}(np)^k\left[\left(1-\frac{np}{n}\right)^{n\frac{n-k}{n}}\right] \\ &\implies \frac{1}{k!}\times 1\times \mu^k\times e^{-\mu}\sim \text{Pois}(\mu)\end{aligned}$$

Since $n(n-1)...(n-k+1)\to n^k$ as $n\to\infty$ and in the last term $\frac{n-k}{n}\to 1$ as $n\to\infty$ □

This last theorem implies the fact that, *the Poisson distribution is often used in situations where we are counting the number of successes in a particular region or interval of time, and there are a large number of trials, each with a small probability of success.*

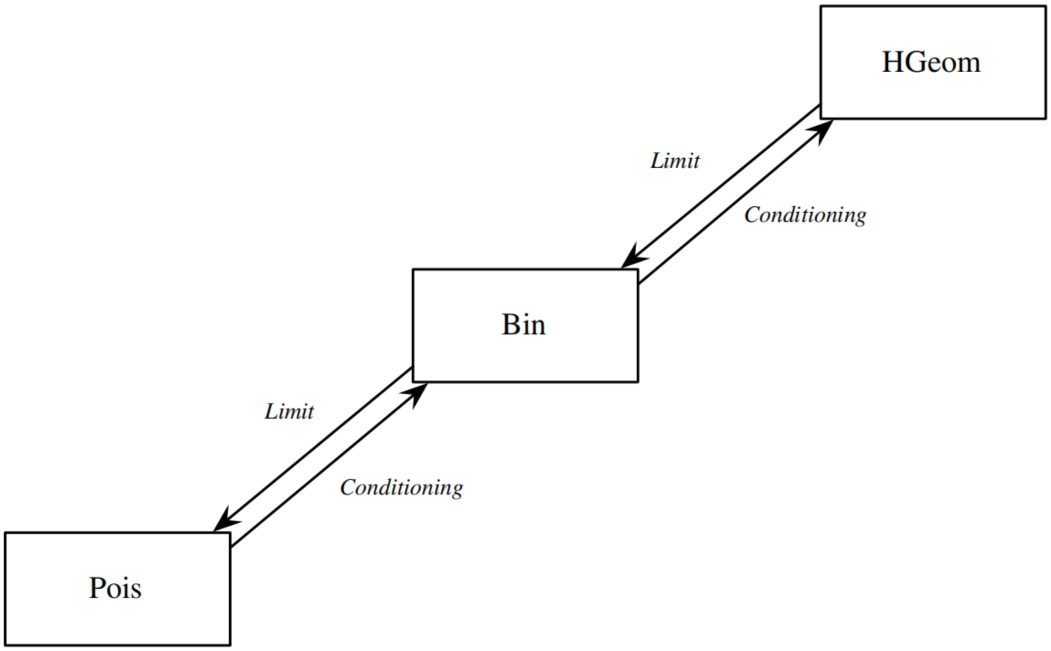


Figure 4.2: Relations between Poisson, Binomial and Hypergeometric

Chapter 5

Examples of Continuous Random Variables

5.1 Uniform

Definition 5.1.1: Uniform Distribution

A continuous r.v. U has **uniform distribution** on (a, b), denoted by $U \sim \text{Unif}(a, b)$ if its PDF is

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases}$$

The CDF is

$$F(x) = \begin{cases} 0, & \text{if } x \leqslant a \\ \dfrac{x-a}{b-a}, & \text{if } a < x < b \\ 1, & \text{if } x \geqslant b \end{cases}$$

Expectation: $E(U) = (a + b)/2$. Variance: $\text{Var}(U) = (b - a)^2/12$.

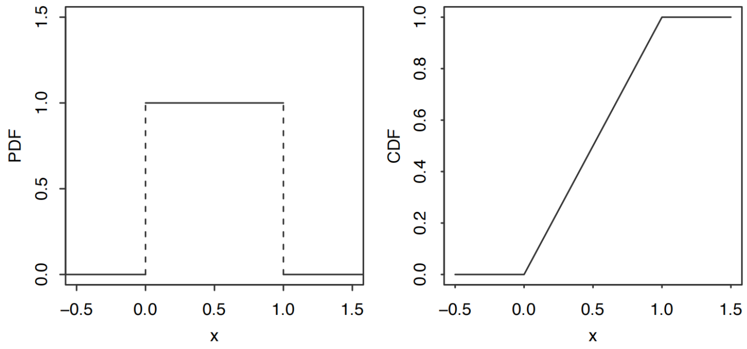


Figure 5.1: Uniform Distribution: PDF and CDF $U \sim \text{Unif}(0, 1)$

Expectation: We first note that for $U_0 \sim \mathrm{Unif}(0,1)$, we have

$$E(U_0) = \int_0^1 x \, \mathrm{d}x = \frac{1}{2}, \quad E(U_0^2) = \int_0^1 x^2 \, \mathrm{d}x = \frac{1}{3}, \quad \mathrm{Var}(U) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Since $U = a + (b-a)U_0$, we have

$$E(U) = a + (b-a)E(U_0) = \frac{a+b}{2}$$

Variance: Similarly, using the transformation and property of variance:

$$\mathrm{Var}(U) = (b-a)^2 \mathrm{Var}(U_0) = \frac{(b-a)^2}{12}$$

5.1.1 Universality of the Uniform

Given a $\mathrm{Unif}(0,1)$ r.v., we can construct an r.v. with any continuous distribution we want. Conversely, given an r.v. with an arbitrary continuous distribution, we can create a $\mathrm{Unif}(0,1)$ r.v.

Theorem 5.1.2: Universality of the Uniform

Let F be a CDF which is a continuous function and strictly increasing on the support. This ensures F^{-1} exists, as a function from $(0,1)$ to $\mathbb{R}$. Then,

1. Let $U \sim \mathrm{Unif}(0,1)$ and $X = F^{-1}(U)$. Then X is an r.v. with CDF F.
2. Let X be an r.v. with CDF F. Then $F(X) \sim \mathrm{Unif}(0,1)$.

Proof. 1. Let $U \sim \mathrm{Unif}(0,1)$ and $X = F^{-1}(U)$. For all real x,

$$P(X \leqslant x) = P(F^{-1}(U) \leqslant x) = P(U \leqslant F(x)) = F(x)$$

so the CDF of X is F.

2. Let X have CDF F. Let $Y = F(X)$. Since Y takes values in $(0,1)$, its CDF is 0 for $y \leqslant 0$ and 1 for $y \geqslant 1$. Then, for $y \in (0,1)$,

$$P(Y \leqslant y) = P(F(X) \leqslant y) = P(X \leqslant F^{-1}(y)) = F(F^{-1}(y)) = y$$

Thus Y has the $\mathrm{Unif}(0,1)$ distribution. □

This can be also generalized to discrete case.

Theorem 5.1.3: Universality of the Uniform: Discrete Case

Let $U \sim \mathrm{Unif}(0,1)$ and Y be a discrete random variable with CDF F. Let

$$F^{-1}(u) = \min\{t \in \mathbb{R} : F(t) \geqslant u\}$$

be the *generalized inverse of* F. Then, $X = F^{-1}(U)$ has CDF F.

Proof. Instead of working with CDF, we work with PMF to solve this. The proof using generalized inverse can be found in Appendix.
Suppose we want to use $U \sim \text{Unif}(0,1)$ to construct a discrete r.v. X with PMF $p_j = P(X = j)$ for $j = 0, 1, 2, \cdots, n$. we can chop up the interval $(0, 1)$ into pieces of lengths $p_0, p_1, \cdots, p_n$. By the properties of a valid PMF, the sum of the p_j's is 1, so this perfectly divides up the interval, without overshooting or undershooting.

Now define X to be the r.v. which equals 0 if U falls into the p_0 interval, 1 if U falls into the p_1 interval, 2 if U falls into the p_2 interval, and so on. Then X is a discrete r.v. taking on values 0 through n. The probability that $X = j$ is the probability that U falls into the interval of length p_j . But for a Unif$(0, 1)$ r.v., probability is length, so $P(X = j)$ is precisely p_j , as desired.

The same trick will work for a discrete r.v. that can take on infinitely many values, such as a Poisson; we'll need to chop $(0, 1)$ into infinitely many pieces, but the total length of the pieces is still 1.

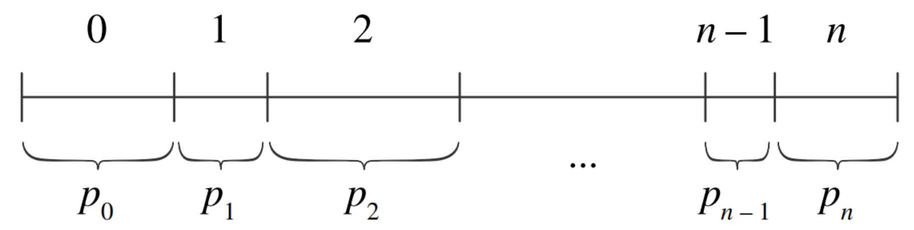


Figure 5.2: Chop into different intervals

□

5.1.2 Expectation by Integrating the Survival Function

Theorem 5.1.4: Expectation via Survival Function

Let X be a nonnegative r.v. Then,

$$E(X) = \int_0^\infty P(X > x)\,\mathrm{d}x$$

Note that this holds for any nonnegative r.v., not just continuous one.

Proof. For any number $x \geqslant 0$, we can write

$$x = \int_0^x \mathrm{d}t = \int_0^\infty I(x > t)\,\mathrm{d}t$$

where $I(x > t) = 1$ if $x > t$ and 0 otherwise. So,

$$X(s) = \int_0^\infty I(X(s) > t)\,\mathrm{d}t$$

for each $s \in S$. We can write this more compactly as

$$X = \int_0^\infty I(X > t)\,\mathrm{d}t$$

Taking the expectation of both sides and swapping the E wit the integral (This can be proved feasible via *Fubini's Theorem* in *Real Analysis*), we have

$$E(X) = E\left(\int_0^\infty I(X > t)\,\mathrm{d}t\right) = \int_0^\infty E(I(X > t))\,\mathrm{d}t = \int_0^\infty P(X > t)\,\mathrm{d}t$$

where the last equation comes from the fundamental bridge of indicator variable. □

5.2 Normal

Definition 5.2.1: Standard Normal Distribution

A continuous r.v. Z is said to have the **standard normal distribution** $Z \sim N(0,1)$, if its PDF φ is

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

The CDF is

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t$$

Expectation: $E(Z) = 0$. Variance: $\mathrm{Var}(Z) = 1$.

CDF integration to 1:

$$\left(\int_{-\infty}^\infty e^{-z^2/2}\,\mathrm{d}z\right)^2 = \left(\int_{-\infty}^\infty e^{-x^2/2}\,\mathrm{d}x\right)\left(\int_{-\infty}^\infty e^{-y^2/2}\,\mathrm{d}y\right) = \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\mathrm{d}y$$
$$= \int_0^{2\pi} \int_0^\infty r e^{-r^2/2}\,\mathrm{d}r\mathrm{d}\theta = \int_0^{2\pi} \left[-e^{-r^2/2}\right]_{r=0}^{\infty} \mathrm{d}\theta = 2\pi$$

Therefore,

$$\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t = 1$$

Expectation:

$$E(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty z e^{-z^2/2}\,\mathrm{d}t$$

the integrand is odd function, thus it leads to 0.

Variance:

$$\begin{aligned}\operatorname{Var}(Z) = E(Z^2) - E(Z)^2 = E(Z^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2}\,\mathrm{d}z = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} z^2 e^{-z^2/2}\,\mathrm{d}z \\ &= \frac{2}{\sqrt{2\pi}} \left(\left[-ze^{-z^2/2} \right]_0^{\infty} + \int_0^{\infty} e^{-z^2/2}\,\mathrm{d}z \right) = \frac{2}{\sqrt{2\pi}} \left(0 + \frac{\sqrt{2\pi}}{2} \right) = 1\end{aligned}$$

Definition 5.2.2: Normal Distribution

If $Z \sim N(0,1)$, then $X = \mu + \sigma Z$ with $\sigma > 0$ is said to have **normal distribution**, denoted by $X \sim N(\mu, \sigma^2)$. The CDF is

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

the PDF is

$$f(x) = \varphi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Expectation: $E(X) = \mu$. Variance: $\operatorname{Var}(X) = \sigma^2$.

It follows the *quantile rule:*

$$P(|X-\mu| < \sigma) \approx 0.68$$

$$P(|X-\mu| < 2\sigma) \approx 0.95$$

$$P(|X-\mu| < 3\sigma) \approx 0.997$$

5.3 Exponential

The Exponential distribution is the continuous counterpart to the Geometric distribution. Recall that a Geometric random variable counts the number of failures before the first success in a sequence of Bernoulli trials. The story of the Exponential distribution is analogous, but we are now waiting for a success in continuous time, where successes arrive at a rate of λ successes per unit of time. An Exponential random variable represents the waiting time until the first arrival of a success.

Definition 5.3.1: Exponential Distribution

A continuous r.v. X is **exponentially distributed** with parameter $\lambda > 0$, $X \sim \operatorname{Exp}(\lambda)$ if its PDF is

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

The CDF is

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0$$

Expectation: $E(X) = 1/\lambda$. Variance: $\operatorname{Var}(X) = 1/\lambda^2$.

Proposition 5.3.2

If $X \sim \text{Exp}(\lambda)$, then $Y = aX \sim \text{Exp}(\lambda/a)$ with $a > 0$.

Proof.

$$P(Y \leqslant y) = P(aX \leqslant y) = P(X \leqslant y/a) = 1 - e^{-\lambda y/a}$$

as desired. □

5.3.1 Memoryless Property

Theorem 5.3.3: Memoryless Property

For Exponential random variable X, we have

$$P(X \geqslant s + t | X \geqslant s) = P(X \geqslant t), \quad s, t \geqslant 0$$

Moreover, exponential r.v. is the only continuous r.v. that has this property.

The proof is similar to that of geometric r.v..

5.3.2 Introduction to Poisson Process

Definition 5.3.4: Poisson Process

A process of arrival in continuous time is called **Poisson Process** with rate λ if

- The number of arrivals that occur in an interval of length t is a $\text{Pois}(\lambda t)$ random variable.
- The number of arrivals that occur in disjoint intervals are independent of each other.

In this section, we will focus on Poisson processes on $(0, \infty)$.

Suppose that the arrivals are emails landing in an inbox according to a Poisson process with rate λ. One question we could ask is:

In one hour, how many emails will arrive?

The answer comes directly from the definition, which tells us that the number of emails in an hour follows a $\text{Pois}(\lambda)$ distribution. But we could also flip the question around and ask:

how long does it take until the first email arrives (measured relative to some fixed starting point)?

Let T_1 be the time until the first email arrives. To find the distribution of T_1, we just need to understand one crucial fact: saying that the waiting time for the first email is greater than t is the same as saying that no emails have arrived between 0 and t. In other words, if N_t is the number of emails that arrive at or before time t, then

$$T_1 > t \text{ is the same event as } N_t = 0$$

Similarly, we have

$$T_n > t \text{ is the same event as } N_t < n$$

This is called the *Count-time duality.* If two events are the same, they have the same probability. Therefore,

$$P(T_1 > t) = P(N_t = 0) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}$$

Therefore T_1 is a exponential distribution with parameter λ!

What about $T_2 - T_1$, the time between the first and second arrivals? Since disjoint intervals in a Poisson process are independent by definition, the past is irrelevant once the first arrival occurs. Thus $T_2 - T_1$ is independent of the time until the first arrival, and by the same argument as before, $T_2 - T_1$ also has an Exponential distribution with rate λ.

Continuing in this way, we deduce that all the interarrival times are i.i.d. Expo(λ) random variables. Thus, Poisson processes tie together two important distributions, one discrete and one continuous.

5.4 Symmetry of i.i.d. Continuous r.v.s

Theorem 5.4.1: Symmetry of i.i.d. Continuous r.v.s

Let $X_1, X_2, \cdots, X_n$ be i.i.d from a continuous distribution. Then,

$$P(X_{a_1} < X_{a_2} < \cdots < X_{a_n}) = \frac{1}{n!}$$

for any permutation $a_1, a_2, \cdots, a_n$ of $1, 2, \cdots, n$.

Proof. Let F be the CDF of X_j. By symmetry, all orderings of $X_1, X_2, \cdots, X_n$ are equally likely. The probability of the tie $P(X_i = X_j) = 0$ since they are independent continuous r.v.s. So the probability of there being at least one tie among $X_1, \cdots, X_n$ is also 0, since

$$P\left(\bigcup_{i \neq j} \{X_i = X_j\}\right) \leqslant \sum_{i \neq j} P(X_i = X_j) = 0$$

Thus, $X_1, X_2, \cdots, X_n$ are distinct with probability 1, and the probability of any permutation is $1/n!$, □

Note: This does not fit in discrete case since there may be ties that with probability larger than 0 for discrete variables.

Example 5.4.2: Record of jumping

Athletes compete one at a time at the high jump. Let X_j be how high the jth jumper jumped, with $X_1, X_2, \cdots$ i.i.d. with a continuous distribution. We say that the jth jumper sets a record if X_j is greater than all of $X_{j-1}, \cdots, X_1$.

1. Is the event '110th jumper sets a record' independent of '111th jumper sets a record'?

2. Find the mean number of records among the first n jumpers. What happens for $n \to \infty$?

3. A *double record* occurs at time j if both the jth and $(j-1)$th jumper sets records. Find the mean number of double records among the first n jumpers. What happens for $n \to \infty$?

Proof. 1. Let I_j be the indicator r.v. for the jth jumper setting a record. By symmetry, $P(I_j = 1) = 1/j$. Also,

$$P(I_{110} = 1, I_{111} = 1) = \frac{109!}{111!} = \frac{1}{100 \times 111}$$

since in order for both jumpers set records, we need the highest of the first 111 jumps to be in position 111 and second highest in position 110, and the remaining 109 can be in any order. So,

$$P(I_{110} = 1, I_{111} = 1) = P(I_{110} = 1)P(I_{111} = 1)$$

thus they are independent.

2. By linearity, the expected number of records is $\sum_{j=1}^{n} \frac{1}{j}$, and

$$\lim_{n\to\infty} \sum_{j=1}^{n} \frac{1}{j} = \infty$$

3. Let J_i be the indicator r.v. for a double record occuring at time j, for $2 \leqslant j \leqslant n$. Then

$$P(J_i = 1) = \frac{1}{j(j-1)}$$

following part (1). So the expected number is

$$\sum_{j=2}^{n} \frac{1}{j(j-1)} = \sum_{j=2}^{n} \left(\frac{1}{j-1} - \frac{1}{j} \right) = 1 - \frac{1}{n}$$

As $n \to \infty$, it goes to 1.

□

Chapter 6

Moments

6.1 Measure of Central Tendency

Definition 6.1.1: Median

c is a **median** of a r.v. X if $P(X \leqslant c) \geqslant 1/2$ and $P(X \geqslant c) \geqslant 1/2$.

Definition 6.1.2: Mode

- For a discrete r.v. X, we say c is a **mode** of X if $P(X = c) \geqslant P(X = x)$ for all x.
- For a continuous r.v. X with PDF f, we say c is the **mode** if $f(c) \geqslant f(x)$ for all x.

Note: A distribution can have multiple medians and multiple modes. Below shows a distribution with median $[-1, 1]$ and modes $-3, 3$.

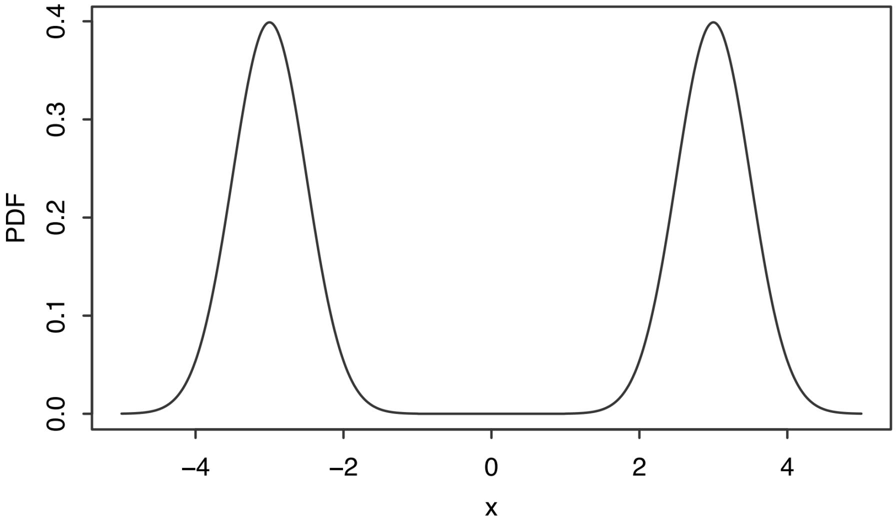


Figure 6.1: Distribution with multiple medians and multiple modes

Proposition 6.1.3

Let X be an r.v. with mean μ and a median m.

1. The value of c that minimizes $E[(X-c)^2]$ is $c=\mu$.
2. The value of c that minimizes $E[|X-c|]$ is $c=m$.

Proof.

1. Note that
$$\mathrm{Var}(X)=\mathrm{Var}(X-c)=E[(X-c)^2]-E(X-c)^2=E[(X-c)^2]-(\mu-c)^2$$
Since $\mathrm{Var}(X)\geqslant 0$, and $(\mu-c)^2\geqslant 0$, to minimize $E[(X-c)^2]$, we need
$$E[(X-c)^2]=\mathrm{Var}(X)+(\mu-c)^2\geqslant \mathrm{Var}(X)$$
with equality if and only if $\mu=c$.

2. Let $a\neq m$, we need to show that $E(|X-m|)\leqslant E(|X-a|)$, which is equivalent to $E(|X-a|-|X-m|)\geqslant 0$. Assume that $m<a$ (the case $m>a$ can be handled similarly), If $X\leqslant m$ then
$$|X-a|-|X-m|=a-X-m+X=a-m$$
and if $X>m$ then
$$|X-a|-|X-m|\geqslant X-a-X+m=m-a$$
Let $Y=|X-a|-|X-m|$, we can split definition of Y into two parts using indicator r.v.s. Let I be the indicator for $X\leqslant m$. Then,
$$\begin{aligned}E(Y)=E(Y)+E(Y(1-I))&\geqslant(a-m)E(I)+(m-a)E(1-I)\\&=(a-m)P(X\leqslant m)+(m-a)P(X>m)=(a-m)P(X\leqslant m)-(a-m)(1-P(X\leqslant m))\\&=(a-m)(2P(X\leqslant m)-1)\end{aligned}$$
By definition of median, we have $2P(X\leqslant m)-1\geqslant 0$. Thus, $E(Y)\geqslant 0$, with equality when $a=m$.

□

6.2 Skewness and Kurtosis

Definition 6.2.1: Moments

Let X be an r.v. with mean μ and variance σ^2. For any positive integer n,

- The **nth moment** of X is $E(X^n)$ if exists.
- The **nth central moment** is $E((X-\mu)^n)$ if exists.

- The **nth standardized moment** is $E\left(\left(\frac{X-\mu}{\sigma}\right)^n\right)$ if exists.

In particular, mean is the first moment, and variance is the second central moment.

Definition 6.2.2: Skewness

The **skewness** of an r.v. X with mean μ and variance σ^2 is the third standardized moment of X:

$$\text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

Positive skewness is indicative of having a long right tail relative to the left tail, and negative skewness is indicative of the reverse.

Definition 6.2.3: Kurtosis/Excess Kurtosis

The **Kurtosis** of an r.v. X with mean μ and variance σ^2 is

$$\text{Kurt}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$$

The **excess kurtosis** is

$$\text{Kurt}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

The reason for substracting 3 is that this make any normal distribution have kurtosis 0 *for comparison.*

6.3 Sample Moments

Definition 6.3.1: Sample Moments

Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables. The kth **sample moment** is the r.v.

$$M_k = \frac{1}{n}\sum_{j=1}^{n} X_j^k$$

By law of large numbers, which will be introduced later, these sample moments will converges to their corresponding true moments, and also the expectation

$$E\left(\frac{1}{n}\sum_{j=1}^{n} X_j^k\right) = \frac{1}{n}(E(X_1^k) + \cdots + E(X_n^k)) = E(X_1^k)$$

Definition 6.3.2: Sample Mean and its variance

Let $X_1, X_2, \cdots, X_n$ be i.i.d. r.v.s. with mean μ and variance σ^2. Then, the **sample mean** is unbiased for estimating μ,

$$E(\bar{X}_n) = \mu$$

The variance of the sample mean is

$$\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

Proof.

$$E(\bar{X}_n) = \frac{1}{n}(E(X_1) + E(X_2) + \cdots + E(X_n)) = E(X_1) = \mu$$

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2}(\text{Var}(X_1) + \cdots + \text{Var}(X_n)) = \frac{n}{n^2}\text{Var}(X_1) = \frac{\sigma^2}{n}$$

□

Definition 6.3.3: Sample variance and its unbiasedness

Let $X_1, X_2, \cdots, X_n$ be i.i.d. r.v.s. with mean μ and variance σ^2. The **sample variance** is the r.v.

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$$

the **sample standard deviation** is the square root of the sample variance. Moreover, the sample variance is unbiased,

$$E(S_n^2) = \sigma^2$$

Proof.

$$\begin{aligned}
E[S_n^2] &= E\left[\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2\right] = \frac{1}{n-1}E\left[\sum_{j=1}^{n}(X_j - \bar{X})^2\right] \overset{\text{linearity of expectation}}{=} \frac{1}{n-1}\sum_{j=1}^{n}E\left[(X_j - \bar{X})^2\right] \\
&= \frac{1}{n-1}\sum_{j=1}^{n}E\left[(X_j - \mu - \bar{X} + \mu)^2\right] = \frac{1}{n-1}\sum_{j=1}^{n}E\left[(X_j - \mu)^2 - 2(X_j - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2\right] \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left(E\left[(X_j - \mu)^2\right] + 2E\left[(X_j - \mu)(\mu - \bar{X})\right] + E\left[(\mu - \bar{X})^2\right]\right)
\end{aligned}$$

Now we discuss these 3 terms separately.

$$E[(X_j - \mu)^2] = Var[X_j] = \sigma^2$$

$$E\left[(X_j - \mu)(\mu - \bar{X})\right] = -(E[X_i\bar{X}] - \mu E[X_i] - \mu E[\bar{X}] + \mu^2) = -(E[X_i\bar{X}] - \mu^2)$$

For $E[X_i\bar{X}]$,

$$E[X_i\bar{X}] = E\left[\frac{1}{n}X_i\sum_{j=1}^{n}X_j\right] = \frac{1}{n}\left(E[X_iX_1] + E[X_iX_2] + ... + E[X_iX_n]\right)$$

when $j \neq i$, X_j, X_i are independent, then

$$E[X_i X_j] = E[X_i]E[X_j] = \mu^2$$

when $j = i$,

$$E[X_i X_j] = E[X_i^2] = Var[X_i] + (E[X_i])^2 = \sigma^2 + \mu^2$$

Thus,

$$E[X_i \bar{X}] = \frac{1}{n}\left((n-1)\mu^2 + \mu^2 + \sigma^2\right) = \frac{1}{n}(n\mu^2 + \sigma^2) = \mu^2 + \frac{\sigma^2}{n}$$

$$E\left[(\mu - \bar{X})^2\right] = Var[\bar{X}] = \frac{\sigma^2}{n}$$

Finally,

$$\begin{aligned} E[S_n^2] &= \frac{1}{n-1}\sum_{j=1}^{n}\left(E\left[(X_j-\mu)^2\right] + 2E\left[(X_j-\mu)(\mu-\bar{X})\right] + E\left[(\mu-\bar{X})^2\right]\right) \\ &= \frac{1}{n-1}\sum_{j=1}^{n}\left(\sigma^2 - 2\mu^2 - 2\frac{\sigma^2}{n} + 2\mu^2 + \frac{\sigma^2}{n}\right) \overset{\text{constant sum}}{=} \frac{1}{n-1}n\frac{n-1}{n}\sigma^2 = \sigma^2 \end{aligned}$$

Thus the sample variance is not biased. □

Similarly, the *sample skewness* is

$$\frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3}{S_n^3}$$

and the *sample kurtosis* is

$$\frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^4}{S_n^4} - 3$$

6.4 Moment Generating Function (MGF)

Definition 6.4.1: Moment Generating Function (MGF)

The **moment generating function** of an r.v. X is

$$M(t) = E(e^{tX})$$

as a function of t, if this is finite on some open interval $(-a, a)$. Otherwise we say MGF of X does not exist.

We see some properties of MGF before seeing examples. Why it is called 'moment generating'? This will be answered in the next theorem.

Theorem 6.4.2: Moment Generating Property

Given MGF of X, we have

$$E(X^n) = M^{(n)}(0)$$

Proof. This can be seen by noting that the Taylor expansion of $M(t)$ at 0 is

$$M(t) = \sum_{n=0}^{\infty} M^{(n)}(0)\frac{t^n}{n!}$$

while on the other hand, we have

$$M(t) = E(e^{tX}) = E\left(\sum_{n=0}^{\infty} X^n \frac{t^n}{n!}\right)$$

We are allowed to interchange the expectation and the infinite sum (uniform convergence), so

$$M(t) = \sum_{n=0}^{\infty} E(X^n)\frac{t^n}{n!}$$

Matching the coefficients, we have $E(X^n) = M^{(n)}(0)$. □

Theorem 6.4.3: MGF determines the distribution

The MGF of a random variable determines its distribution. If two r.v.s have the same MGF, they must have the same distribution.

The proof is difficult analysis problem, using *Laplace Transform*. We will not address it here.

Theorem 6.4.4: MGF of sum of independent r.v.s

If X and Y are independent, then

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

This is true by definition:

$$E(e^{tX+tY}) = E(e^{tX})E(e^{tY})$$

Theorem 6.4.5: Transformation of MGF

If X has MGF $M(t)$, then the MGF of $a + bX$ is

$$E(e^{t(a+bX)}) = e^{at}E(e^{btX}) = e^{at}M(bt)$$

Now we see examples:

PMF for Bernoulli: $X \sim \text{Ber}(p)$,

$$M(t) = pe^t + (1-p)$$

Proof. e^{tX} takes on the value e^t with probability p and the value 1 with proability $1-p$, thus,

$$M(t) = E(e^{tX}) = pe^t + (1-p)$$

□

PMF for Binomial: $X \sim \text{Bin}(n, p)$

$$M(t) = (pe^t + (1-p))^n$$

Proof. This directly follows from Theorem 6.4.4. □

PMF for Geometric: $X \sim \text{Geom}(p)$,

$$M(t) = \frac{p}{1-(1-p)e^t}, \quad t < -\ln(1-p)$$

Proof.

$$M(t) = E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}(1-p)^k p = p\sum_{k=0}^{\infty}((1-p)e^t)^k = \frac{p}{1-(1-p)e^t}$$

□

PMF for Negative Binomial: $X \sim \text{NegBin}(r, p)$,

$$M(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \quad t < -\ln(1-p)$$

Proof. This directly follows from Theorem 6.4.4. □

PMF for Uniform: $U \sim \text{Unif}(a, b)$,

$$M(t) = \begin{cases} \dfrac{e^{tb}-e^{ta}}{t(b-a)}, & \text{if } t \neq 0 \\ 1, & \text{if } t = 0 \end{cases}$$

Proof.

$$M(t) = E(e^{tU}) = \frac{1}{b-a}\int_a^b e^{tu}\,\mathrm{d}u = \frac{e^{tb}-e^{ta}}{t(b-a)}$$

□

PMF for Poisson: $X \sim \text{Pois}(\lambda)$,

$$M(t) = e^{\lambda(e^t-1)}$$

Proof.

$$E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} = e^{-\lambda}e^{\lambda e^t}$$

□

PMF for Normal: $X \sim N(\mu, \sigma^2)$,

$$M(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Proof. The MGF for standard normal is

$$M_Z(t) = E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\,\mathrm{d}z = e^{t^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(z-t)^2/2}\,\mathrm{d}z = e^{t^2/2}$$

since the $N(t, 1)$ PDF integrates to 1. Thus, the MGF of X will be

$$M(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

which follows from Theorem 6.4.5 □

PMF for Exponential: $X \sim \text{Exp}(\lambda)$,

$$M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda$$

Proof. The MGF of $Y \sim \text{Exp}(1)$ is

$$M_Y(t) = E(e^{tY}) = \int_0^\infty e^{tx} e^{-x} \, \mathrm{d}x = \int_0^\infty e^{-x(1-t)} \, \mathrm{d}x = \frac{1}{1-t}, \quad t < 1$$

So $X = Y/\lambda$ MGF is

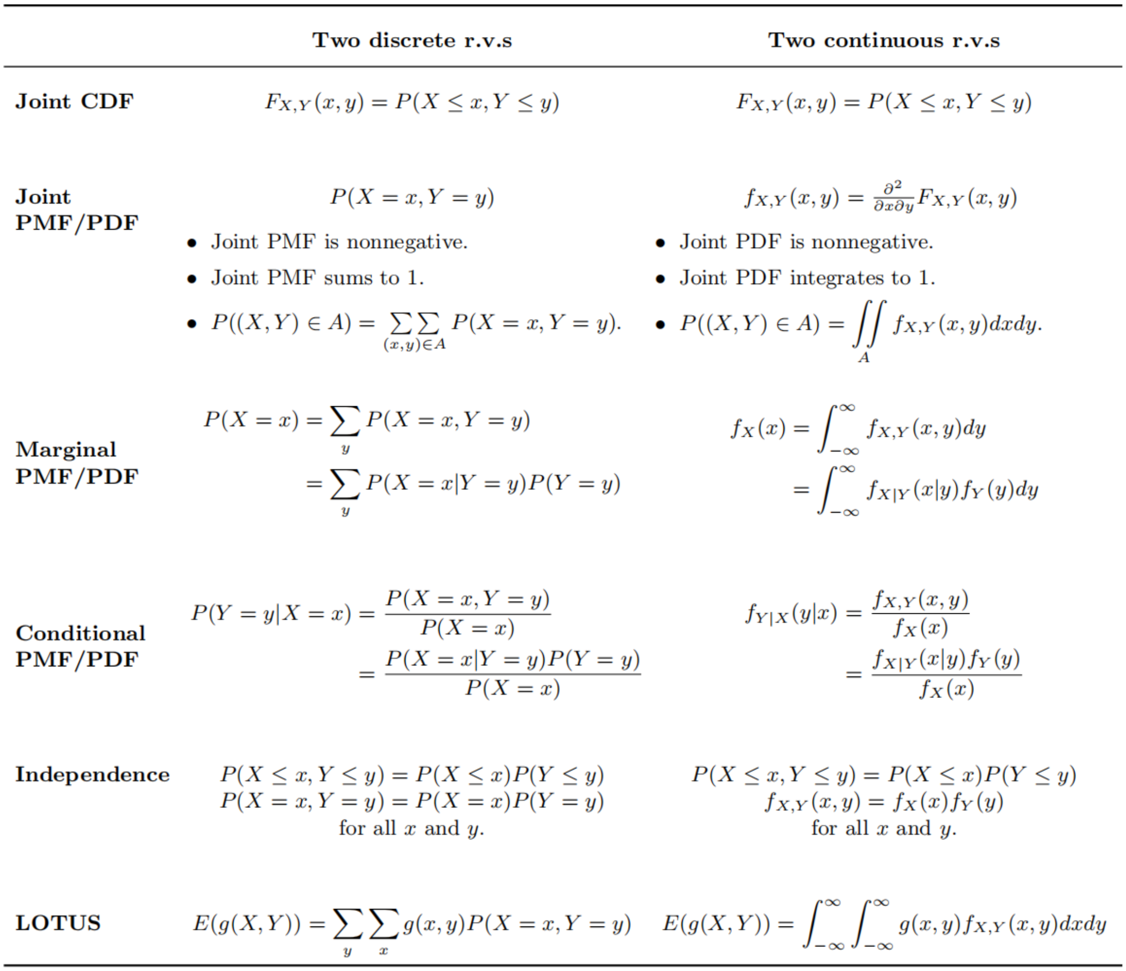
$$M(t) = M_Y\left(\frac{t}{\lambda}\right) = \frac{\lambda}{\lambda - t}$$

□

Chapter 7

Joint Distributions

7.1 Joint, Marginal and Conditional



	Two discrete r.v.s	**Two continuous r.v.s**
Joint CDF	$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$	$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$
Joint PMF/PDF	$P(X = x, Y = y)$ • Joint PMF is nonnegative. • Joint PMF sums to 1. • $P((X,Y) \in A) = \sum\sum_{(x,y)\in A} P(X = x, Y = y)$.	$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$ • Joint PDF is nonnegative. • Joint PDF integrates to 1. • $P((X,Y) \in A) = \iint_A f_{X,Y}(x,y)dxdy$.
Marginal PMF/PDF	$P(X = x) = \sum_y P(X = x, Y = y)$ $= \sum_y P(X = x \mid Y = y)P(Y = y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ $= \int_{-\infty}^{\infty} f_{X\mid Y}(x\mid y)f_Y(y)dy$
Conditional PMF/PDF	$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$ $= \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$	$f_{Y\mid X}(y\mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ $= \frac{f_{X\mid Y}(x\mid y)f_Y(y)}{f_X(x)}$
Independence	$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x and y.	$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x and y.
LOTUS	$E(g(X,Y)) = \sum_y \sum_x g(x,y)P(X = x, Y = y)$	$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy$

Figure 7.1: Summary of 2D case

Four versions of Bayes' Rule:

	Y **discrete**	Y **continuous**
X **discrete**	$P(Y=y\|X=x)=\frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y\|X=x)=\frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X **continuous**	$P(Y=y\|X=x)=\frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y\|x)=\frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

Four versions of LOTP:

	Y **discrete**	Y **continuous**
X **discrete**	$\sum_y P(X=x\|Y=y)P(Y=y)$	$\int_{-\infty}^{\infty} P(X=x\|Y=y)f_Y(y)dy$
X **continuous**	$\sum_y f_X(x\|Y=y)P(Y=y)$	$\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)f_Y(y)dy$

Sometimes we have a joint PDF for X and Y that factors as a function of x times a function of Y, without knowing in advance whether these functions are the marginal PDFs, or even whether they are valid PDFs. The next result addresses this situation.

Theorem 7.1.1: Factoring joint PDF

Suppose that the joint PDF $f_{X,Y}$ of X and Y factors as

$$f_{X,Y}(x,y)=g(x)h(y)$$

for all x and y, where g and h are nonnegative functions. Then X and Y are independent. Also, if either g or h is a valid PDF, then the other one is a valid PDF too, and g and h are the marginal PDFs of X and Y, respectively. (The analogous result in discrete case also holds)

Proof. Let $c=\int_{-\infty}^{\infty} h(y)\,\mathrm{d}y$. Multiplying and dividing by c, we can write as

$$f_{X,Y}(x,y)=cg(x)\cdot\frac{h(y)}{c}$$

Then $h(y)/c$ is a valid PDF. Then the marginal PDF of X is

$$f_X(x)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y=cg(x)\int_{-\infty}^{\infty}\frac{h(y)}{c}\,\mathrm{d}y=cg(x)$$

It follows that $\int_{-\infty}^{\infty} cg(x) = 1$ since a marginal PDF is a valid PDF. Then the marginal PDF of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \frac{h(y)}{c}\int_{-\infty}^{\infty} cg(x)\,\mathrm{d}x = \frac{h(y)}{c}$$

Thus, X and Y are independent. □

7.2 Covariance and Correlation

Definition 7.2.1: Covariance

The **covariance** between r.v.s. X and Y is

$$\mathrm{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

Theorem 7.2.2: Independence and Correlation

If X and Y are independent, then they are uncorrelated.

Proof. We just show the continuous case. Discrete case is similar. Use LOTUS,

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\,\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty} yf_Y(y)\left(\int_{-\infty}^{\infty} xf_X(x)\,\mathrm{d}x\right)\mathrm{d}y \\ &= \int_{-\infty}^{\infty} xf_X(x)\,\mathrm{d}x \int_{-\infty}^{\infty} yf_Y(y)\,\mathrm{d}y = E(X)E(Y) \end{aligned}$$

□

Note: The converse is false. Let $X \sim N(0,1)$ and $Y = X^2$, then $E(XY) = E(X^3) = 0$. Thus they are uncorrelated. But they are dependent. *Covariance is a measure of linear association.*

Few properties of Correlation: (Proofs are easy and omitted)

- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(X,c) = 0$, $\forall\, c \in \mathbb{R}$
- $\mathrm{Cov}(aX,Y) = a\mathrm{Cov}(X,Y)$, $\forall\, a \in \mathbb{R}$
- $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$
- $\mathrm{Cov}(X+Y,Z+W) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z) + \mathrm{Cov}(X,W) + \mathrm{Cov}(Y,W)$
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X+Y)$
- $\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) + 2\sum_{i<j}\mathrm{Cov}(X_i,X_j)$

- $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X + Y)$

Definition 7.2.3: Correlation

The **correlation** between r.v.s X and Y is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Note that for correlation:

- $\text{Corr}(cX, Y) = \text{Corr}(X, Y)$
- $-1 \leqslant \text{Corr}(X, Y) \leqslant 1$

7.3 Multinomial

Whereas the Binomial distribution counts the successes in a fixed number of trials that can only be categorized as success or failure, the Multinomial distribution keeps track of trials whose outcomes can fall into multiple categories, such as excellent, adequate, poor; or red, yellow, green, blue.

Definition 7.3.1: Multinomial Distribution

Each of n objects is independently placed int one of k categories. An object is placed into category j with probability p_j, where p_j are nonnegative and $\sum_{j=1}^{k} p_j = 1$. Let X_1 be the number of objects in category 1, X_2 be the number of objects in category 2, etc, so that $X_1 + \cdots + X_k = n$. Then,

$$\mathbf{X} = (X_1, X_2, \cdots, X_k)$$

is said to have **multinomial distribution** with parameter n and $\mathbf{p} = (p_1, p_2, \cdots, p_k)$, written as $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$. The PMF is

$$P(X_1 = n_1, \cdots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} \cdot p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

for $n_1, \cdots, n_k$ satisfying $n_1 + \cdots + n_k = n$.

The derivation of this PMF should remind you of the permutation of letters in STATISTICS.

Proposition 7.3.2: Marginal Distribution of Multinomial

The marginals of a Multinomial are binomial. Specifically, if $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then $X_j \sim \text{Bin}(n, p_j)$.

Proof. X_j is the number of objects in category j, where each of the n objects independently belongs to category j with probability p_j. Define success as landing in category j. Then we just have n independent Bernoulli trials, so the marginal distribution of X_j is $\text{Bin}(n, p_j)$. □

This thought leads to another corollary.

Corollary 7.3.3: Multinomial Lumping

If $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then for any distinct i and j, $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$. The random vector of counts obtained from merging categories i and j is still Multinomial. For example, merging categories 1 and 2 gives

$$(X_1 + X_2, X_3, \cdots, X_k) \sim \text{Mult}_{k-1}(n, (p_1 + p_2, p_3, \cdots, p_k))$$

Now we examine the conditional distribution.

Proposition 7.3.4: Multinomial Conditioning

If $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then

$$(X_2, X_3, \cdots, X_k)|X_1 = n_1 \sim \text{Mult}_{k-1}(n - n_1, (p_2', \cdots, p_k'))$$

where $p_j' = p_j/(p_2 + \cdots + p_k)$.

This is clear from intuition and story of Multinomial. Finally we examine the covariance.

Proposition 7.3.5: Multinomial Covariance

Let $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$. Then, for $i \neq j$,

$$\text{Cov}(X_i, X_j) = -np_ip_j$$

Proof. For concreteness, let $i = 1$ and $j = 2$. We know $X_1 + X_2 \sim \text{Bin}(n, p_1 + p_2)$, $X_1 \sim \text{Bin}(n, p_1)$ and $X_2 \sim \text{Bin}(n, p_2)$. Therefore,

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$

becomes

$$n(p_1 + p_2)(1 - (p_1 + p_2)) = np_1(1 - p_1) + np_2(1 - p_2) + 2\text{Cov}(X_1, X_2)$$

solve this we can have $\text{Cov}(X_1, X_2) = -np_1p_2$. Similarly, this can be done for arbitrary i and j. □

7.4 Multivariate Normal (MVN)

Definition 7.4.1: Multivariate Normal Distribution

A k-dimensional random vector $\mathbf{X} = (X_1, X_2, \cdots, X_k)$ is **multivariate normal distribution**, denoted as $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its PDF is

$$f(\mathbf{x}) = (2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu} = (E(X_1), E(X_2), \cdots, E(X_k))^T$, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix $(\boldsymbol{\Sigma})_{i,j} = \text{Cov}(X_i, X_j)$, where each X_i is a normal distribution.

Specifically, we have the *Bivariate Normal* (X, Y) with $N(0,1)$ marginal distributions and correlation ρ have PDF

$$f_{X,Y}(x,y) = \frac{1}{2\pi\tau} \exp\left(-\frac{1}{2\tau^2}(x^2 + y^2 - 2\rho xy)\right)$$

where $\tau = \sqrt{1-\rho^2}$.

A property, which is always as the definition of multivaraite normal, is in the theorem below.

Theorem 7.4.2

A k-dimensional random vector $\mathbf{X} = (X_1, X_2, \cdots, X_k)$ is mutivariate normal if every linear combination of the X_j has a Normal Distribution. That is, we require

$$t_1X_1 + \cdots + t_kX_k$$

to have normal distribution for any constants t_i.

The marginal of MVN is normal. Converse is false: *it is possible to have normally distributed r.v.s.* $X_1, \cdots, X_k$ *such that* $(X_1, \cdots, X_k)$ *is not multivariate normal.* Let $X \sim N(0,1)$ and let

$$S = \begin{cases} 1, & \text{with probability } 1/2 \\ -1, & \text{with probability } 1/2 \end{cases}$$

be independent of X. Then $Y = SX$, can be seen, is a standard normal r.v., since

$$\begin{aligned} P(Y \leqslant y) &= P(SX \leqslant y) = P(S = 1, X \leqslant y) + P(S = -1, X \geqslant -y) \\ &= P(S = 1)P(X \leqslant y) + P(S = -1)P(X \geqslant -y) = P(S = 1)P(X \leqslant y) + P(S = -1)P(X \leqslant y) \\ &= \frac{1}{2}P(X \leqslant y) + \frac{1}{2}P(X \leqslant y) = P(X \leqslant y) \end{aligned}$$

However, $(X + Y)$ is not bivariate normal because

$$P(X + Y = 0) = P(S = -1) = \frac{1}{2}$$

which indicates that it cannot be a continuous distribution.

Moreover, we can use this example to show that uncorrelated cannot lead to independence. They are uncorrelated since

$$Cov(X,Y) = E[XY] - E[X]E[Y] = E[SX^2] - 0 \times 0 = E[S]E[X^2] \text{ by independence} = 0 \text{ since } E[S] = 0$$

Now consider $\mathbb{P}(|Y| \geqslant 10, |X| \leqslant 5)$

$$P(|Y| \geqslant 10, |X| \leqslant 5) = P(|S||X| \geqslant 10, |X| \leqslant 5) = P(|X| \geqslant 10, |X| \leqslant 5) = 0$$

However,

$$P(|Y| \geqslant 10)P(|X| \leqslant 5) > 0 \text{ since they are all normally distributed}$$

Thus X and Y are not independent.

Therefore, generally the converse of Theorem 7.2.2 is false. However, we will next show that the converse is true for multivariate normal random variables. Let's first introduce joint MGF.

Definition 7.4.3: Joing MGF

The **joint moment generating function** of a random vector $\mathbf{X} = (X_1, \cdots, X_k)$ is the function M defined by

$$M(\mathbf{t}) = E(e^{\mathbf{t}^T\mathbf{X}}) = E(e^{t_1X_1+t_2X_2+\cdots+t_kX_k})$$

for $\mathbf{t} = (t_1, t_2, \cdots, t_k) \in \mathbb{R}^k$.

Proposition 7.4.4: Multivariate Normal Joint MGF

The joint MGF for multivariate normal is

$$M(\mathbf{t}) = \exp\left(\boldsymbol{\mu}^T\mathbf{t} + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}\right)$$

Proof. Recall that for any normal r.v. W,

$$E(e^W) = e^{E(W)+\frac{1}{2}\mathrm{Var}(W)}$$

For a multivariate normal random vector, the exponent $t_1X_1 + \cdots t_kX_k$ is normal by definition. Therefore, the joint MGF is

$$E(e^{t_1X_1+\cdots t_kX_k}) = \exp\left(t_1E(X_1) + t_2E(X_2) + \cdots + t_kE(X_k) + \frac{1}{2}\mathrm{Var}(t_1X_1 + \cdots + t_kX_k)\right)$$

□

Now we are ready to show this important conclusion.

Theorem 7.4.5: Uncorrelation implies independence in multivariate normal

Within an MVN random vector, **uncorrelated implies independent**. That is, if $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, and every component of $\mathbf{X_1}$ is uncorrelated with every component of $\mathbf{X_2}$, then they $\mathbf{X_1}$ and $\mathbf{X_2}$ are independent.

Proof. We will prove for 2-dim case. The proof for higher dimension is analogous. Let (X, Y) be bivariate normal with $E(X) = \mu_1$, $E(Y) = \mu_2$, $\mathrm{Var}(X) = \sigma_1^2$, $\mathrm{Var}(Y = \sigma_2^2$ and $\mathrm{Corr}(X, Y) = \rho$. The joint MGF is then

$$M_{X,Y}(s,t) = \exp\left(s\mu_1 + t\mu_2 + \frac{1}{2}\mathrm{Var}(sX + tY)\right) = \exp\left(s\mu_1 + t\mu_2 + \frac{1}{2}(s^2\sigma_1^2 + t^2\sigma_2^2 + 2st\sigma_1\sigma_2\rho)\right)$$

If $\rho = 0$, the joint MGF reduces to

$$M_{X,Y}(s,t) = \exp\left(s\mu_1 + t\mu_2 + \frac{1}{2}(s^2\sigma_1^2 + t^2\sigma_2^2)\right)$$

But this is also the joint MGF of (Z, W) where Z is independent of W with $Z \sim N(\mu_1, \sigma_1^2)$ and $W \sim N(\mu_2, \sigma_2^2)$. Since MGF determines the joint distribution, it must be that (X, Y) is independent. $\square$

Chapter 8

Transformation

8.1 Change of Variables

Theorem 8.1.1: Change of variable in 1-dim

Let X be continuous r.v. with PDF f_X, and let $Y = g(X)$, where g is differentiable and strictly monotone. Then, the PDF of Y is

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

where $x = g^{-1}(y)$. The support of Y is all $g(x)$ with x in the support of X.

Proof.

- Let g be strictly increasing. The CDF of Y is then

$$F_Y(y) = P(Y \leqslant y) = P(g(X) \leqslant y) = P(X \leqslant g^{-1}(y)) = F_X(g^{-1}(y)) = F_X(x)$$

  By Chain Rule, the PDF of Y is

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

- The proof for g strictly decreasing is analogous, with PDF ends up in $-f_X(x)\frac{dx}{dy}$.

Using absolute value, as in the statement of the theorem, covers both cases. □

Then, a corollary naturally follows for linear transformations of variable.

Corollary 8.1.2: Linear Transformation of Variable

Let X be continuous r.v. with PDF f_X, and let $Y = a + bX$, with $b \neq 0$. Let $y = a + bx$. Then, the PDF of Y is

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = f_X\left(\frac{y-a}{b}\right)\frac{1}{|b|}$$

Example 8.1.3: Log-Normal Distribution

Let $X \sim N(0,1)$, $Y = e^X$. Then Y follows the **log-normal distribution**. Find the PDF of Y.

Proof. $g(x) = e^x$ is strictly increasing, and $x = \log y$. We then can apply Theorem 8.1.1 to have that

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \varphi(x)\frac{1}{y} = \varphi(\log y)\frac{1}{y}, y > 0$$

Note that the support of log-normal is $y > 0$. □

This can be generalized into more than one dimension.

Theorem 8.1.4: Change of Variables

Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ be a continous random vector with joint PDF $f_{\mathbf{X}}$. Let $g : A \to B$ be an invertible function where A and B are open subsets of $\mathbb{R}^n$, A contains the support of $\mathbf{X}$, and B is the range of g. Let $\mathbf{Y} = g(\mathbf{X})$. Since g is invertible, we have $\mathbf{X} = g^{-1}(\mathbf{Y})$. Suppose all the partial derivatives $\frac{\partial x_i}{\partial y_j}$ exist and are continuous, then we can form the **Jacobian Matrix**

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$$

Also assume that the determinant of this Jacobian matrix is never 0, then the joint PDF of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))\left|\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)\right|$$

The proof of this theorem is just a generalization in multivariable calculus.

Example 8.1.5: Box-Muller Method for Generating Normal r.v.s

Let $U \sim \text{Unif}(0, 2\pi)$, and let $T \sim \text{Exp}(1)$ be independent of U. Define

$$X = \sqrt{2T}\cos U \quad \text{and} \quad Y = \sqrt{2T}\sin U$$

Then, X,Y are independent standard normal r.v.s.

Proof. The PDF of U and V are

$$f_U(u) = \frac{1}{2\pi}, u \in (0, 2\pi), \quad f_T(t) = e^{-t}, t > 0$$

respectively. Since they are independent, the joint PDF is

$$f_{U,T}(u,t) = \frac{1}{2\pi}e^{-t}, \quad u \in (0, 2\pi), t > 0$$

Viewing (X, Y) as a point in the plane,

$$X^2 + Y^2 = 2T(\cos^2 U + \sin^2 U) = 2T$$

is the squared distance from the origin and U is the angle; That is, $(\sqrt{2T}, U)$ expresses (X, Y) in polar coordinates. Since we can recover from polar to Cartesian coordinate, the transformation is invertible. The Jacobian matrix

$$\frac{\partial(x, y)}{\partial(u, t)} = \begin{pmatrix} -\sqrt{2t}\sin u & \frac{1}{\sqrt{2t}}\cos u \\ \sqrt{2t}\cos u & \frac{1}{\sqrt{2t}}\sin u \end{pmatrix}$$

the determinant is 1. Therefore, we have

$$f_{X,Y}(x, y) = f_{U,T}(u, t) \times 1 = \frac{1}{2\pi}e^{-t} = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)} = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$$

By Theorem 7.1.1, X and Y are independent normal distributions. □

8.2 Convolution

Theorem 8.2.1: Convolution of Independent r.v.s

Let X, Y be independent r.v.s and $T = X + Y$.

1. If X and Y are discrete, the PMF of T is

$$P(T = t) = \sum_x P(Y = t - x)P(X = x) = \sum_y P(X = t - y)P(Y = y)$$

2. If X and Y are continuous, the PDF of T is

$$f_T(t) = \int_{-\infty}^{\infty} f_Y(t - x)f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f_X(t - y)f_Y(y)\,\mathrm{d}y$$

The proof can be simply derived from LOTP.

To see an example, let $X, Y \sim \text{Unif}(0, 1)$. Let $T = X + Y$. The PDF of X and Y is

$$g(x) = \begin{cases} 1, & x \in (0, 1) \\ 0, & \text{otherwise} \end{cases}$$

The convolution gives

$$f_T(t) = \int_{-\infty}^{\infty} g(t - x)g(x)\,\mathrm{d}x$$

The integrand is 1 if and only if $0 < t - x < 1$ and $0 < x < 1$. Then, for $0 < t \leqslant 1$ we have $0 < x < t$, and for $1 < t < 2$

we have $t-1<x<1$. Therefore, the PDF of T is a piecewise linear function

$$f_T(t) = \begin{cases} \displaystyle\int_0^t \mathrm{d}x = t, & t \in (0,1] \\ \displaystyle\int_{t-1}^1 , \mathrm{d}x = 2-t & 1<t<2 \end{cases}$$

8.3 Gamma Distribution

Gamma distribution is a generalization of the Exponential distribution. While an Exponential r.v. represents the waiting time for the first success under conditions of memorylessness, we shall see that a Gamma r.v. represents the total waiting time for multiple successes.

Definition 8.3.1: Gamma Function

The **gamma function** Γ is defined by

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\,\mathrm{d}x, \quad a>0$$

It has two important properties:

- $\Gamma(a+1) = a\Gamma(a)$ for all $a>0$.
- $\Gamma(n) = (n-1)!$ for positive integer n.

Definition 8.3.2: Gamma Distribution

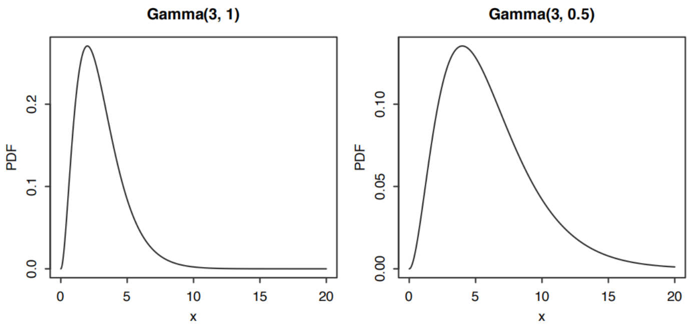
An r.v. X is said to have the **Gamma distribution** with parameters a and λ, with $a>0$ and $\lambda>0$, if its PDF is

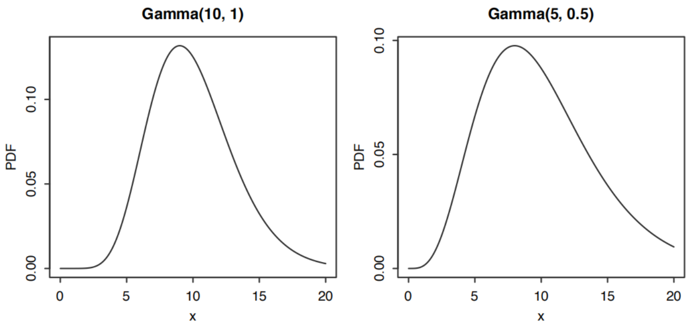
$$f(x) = \frac{\lambda^a}{\Gamma(a)}x^{a-1}e^{-\lambda x}, \quad x>0$$

We write $X \sim \text{Gamma}(a,\lambda)$. a is called the **shape parameter**, and λ is called the **rate parameter**. Expectation: $E(X) = a/\lambda$. Variance: $\text{Var}(X) = a/\lambda^2$.

Special Case: For $a=1$, $\text{Gamma}(1,\lambda) = \text{Exp}(\lambda)$.

- a increase, PDF looks more symmetric.
- λ increase, compress the PDF toward smaller values.





Expectation:

$$E(X) = \int_0^\infty x \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}\,\mathrm{d}x = \frac{\lambda^a}{\Gamma(a)} \frac{\Gamma(a+1)}{\lambda^{a+1}} \int_0^\infty \frac{\lambda^{a+1}}{\Gamma(a+1)} x^a e^{-\lambda x}\,\mathrm{d}x$$
$$= \frac{a!}{(a-1)!} \frac{1}{\lambda} = \frac{a}{\lambda}$$

where the integral equals 1 since it is the integration of Gamma$(a+1, \lambda)$ distribution.

Variance:

$$E(X^2) = \int_0^\infty x^2 \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}\,\mathrm{d}x = \frac{\lambda^a}{\Gamma(a)} \frac{\Gamma(a+2)}{\lambda^{a+2}} \int_0^\infty \frac{\lambda^{a+2}}{\Gamma(a+2)} x^{a+1} e^{-\lambda x}\,\mathrm{d}x = \frac{a(a+1)}{\lambda^2}$$

Therefore,

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{a(a+1)}{\lambda^2} - \frac{a^2}{\lambda^2} = \frac{a}{\lambda^2}$$

MGF:

$$E(e^{tX}) = \int_0^\infty e^{tx} \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}\,\mathrm{d}x = \frac{\lambda^a}{\Gamma(a)} \frac{\Gamma(a)}{(\lambda-t)^a} \int_0^\infty \frac{(\lambda-t)^a}{\Gamma(a)} x^{a-1} e^{-(\lambda-t)x}\,\mathrm{d}x = \left(\frac{\lambda}{\lambda-t}\right)^a$$

for $t < \lambda$.

Recall The MGF of Exponential distribution, and recall Theorem 6.4.3, we have the following result.

Theorem 8.3.3: Gamma-Exponential Relationship

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Exp(λ). Then,

$$X_1 + X_2 + \cdots + X_n \sim \text{Gamma}(n, \lambda)$$

The other two properties of gamma are similar to derive:

- If $X_1 \sim$ Gamma(a, λ), and $X_2 \sim$ Gamma(b, λ), we have $X_1 + X_2 \sim$ Gamma$(a+b, \lambda)$.
- If $X \sim$ Gamma(a, λ), then $cX \sim$ Gamma$(a, \lambda/c)$.

In *Poisson Process*, Gamma distribution is the time to wait n success. It is the continuous analog of *negative binomial distribution* since exponential is the analog of geometric distribution.

8.4 Beta Distribution

The Beta distribution is a continuous distribution on the interval $(0, 1)$. It is a generalization of the $\mathrm{Unif}(0, 1)$ distribution, allowing the PDF to be non-constant on $(0, 1)$.

Definition 8.4.1: Beta Distribution

An r.v. X has **Beta Distribution** with parameter a and b, where $a > 0$ and $b > 0$, if its PDF is

$$f(x) = \frac{1}{\mathcal{B}(a, b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$$

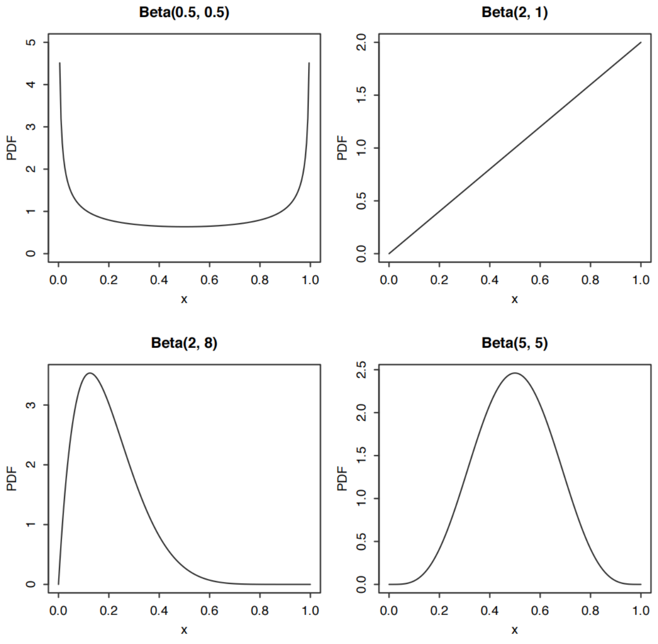
where $\mathcal{B}(a, b)$ is **beta function**,

$$\mathcal{B}(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

which makes the PDF integrate to 1. We write as $X \sim \mathrm{Beta}(a, b)$.
Expectation: $E(x) = a/(a+b)$. Variance: $\mathrm{Var}(X) = ab/(a+b)^2(a+b+1)$.

Special Case: For $a = b = 1$, $\mathrm{Beta}(1, 1) = \mathrm{Unif}(0, 1)$.

- If $a < 1$ and $b < 1$, the PDF is U-shaped and opens upwards. If $a > 1$ and $b > 1$, the PDF opens down.
- If $a = b$, the PDF is symmetric about $1/2$. If $a > b$, the PDF favors values bigger than $1/2$. If $a < b$, the PDF favors values smaller than $1/2$.



To figure out how to calculate mean and variance, and to prove the relation between gamma and beta function, we need to show the relation between beta and gamma distribution.

Theorem 8.4.2: Beta-Gamma Relation

Let $X \sim \text{Gamma}(a, \lambda)$ and $Y \sim \text{Gamma}(b, \lambda)$ be independent. Then

$$T = X + Y \sim \text{Gamma}(a + b, \lambda)$$

and

$$W = \frac{X}{X+Y} \sim \text{Beta}(a, b)$$

Proof. We will do change of variables to solve this. Let $t = x + y$ and $w = x/(x + y)$. Then, $x = tw$ and $y = t(1 - w)$, and

$$\frac{\partial(x,y)}{\partial(t,w)} = \begin{pmatrix} w & t \\ 1-w & -t \end{pmatrix}$$

The determinant is then $-t$. Therefore,

$$\begin{aligned} f_{T,W}(t,w) = f_{X,Y}(x,y)|-t| = f_X(x)f_Y(y)t &= \frac{\lambda^a}{\Gamma(a)}x^{a-1}e^{-\lambda x}\frac{\lambda^b}{\Gamma(b)}y^{b-1}e^{-\lambda y}t \\ &= \frac{\lambda^a}{\Gamma(a)}(tw)^{a-1}e^{-\lambda tw}\frac{\lambda^b}{\Gamma(b)}(t(1-w))^{b-1}e^{-\lambda t(1-w)}t \\ &= \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}w^{a-1}(1-w)^{b-1}\right)\left(\frac{\lambda^{a+b}}{\Gamma(a+b)}t^{a+b-1}e^{-\lambda t}\right) = f(w)g(t) \end{aligned}$$

By Theorem 7.1.1, W and T are independent, and since we can see from the formula that $T \sim \text{Gamma}(a + b, \lambda)$ is a valid PDF, we have the formula for W is also a valid PDF. Then, it must be beta distribution and the Beta function is

$$\frac{1}{\mathcal{B}(a,b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

□

Expectation: Since T and W are independent, they are uncorrelated, thus $E(TW) = E(T)E(W)$. Then,

$$E\left((X+Y)\frac{X}{X+Y}\right) = E(X) = E(X+Y)E\left(\frac{X}{X+Y}\right)$$

Therefore,

$$E\left(\frac{X}{X+Y}\right) = \frac{E(X)}{E(X+Y)} = \frac{a/\lambda}{(a+b)/\lambda} = \frac{a}{a+b}$$

Variance:

$$\begin{aligned} E(W^2) &= \int_0^1 w^2\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}w^{a-1}(1-w)^{b-1}\,\mathrm{d}w = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}\int_0^1\frac{\Gamma(a+b+2)}{\Gamma(a+2)\Gamma(b)}w^{a+1}(1-w)^{b-1}\,\mathrm{d}w \\ &= \frac{\Gamma(a+b)}{\Gamma(a)}\frac{a(a+1)\Gamma(a)}{(a+b)(a+b+1)\Gamma(a+b)} = \frac{a(a+1)}{(a+b)(a+b+1)} \end{aligned}$$

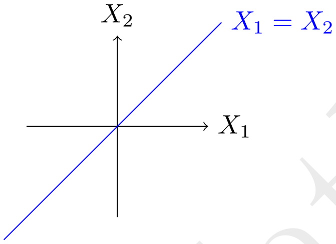
Therefore, the variance

$$\mathrm{Var}(W) = E(W^2) - E(W)^2 = \frac{a(a+1)(a+b)}{(a+b)^2(a+b+1)} - \frac{a^2(a+b+1)}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)}$$

8.5* Order Statistics

Note: This section's style is a little bit different since I directly copied this from my undergraduate notes.

Given a random sample $X_1, ...X_n$ of size n, and suppose that they are continuous rvs. As a result, we have $\mathbb{P}(X_j = X_k) = 0$ for $j \neq k$. This can be explained by the figure shown below. Since these random variable is continuous, the domain of these continuous rvs must cover some region of the Cartesian space. Therefore, the probability of $X_1 = X_2$, which is a straight line, can be seen as 0.



With this finding, we can arrange these random variables in a strictly increasing manner. This is called **Ordered Statistics**. The PDF and CDF are different from the non-ordered rvs since different permutation can lead to different result. Below is the rigorous definition.

Definition 8.5.1 (Order Statistics)**.** Let $X_1, ...X_n$ be a random sample of size n (i.i.d. rvs). For a fixed sample point $\omega \in S$, we can order $X_1(\omega), ..., X_n(\omega)$ in a non-decreasing manner and labelled as

$$Y_1(\omega), ..., Y_n(\omega) \text{ where } Y_1 < Y_2 < ... < Y_n$$

In this way, we get n new random variables. We call Y_1 the **smallest order statistics** of the random sample $X_1, ...X_n$ and Y_n the **largest order statistics** of the random sample. Also, Y_k the $\mathbf{k-th}$ **order statistics** of the random sample.

The first thing we want to know is the PDFs of these ordered statistics. To achieve this, we can use the so-called **CDF technique**, which allow us to first generate the CDF of some random variable and obtain the PDF by differentiation.

Theorem 8.5.1 (joint PDF of order statistics)**.** Let $Y_1, ..., Y_n$ be the order statistics based on a random sample of size n from a population distribution with PDF $f(x)$. Then, the joint PDF of $Y_1, ..., Y_n$ is given by

$$g(y_1, y_2, ...y_n) = n! f(y_1)...f(y_n)$$

for $y_1 < y_2 < ... < y_n$ and $g(y_1, y_2, ...y_n) = 0$ elsewhere.

Proof. Let $X_1, ...X_n$ be a random sample of size n with population PDF f. Then, one can view the order statistics as the ordered version of $X_1, ...X_n$ and write

$$Y_1 = X_{\sigma(1)}, \quad Y_2 = X_{\sigma(2)}, \quad ..., \quad Y_n = X_{\sigma(n)}$$

where $\sigma : \{1, ..., n\} \to \{1, ..., n\}$ is some bijection (i.e., permutation) such that

$$X_{\sigma(1)} < X_{\sigma(2)} < ... < X_{\sigma(n)}$$

Given such a random sample, we can always find a unique permutation σ satisfying the above relation. As a result, we can write

$$\begin{aligned}\mathbb{P}(Y_1 \leqslant t_1, ..., Y_n \leqslant t_n) &= \mathbb{P}\left(\sum_{\sigma} \left\{X_{\sigma(1)} \leqslant t_1, ..., X_{\sigma(n)} \leqslant t_n\right\}\right) = n!\mathbb{P}(X_1 \leqslant t_1, ..., X_n \leqslant t_n) \\ &= \int_{-\infty}^{t_1} ... \int_{-\infty}^{t_n} [n!f(y_1)f(y_2)...f(y_n)] \, \mathrm{d}y_n...\mathrm{d}y_1\end{aligned}$$

The function inside the integration is clearly the joint PDF of order statistics. □

With this joint PDF we are able to find the PDF of single order statistics.

Theorem 8.5.2 (PDF of kth order statistics)**.** Let Y_k be the kth order statistics based on a random sample of size n from a population distribution with PDF $f(x)$ and CDF $F(x)$. Then, Y_k is a continuous random variable with PDF

$$g_k(y) = \frac{n!}{(k-1)!(n-k)!} F(y)^{k-1}[1 - F(y)]^{n-k} f(y)$$

Proof. We can obtain the marginal from the joint PDF got from Theorem 8.5.1 by integration over other variables

$$\begin{aligned}g_k(y) &= \int_{\mathbb{R}^{n-1}} f_{Y_1,...,Y_n}(y_1, ..., y_{k-1}, y, y_{k+1}, ..., y_n) \, \mathrm{d}y_1...\mathrm{d}y_{k-1}\mathrm{d}y_{k+1}...\mathrm{d}y_n \\ &\overset{\text{Thm2.4}}{=\!=\!=} n!f(y) \left(\int_{y_1<y_2<...<y_{k-1}<y} f(y_1)...f(y_{k-1})\mathrm{d}y_1...\mathrm{d}y_{k-1}\right) \times \left(\int_{y<y_{k+1}<...<y_n} f(y_{k+1})...f(y_n)\mathrm{d}y_{k+1}...\mathrm{d}y_n\right)\end{aligned}$$

We can see that this function is symmetric. By symmetric we mean that for any permutation σ on $\{1, ..., n\}$, we have

$$f(x_{\sigma(1)}, ..., x_{\sigma(n)}) = f(x_1, ..., x_n)$$

For example, $g(x_1, x_2) = x_1 + x_2$ is symmetric and $h(x_1, x_2) = x_1^2 + x_2$ is not symmetric.
In case $f(x_1, ...x_n)$ is symmetric, we actually have

$$\int_{x_1<x_2<...<x_n} f(x_1, ...x_n) \, \mathrm{d}x_1...\mathrm{d}x_n = \frac{1}{n!} \int_{\mathbb{R}^n} f(x_1, ...x_n) \, \mathrm{d}x_1...\mathrm{d}x_n$$

If this is not intuitive, consider the case that we have two variables and $g(x, y) = xy$. Then the integration over $[0, 1]^2$

is a square and the integration over $x < y \in (0,1)$ is a triangle. Therefore, we can have

$$\int_{x<y} xy\,\mathrm{d}x\mathrm{d}y = \frac{1}{2}\int_0^1\int_0^1 xy\,\mathrm{d}x\mathrm{d}y = \frac{1}{2!}\int_0^1\int_0^1 xy\,\mathrm{d}x\mathrm{d}y$$

Coming back to the proof of $g_k(y)$. If we substitute this property of symmetric function into the original equation, we have

$$\begin{aligned} g_k(y) &= n!f(y)\left(\int_{y_1<y_2<...<y_{k-1}<y} f(y_1)...f(y_{k-1})\mathrm{d}y_1...\mathrm{d}y_{k-1}\right) \times \left(\int_{y<y_{k+1}<...<y_n} f(y_{k+1})...f(y_n)\mathrm{d}y_{k+1}...\mathrm{d}y_n\right) \\ &= f(y)\frac{n!}{(k-1)!(n-k)!}\underbrace{\int_{-\infty}^{y}...\int_{-\infty}^{y}}_{k-1 \text{ terms}} f(y_1)...f(y_{k-1})\mathrm{d}y_1...\mathrm{d}y_{k-1} \times \underbrace{\int_{y}^{\infty}...\int_{y}^{\infty}}_{n-k \text{ terms}} f(y_{k+1})...f(y_n)\mathrm{d}y_{k+1}...\mathrm{d}y_n \\ &= f(y)\frac{n!}{(k-1)!(n-k)!}F(y)^{k-1}[1-F(y)]^{n-k} \end{aligned}$$

□

Example 8.5.1. Use the **CDF technique** to find out the PDF of Y_1 to verify Thm 8.5.2.

Sol. Consider the tail probability

$$\mathbb{P}(Y_1 \geqslant t) = \mathbb{P}(X_1 \geqslant t, ..., X_n \geqslant t)$$

since Y_1 is the minimum. Then, using independence, we have

$$\mathbb{P}(Y_1 \geqslant t) = \mathbb{P}(X_1 \geqslant t)...\mathbb{P}(X_n \geqslant t) = (1-F(t))^n \text{ where } F \text{ is the CDF of } X_1$$

The PDF of Y_1 is obtained then by differentiation

$$f_{Y_1}(t) = -\frac{d}{dt}\mathbb{P}(Y_1 \geqslant t) = -\frac{d}{dt}(1-F(t))^n = n(1-F(t))^{n-1}f(t)$$

which is exactly the case when $k = 1$ in Thm 8.5.2. □

Now we mention two ways of sampling for the use of order statistics.

Definition 8.5.2 (Type II censored sampling)**.** Terminate the experiment after only first r ordered observations have occurred among n random samples.

Theorem 8.5.3 (joint PDF of Type II censored sampling)**.** Let $Y_1, ...Y_n$ be the order statistics based on a random sample of size n from a population distribution with PDF $f(x)$. Fix $r \in \{1, 2, ..., n\}$. Then, the joint PDF of $Y_1, ..., Y_r$ is given by

$$f_{Y_1,...,Y_r}(y_1, ..., y_r) = \frac{n!}{(n-r)!}(1 - F(y_r))^{n-r} \prod_{i=1}^{r} f(y_i)$$

for $y_1 < y_2 < ... < y_r$ and $g(y_1, ..., y_r) = 0$ elsewhere.

Proof. If follows from Thm 8.5.1 that for $y_1, ..., y_r$,

$$\begin{aligned} f_{Y_1,...,Y_r}(y_1, ..., y_r) &= \int_{\mathbb{R}^{n-r}} f_{Y_1,...,Y_n}(y_1, ..., y_n)\, \mathrm{d}y_{r+1}...\mathrm{d}y_n = n! f(y_1)...f(y_r) \int_{y_r < y_{r+1} < ... < y_n} f(y_{r+1})...f(y_n)\, \mathrm{d}y_{r+1}...\mathrm{d}y_n \\ &\overset{symmetric}{=\!=\!=\!=} \frac{n! f(y_1)...f(y_r)}{(n-r)!} \underbrace{\int_{y_r}^{\infty} ... \int_{y_r}^{\infty}}_{n-r \text{ terms}} f(y_{r+1})...f(y_n)\, \mathrm{d}y_{r+1}...\mathrm{d}y_n \\ &= \frac{n! f(y_1)...f(y_r)}{(n-r)!}[1 - F(y_r)]^{n-r} \end{aligned}$$

□

Definition 8.5.3 (Type I censored sampling)**.** Terminate the experiment after time t_0. Given a random sample $X_1, ..., X_n$ of size n from some population distribution with CDF F and PDF f. The number of observations, denoted by R, is $BIN(n, p)$ distributed with

$$p = F(t_0)$$

Theorem 8.5.4 (Truncated PDF of type I censored samples)**.** The truncated density of type I censored samples $X_1, ..., X_n$ is

$$f_{t_0}(x) = \frac{1}{F(t_0)} f(x)$$

Proof. This can be obtained by the **CDF technique**. The CDF of X_1 is given by

$$\mathbb{P}(X_1 \leqslant x | X_1 \leqslant t_0) = \frac{\mathbb{P}(X_1 \leqslant x)}{\mathbb{P} \leqslant t_0} = \frac{F(x)}{p} = \frac{1}{F(t_0)} F(x)$$

Taking derivative w.r.t. x, we can yield the formula. □

Theorem 8.5.5 (Type I censored samples joint conditional PDF)**.** Let $R \sim BIN(n, F_{t_0})$ be the number of observations before time t_0. Then, when $R = r$, $Y_1, ..., Y_r$ have joint conditional PDF:

$$g(y_1, ..., y_r | R = r) = \frac{r!}{F(t_0)^r} \prod_{i=1}^{r} f(y_i)$$

Proof. This follows from Thm 8.5.1 with PDF using the truncated one, which is $f_{t_0}(x) = \frac{1}{F(t_0)} f(x)$. □

Theorem 8.5.6 (Type I censored samples joint PDF)**.** The joint distribution of $(Y_1, ..., Y_R)$ for type I censored sampling can be described by

$$f_{Y_1,...,Y_R}(y_1, ..., y_r) = \frac{n!}{(n-r)!}[1 - F(t_0)]^{n-r} \prod_{i=1}^{r} f(y_i)$$

Proof.

$$\mathbb{P}(Y_1 \leqslant t_1, ..., Y_R \leqslant t_r, R = r) = \mathbb{P}(R = r)\mathbb{P}(Y_1 \leqslant t_1, ..., Y_R \leqslant t_r | R = r)$$

$$= \int_{-\infty}^{t_1} ... \int_{-\infty}^{t_r} \underbrace{\binom{n}{r} F(t_0)^r (1 - F(t_0))^{n-r}}_{\text{binomial distribution}} \times \underbrace{\frac{r!}{F(t_0)^r} \prod_{i=1}^{r} f(y_i)}_{\text{joint conditional PDF}} \mathbb{1}_{y_1<y_2<...<y_r} \, \mathrm{d}y_1 ... \mathrm{d}y_r$$

$$= \int_{-\infty}^{t_1} ... \int_{-\infty}^{t_r} \frac{n!}{(n-r)!} (1 - F(t_0))^{n-r} \prod_{i=1}^{r} f(y_i) \mathbb{1}_{y_1<y_2<...<y_r} \, \mathrm{d}y_1 ... \mathrm{d}y_r$$

where the function inside the integral is exactly the joint PDF. □

Chapter 9

Inequalities and Limit Theorems

9.1 Inequalities

Cauchy-Schwarz Inequality in Probability

$$|E(XY)| \leqslant \sqrt{E(X^2)E(Y^2)}$$

Proof. For any t,

$$0 \leqslant E(Y - tX)^2 = E(Y^2) - 2tE(XY) + t^2E(X^2)$$

Differentiate right hand side w.r.t t and set it equals to 0, we get

$$t = \frac{E(XY)}{E(X^2)}$$

minimizes the right hand side. Plugging in this value t, we have

$$E(Y^2) - 2\frac{E(XY)^2}{E(X^2)} + \frac{E(XY)^2}{E(X^2)} \geqslant 0 \quad \Longrightarrow \quad |E(XY)| \leqslant \sqrt{E(X^2)E(Y^2)}$$

□

Jenson's Inequality in Probability

For convex function g, $E(g(X)) \geqslant g(E(X))$. For concave function, $E(g(X)) \leqslant g(E(X))$

In both cases, the only way that equality holds is if there are constants a and b such that $g(X) = a + bX$ with probability 1.

Proof. If g is convex, then all lines that are tangent to g lie below g. In particular, let $\mu = E(X)$, and consider the tangent line at the point $(\mu, g(\mu))$. If the tangent line is not unique, choose any one. Denote this tangent line by $a + bx$, we have $g(x) \geqslant a + bx$ for all x by convexity, so $g(X) \geqslant a + bX$. Taking expectation,

$$E(g(X)) \geqslant E(a + bX) = a + b\mu = g(\mu) = g(E(X))$$

Similarly, if g is concave, then $h = -g$ is convex, so we can apply previous steps to h.

Let's assume the equality holds in convex case. Let $Y = g(X) - a - bX$. Then Y is nonnegative .r.v. with $E(Y) = 0$. So $P(Y = 0) = 1$. For the concave case, we can use $Y = a + bX - g(X)$. □

Markov's Inequality

$$P(|X| \geqslant a) \leqslant \frac{E(|X|)}{a}, \quad a > 0$$

Proof. Let $Y = |X|/a$. We need to show that $P(Y \geqslant 1) \leqslant E(Y)$. Note that

$$I(Y \geqslant 1) \leqslant Y$$

since if $I(Y \geqslant 1) = 0$ then the inequality reduces to $Y \geqslant 0$, and if $I(Y \geqslant 1) = 1$ then $Y \geqslant 1$. Taking expectation of both sides, we have Markov's Inequality. □

Chebyshev's Inequality

X has mean μ and variance σ^2, $\quad P(|X - \mu| \geqslant a) \leqslant \frac{\sigma^2}{a^2}$ or equivalently $P(|X - \mu| \geqslant c\sigma) \leqslant \frac{1}{c^2}, \quad a, c > 0$

Proof. By Markov's Inequality,

$$P(|X - \mu| \geqslant a) = P((X - \mu)^2 \geqslant a^2) \leqslant \frac{E[(X - \mu)^2]}{a^2} = \frac{\sigma^2}{a^2}$$

Substituting $c\sigma$ for a, we have another inequality. □

9.2 Convergence Almost Surely, in Probability, and in Distribution

There are different modes in convergence due to some analysis rigor.

Definition 9.2.1: Convergence Almost Surly

A sequence of random variables X_n **converges almost surely** to X, written as $X_n \xrightarrow{a.s.} X$, if

$$P(X_n \to X) = 1 \quad (i.e., P(s \in S : X_n(s) \to X(s)) = 1)$$

This is the strongest condition of convergence.

Example 9.2.1. Let the sample space $S = [0,1]$, and $X_n(s) = s^n$.

$$\forall\, s \in [0,1), X_n(s) \to 0 \text{ as } n \to \infty$$

$$s = 1, X_n(s) \nrightarrow 0, \text{ but } P(\{1\}) = 0$$

Thus $X_n \xrightarrow{a.s.} 0$.

Definition 9.2.2: Convergence in Probability

A sequence of random variables X_n **converges almost surely** to X, written as $X_n \xrightarrow{P} X$, if

$$\forall\, \epsilon > 0, \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 0 \quad \left(i.e., \lim_{n\to\infty} P(s \in S : |X_n(s) - X(s)| > \epsilon) = 0\right)$$

We state without proof that convergence almost surely is a *stronger* convergence than convergence in probability. The proof would be addressed in *Measure Theoretic Probability Theory.*

Theorem 9.2.3: Almost Surely implies Probability

If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$.

The converse is not true,

Counterexample: Walking Boxes Let $S = [0,1]$ be the sample space, and let

$$X_1 = I_{[0,1]}(s), \quad X_2 = I_{[0,1/2]}(s), \quad X_3 = I_{[1/2,1]}(s), \quad X_4 = I_{[0,1/4]}(s), \quad X_5 = I_{[1/4,1/2]}(s), \cdots$$

Let $\epsilon > 0$, then

$$\lim_{n\to} P(|X_n \to 0| > \epsilon) = \lim_{n\to\infty} \frac{1}{2^{\lfloor \log_2 n \rfloor}} = 0$$

It converges in probability to 0. However, for example,

$$X_n(0.02) \nrightarrow 0$$

since there exists infinitely many n such that $X_n(0.02) = 1$.

We state some properties of convergence in probability. Note that these can be applied also to convergence almost surely, since it is the stronger version.

Theorem 9.2.4: Properties of Convergence in Probability

If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then

- $X_n + Y_n \xrightarrow{P} X + Y$.
- $X_n Y_n \xrightarrow{P} XY$.
- $X_n / Y_n \xrightarrow{P} X/Y$, provided that $Y \neq 0$.
- For any continuous function g, we have $g(X_n) \xrightarrow{P} g(X)$.

Here comes the final one.

Definition 9.2.5: Convergence in Distribution

Let X_n be a sequence of r.v. with CDF $F_n(x)$. Let X be an r.v. with CDF $F(x)$. Then, X_n **converges in distribution** to X, written as $X_n \xrightarrow{d} X$, if

$$\lim_{n \to \infty} F_n(x) = F(x), \quad \forall x \text{ where } F \text{ is continuous}$$

This is an even *weaker* condition.

Theorem 9.2.6: Probability implies Distribution

If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$.

The converse is not true,
Counterexample: Coin Toss Let $S = \{H, T\}$ be the sample space of a fair coin toss. Let

- $X_1, X_3, X_5, \cdots$ be 1 if toss is H, 0 otherwise.
- $X_2, X_4, X_6, \cdots$ be 0 if toss is H, 1 otherwise.

All X_i have the same distribution, the same PDF, but

$$\{X_n = X_{n+1}\} = \emptyset, \quad P\left(|X_n - X_{n+1}| \geqslant \frac{1}{2}\right) = 1 \text{ for all } n$$

If a sequence of random variable converges in distribution to a constant, it also converges in probability.

Theorem 9.2.7: Distribution implies Probability for Constants

If $X_n \xrightarrow{d} c$, where c is a constant, then $X_n \xrightarrow{P} c$.

There is still one property inherited from convergence in probability.

Theorem 9.2.8: Continuous Transformation

For any continuous function g, if $X_n \xrightarrow{d} X$, then we have $g(X_n) \xrightarrow{d} g(X)$.

Note that the other algebraic limit theorems do not work in the context of convergence in distribution. For example, Let $S = \{H, T\}$ still be the sample space of coin toss outcomes. Let

$$X_n(H) = 1, X_n(T) = 0. \quad Y_n(H) = -1, Y_n(T) = 0$$

and

$$X(H) = 1, X(T) = 0. \quad Y(H) = 0, Y(T) = -1$$

Then, both X_n and Y_n converges in distribution, but $X_n + Y_n$ does not converge in distribution to $X + Y$.

Other useful properties of convergence in distribution is listed below.

Theorem 9.2.9: Slutsky's Theorem

Suppose $X_n \xrightarrow{P} a$, and $Y_n \xrightarrow{d} Y$, where a is a constant. Then,

1. $X_n + Y_n \xrightarrow{d} a + Y$.
2. $X_n Y_n \xrightarrow{d} aY$.
3. $Y_n / X_n \xrightarrow{d} Y/a$, provided that $a \neq 0$.

Theorem 9.2.10: MGF convergence implies Distribution Convergence

Let X_n be a sequence of r.v.s with MGF M_{X_n}. Let X be the r.v. with MGF M_X. If

$$\lim_{n \to \infty} M_{X_n}(t) = M_X(t)$$

then $X_n \xrightarrow{d} X$.

9.3 Law of Large Numbers

Theorem 9.3.1: Strong Law of Large Number

The sample mean $\bar{X}_n \xrightarrow{a.s.} \mu$.

Theorem 9.3.2: Weak Law of Large Number

The sample mean $\bar{X}_n \xrightarrow{P} \mu$.

Proof. For fixed $\epsilon > 0$, by Chebyshev's Inequality,

$$P(|\bar{X}_n - \mu| > \epsilon) \leqslant \frac{\sigma^2}{n\epsilon^2}$$

As $n \to \infty$, the right hand side goes to 0. □

9.4 Central Limit Theorem (CLT)

Theorem 9.4.1: Central Limit Theorem

As $n \to \infty$,

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$$

Proof. We will prove CLT assuming that MGF of X_j exists, though the theorem holds much more generally.

Let $M(t) = E(e^{tX_j})$, and without loss of generality let $\mu = 0$ and $\sigma = 1$ (since we end up standardizing $\bar{X}_n$ for the theorem, we might as well standardize X_j in the first place). Then $M(0) = 1$, $M'(0) = \mu = 0$, and $M''(0) = \sigma^2 = 1$.

We wish to show that the MGF of $\sqrt{n}\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/\sqrt{n}$ converges to the MGF of the $N(0,1)$ distribution, which is $e^{t^2/2}$. By properties of MGF,

$$E(e^{t(X_1+\cdots+X_n)/\sqrt{n}}) = E(e^{tX_1/\sqrt{n}})E(e^{tX_2/\sqrt{n}})\cdots E(e^{tX_n/\sqrt{n}}) = \left(M\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Let $n \to \infty$ for the logarithm, we have

$$\begin{aligned}\lim_{n\to\infty} n\log M\left(\frac{t}{\sqrt{n}}\right) &\overset{y=1/\sqrt{n}}{=\!=\!=} \lim_{y\to 0}\frac{\log M(yt)}{y^2} \overset{\text{L'Hôpital Rule}}{=\!=\!=\!=\!=} \lim_{y\to 0}\frac{tM'(yt)}{2yM(yt)}\\ &\overset{yt\to 0,\text{ thus } M(yt)\to 1}{=\!=\!=\!=\!=\!=\!=} \frac{t}{2}\lim_{y\to 0}\frac{M'(yt)}{y} \overset{\text{L'Hôpital Rule}}{=\!=\!=\!=\!=} \frac{t^2}{2}\lim_{y\to 0}M''(yt) = \frac{t^2}{2}\end{aligned}$$

Therefore the exponential approaches to $e^{t^2/2}$. □

The most widely used normal approximation is statistics is *Binomial Normal Approximation*. To account for the discreteness of X, a binomial r.v., we write the probability $P(Y = k)$ as

$$P(k - 1/2 < Y < k + 1/2)$$

and apply normal approximation to the latter. This is called **continuity correlation**. The approximation then, is

$$P(Y = k) \approx \Phi\left(\frac{k + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 1/2 - np}{\sqrt{np(1-p)}}\right)$$

Final Warining: The central limit theorem and Law of Large Number requires that the mean and variance of X_j are finite. However, the *Cauchy distribution* has no mean or variance exists.

Appendices

Appendix A

Proofs of some Theorems

A.1 Chapter 3 Remaining Proofs

Proof. **Proof for Proposition 3.1.2, 3.2.3 and 3.3.4.**
We first consider **??**. The proof for the three properties are easy and tedious (separating discrete and continuous case), we here prove that given any function F, we can construct a random variable whose CDF is F.

This is trivial if we think about the universality of uniform 5.1.2. Given a function F and the standard uniform distribution $U \sim \text{Unif}(0,1)$, we have the variable $F^{-1}(U)$ satisfying the CDF F distribution.

Similarly, the other two can be proved by universality of the uniform. □

A.2 Chapter 4 Remaining Proofs

Proof. **Computational Proof for Theorem 4.5.2**
For $X \sim GEO(p)$, we have

$$\begin{aligned} M_X(t) = E[e^{tX}] &= \sum_{k=1}^{\infty} e^{tk} p(1-p)^{k-1} = pe^t \sum_{k=1}^{\infty} e^{t(k-1)}(1-p)^{k-1} = pe^t \sum_{k=1}^{\infty} \left[e^t(1-p)\right]^{k-1} \\ &\xlongequal{rewrite} pe^t \sum_{k=0}^{\infty} \left[e^t(1-p)\right]^{k} \text{ (we need } e^t(1-p) < 1 \text{ to compute this sum)} \\ &= pe^t . \frac{1}{1-e^t(1-p)} = \frac{pe^t}{1-e^t(1-p)} \end{aligned}$$

For $Y \sim NegBIN(r,p)$, we have

$$\begin{aligned} M_Y(t) &= \sum_{k=r}^{\infty} e^{tk} \binom{k-1}{r-1} (1-p)^{k-r} p^r = \left(pe^t\right)^r \sum_{k=r}^{\infty} e^{t(k-r)} \frac{(k-1)!}{(r-1)!(k-r)!} (1-p)^{k-r} \\ &\xlongequal{rewrite\, n=k-r} \left(pe^t\right)^r \sum_{n=0}^{\infty} \left[e^t(1-p)\right]^n \frac{(n+r-1)!}{n!(r-1)!} \end{aligned}$$

Now we put $x = e^t(1-p)$ and assume that $0 < x < 1$. We need two facts to help us to solve this sum.

FACT A:

$$\sum_{n=0}^{\infty} x^n \frac{(n+r-1)}{n!} = \sum_{n=0}^{\infty} \frac{d}{dx^{r-1}} x^{n+r-1} \overset{\text{We can do so for } |x|<1}{=\!=\!=\!=} \frac{d}{dx^{r-1}} \left(\sum_{n=0}^{\infty} x^{n+r-1} \right) = \frac{d}{dx^{r-1}} \frac{x^{r-1}}{1-x}$$

FACT B: For any $r \geqslant 1$,

$$\frac{1}{(r-1)!} \frac{d}{dx^{r-1}} \frac{x^{r-1}}{1-x} = \frac{1}{(1-x)^r}$$

This FACT B can be proved by induction:

- $r = 1$, RHS $= 1 \times \frac{1}{1-x}$ = LHS, true.

- Assume it holds for $r = 1, 2, ..., k$, that is:

$$\frac{1}{(k-1)!} \frac{d}{dx^{k-1}} \frac{x^{k-1}}{1-x} = \frac{1}{(1-x)^k}$$

- Now we need to prove that it hold for $r = k+1$. Note first that

$$\begin{aligned}\frac{1}{(1-x)^k} &= \frac{1}{(k-1)!} \frac{d}{dx^{k-1}} \frac{x^{k-1}}{1-x} = \frac{1}{(k-1)!} \frac{d}{dx^{k-1}} \left(\frac{x^{k-1}(1-x+x)}{1-x} \right) = \frac{1}{(k-1)!} \frac{d}{dx^{k-1}} \left(x^{k-1} + \frac{x^k}{1-x} \right) \\ &= \underbrace{\frac{1}{(k-1)!} \frac{d}{dx^{k-1}} \left(x^{k-1} \right)}_{=1} + \frac{1}{(k-1)!} \frac{d}{dx^{k-1}} \left(\frac{x^k}{1-x} \right)\end{aligned}$$

Thus we can get from the $r = k$ condition that:

$$\frac{1}{(k-1)!} \frac{d}{dx^{k-1}} \left(\frac{x^k}{1-x} \right) = \frac{1}{(1-x)^k} - 1$$

Therefore, for $r = k+1$, we have

$$\begin{aligned}\frac{1}{k!} \frac{d}{dx^k} \frac{x^k}{1-x} &= \frac{1}{k} \frac{d}{dx} \left(\frac{1}{(k-1)!} \frac{d}{dx^{k-1}} \left(\frac{x^k}{1-x} \right) \right) = \frac{1}{k} \frac{d}{dx} \left(\frac{1}{(1-x)^k} - 1 \right) \\ &= \frac{1}{k}(-k) \frac{-1}{(1-x)^{k+1}} = \frac{1}{(1-x)^{k+1}}\end{aligned}$$

Hence the original equation is verified by the induction process.

Now we return to the calculation of the MGF of negative binomial distribution. With FACT A and FACT B, we have that with $x = e^t(1-p)$,

$$\begin{aligned}M_Y(t) &= \left(pe^t\right)^r \frac{1}{(r-1)!} \sum_{n=0}^{\infty} x^n \frac{(n+r-1)!}{n!} \overset{FACTA}{=\!=\!=} \left(pe^t\right)^r \frac{1}{(r-1)!} \frac{d}{dx^{r-1}} \frac{x^{r-1}}{1-x} \\ &\overset{FACTB}{=\!=\!=} \frac{(pe^t)^r}{(1-x)^r} = \left(\frac{pe^t}{1-e^t(1-p)} \right)^r\end{aligned}$$

Thus, we can read from the MGF that the statement holds. □

A.3 Chapter 5 Remaining Proofs

Proof. **Generalized Inverse Proof for Theorem 5.1.3**

Proof. Let us first convince ourselves that the minimum in $\min\{t \in \mathbb{R} : F(t) \geq u\}$ is in fact attained.

Consider for a given $u \in (0,1)$ the set $I_u = \{t \in \mathbb{R} : F(t) \geq u\}$. Note that I_u is non-empty, since $u < 1$ and $F(y) \to 1$ as $y \to \infty$. I_u has a finite left endpoint, say η_u, because $u > 0$ and $F(y) \to 0$ as $y \to -\infty$. Finally, $\eta_u \in I_u$, since F is a cdf and therefore right-continuous (consider $y_n = \eta_u + 1/n$, $n = 1, 2, \ldots$. Then, $u \leq F(y_n)$ for all n, hence $u \leq \lim_n F(y_n) = F(\eta_u)$ implying $\eta_u \in I_u$). In summary, the minimum in $\min\{t \in \mathbb{R} : F(t) \geq u\}$ is indeed attained.

We claim

$$\{(t,u) \in \mathbb{R} \times (0,1) : F^{-}(u) \leq t\} = \{(t,u) \in \mathbb{R} \times (0,1) : u \leq F(t)\}. \tag{3.4}$$

Taking an element from the left set, i.e., $(t,u) \in \mathbb{R} \times (0,1)$ satisfying $F^{-}(u) \leq t$, we have

$$F(t) \overset{\text{non-decr.}}{\geq} F(F^{-}(u)) \overset{\text{def.}}{=} F(\min\{t \in \mathbb{R} : F(t) \geq u\}) \overset{\text{def.}}{\geq} u$$

and (t,u) is contained in the right set. Conversely, for $(t,u) \in \mathbb{R} \times (0,1)$ satisfying $u \leq F(t)$ we have

$$F^{-}(u) \overset{\text{non-decr.}}{\leq} F^{-}(F(t)) \overset{\text{def.}}{=} \min\{r \in \mathbb{R} : F(r) \geq F(t)\} \overset{t \text{ in set}}{\leq} t,$$

proving (3.4).

Now we can complete the proof:

$$\Pr(X \leq t) \overset{\text{def.}}{=} \Pr(F^{-}(U) \leq t) \overset{(3.4)}{=} \Pr(U \leq F(t)) \overset{\text{def.}}{=} F_U(F(t)) \overset{(1.1)}{=} F(t).$$

□

□

Bibliography

[1] Blitzstein, J. K. and Hwang, J. (2019). *Introduction to probability*. Chapman and Hall/CRC.