STAT 230 Probability

University of Waterloo - Spring 2024

11th August 2024

LaTeX : Xing Liu

Contents

1 Introduction to Probability

1.1 Definitions of Probability

Definition (**Experienment**). An **experiment** is a situation involving chance or uncertainty that leads to results called outcomes.

Definition (**Outcome**). An **outcome** is the result of a single trial (attempt) of an experiment.

Definition (**Event**). An **event** is one or more outcomes of an experiment.

Definition (**Sample Space,** S). The set of ALL possible distinct outcomes in a random experiment is called the **sample space**, S.

Definition (**Probability**).

1. **Classical** definition:

$$P(\text{event}) = \frac{\text{\# of ways the event can occur}}{\text{\# of all possible outcomes}}$$

   provided all outcomes are equally likely.

2. **Relative frequency** definition:

   $P(\text{event}) = \text{proportion of times the event occurs in a long series of repeated experiment}$

3. **Subjective probability** definition:

$$P(\text{event}) = \text{how certain we are that the event will occur}$$

Note. All three definitions have serious limitations.

Example (**Rolling a die**). $S = \{1, 2, \dots, 6\}$. If a die is rolled once, the number 2 can be observed in exactly 1 out of 6 ways.

2 Mathematical Probability Models

2.1 Sample Spaces and Probability

Definition (**Sample Space**). A **sample space** S is a set of distinct outcomes for an experiment or process, with the property that in a single trial, one and only one of these outcomes occurs.

Note.

1. A sample space is NOT necessarily unique.
2. A discrete sample space consists of a finite or countable infinite set of outcomes.

Example. Roll a six-sided die, then $S = \{1, 2, 3, 4, 5, 6\}$, which is discrete.

Definition (**Simple/Compound Event**). An event in a discrete sample space is a subset $A \subset S$. If the event is indivisible so it contains only one point, e.g. $A_1 = \{a_1\}$, we call it a **simple event**. An event A made up of two or more simple events such as $A = \{a_1, a_2\}$, is called a **compound event**.

Definition. Let $S = \{a_1, a_2, ...\}$ be a discrete sample space. The **probabilities** $P(a_i)$, for $i = 1, 2, ...$ must satisfy the following two conditions:

(1) $0 \leq P(a_i) \leq 1$

(2) $\sum_{\text{all } i} P(a_i) = 1$

The set of probabilities $\{P(a_i), i = 1, 2, ...\}$ is called a **probability distribution** on S.

Note. $P(*)$ is a function where domain $= S$.

Definition (**Probability** $P(A)$ **of an Event** A). The probability $P(A)$ of an event A is the sum of the probabilities for all the simple events that make up A or $P(A) = \sum_{a \in A} P(a)$.

Example. For a fair die, each number is equally likely to occur. Therefore, $P(i) = \frac{1}{6}, i = 1, 2, ..., 6$. We can define the compound event $B =$ an even number is obtained. Then $B = \{2, 4, 6\}$, and $P(B) = P(2) + P(4) + P(6) = \frac{1}{2}$.

Example (**Card example**). Randomly draw one card from a standard deck of 52 cards. Find the probability of the card is a club.

Solution 1: Let $S = \{\text{spade, heart, diamond, club}\}$. Then $P(\text{club}) = \frac{1}{4}$.

Solution 2: Let $S =$ all 52 cards. Then $P(\text{club}) = \frac{13}{52} = \frac{1}{4}$ (the "club" event is compound here).

Example (**Coin example**). Toss a coin twice. Find the probability of getting exactly one head.

Solution: Let $S = \{\text{HH, HT, TH, TT}\}$ with all outcomes equally likely to happen. Then, $P(1H) = P(\text{TH}) + P(\text{HT}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

(Note that the sample space {0 heads, 1 head, 2 heads} will lead to a wrong answer as the outcomes are not equally likely)

We can use the term "odds" to describe probabilities in the following way.

Definition (**Odds**). The **odds in favour** of an event A occuring is the ratio $\frac{P(A)}{1-P(A)}$. The **odds against** the event is the reciprocal, $\frac{1-P(A)}{P(A)}$.

In the card example above, the odds in favour of clubs are $1 : 3$, we could also say the odds against clubs are $3 : 1$.

3 Probability and Counting Techniques

Remark. $P(A) = \dfrac{\#\text{ of oucomes in } A}{\#\text{ of outcomes in } S}$

3.1 Addition and Multiplication Rules

Definition (**Counting Rules**).

1. **Addition Rule:** Suppose we can do job 1 in p ways and job 2 in q ways. Then we can do either job 1 **OR** job 2 (but not both), in $p + q$ ways.
2. **Multiplication Rule:** Suppose we can do job 1 in p ways and, for each of these ways, we can do job 2 in q ways. Then we can do both job 1 **AND** job 2 in $p \times q$ ways.

Note.

- "OR": interpreted as **addition**.
- "AND": interpreted as **multiplication**.
- "With" replacement: every time an object is selected, we put it back into the pool of possible objects (could be picked again).
- "Without" replacement: every time an object is selected, it is NOT put back.

Example. A bag contains 3 blue marbles and 5 red marbles. Find the probability of selecting two blue marbles **with and without** replacement.

Solution: With replacement, $P(BB) = P(B \text{ AND } B) = \frac{3}{8} \times \frac{3}{8} = \frac{9}{64}$.
Without replacement, $P(BB) = P(B \text{ AND } B) = \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$.

3.2 Counting Arrangements or Permutations

Definition (Permutation). A **permutation** is an arrangement of objects in a definite order (order matters and selection is done without replacement).

Example. How many different ordered arrangements of the letters a, b, and c are possible.

Solution: $\underbrace{3\ 2\ 1}_{\text{letters available for each selection}}$ So, $3 \times 2 \times 1 = 6$ ways.

To generalize, in each case we count the # of arrangements by counting the # of ways we can fill the positions in the arrangement. Suppose we have n symbols, we can make:

- $n! = n \times (n-1) \times \cdots \times 1$ arrangements of length n using each symbol once and only once (the # of permutations of n distinct objects).
- $n^{(k)} = n \times (n-1) \times \cdots \times (n-k+1)$ arrangements of length k using each symbol at most once. Note that $n^{(k)} = \frac{n!}{(n-k)!} = {}^nP_k$ (read "n to k factors").
- $n^k = n \times n \times \cdots \times n$ arrangements of length k using each symbol more than once.

Theorem (**Stirling's Approximation**).

For large n there is an approximation to $n!$, it states that $n!$ is **asymptotically equivalent** to $(\frac{n}{e})^n\sqrt{2\pi n}$. Note that the sequence $\{a_n\}$ is asymptotically equivalent to the sequence $\{b_n\}$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$.

Note. This implies that as $n \to \infty$, the approximation becomes better and better (i.e. $\lim_{n\to\infty} \frac{(\frac{n}{2})^n\sqrt{2\pi n}}{n!} = 1$).

Example. A pin number of length four is formed by randomly selecting four digits from the set $\{0, 1, 2, \ldots, 9\}$ **with replacement**. Find the probability of the events:

- A: the pin number is even.
- B: the pin number contains at least one 1.

Solution:

(a) $P(A) = \frac{\text{\# of pins that are even}}{\text{\# of pins in the sample space}} = \frac{10\times10\times10\times5}{10\times10\times10\times10} = \frac{1}{2}$.

(b) We can use the complement. Find the pin numbers that do not contain 1.
$P(B) = 1 - P(\overline{B}) = 1 - \frac{9^4}{10^4} = 0.3439$.

Exercise: try this **without replacement**.

Example. Five separate awards (best scholarship, best leadership qualities, and so on) are to be presented to selected students from a class of 30. How many different outcomes are possible if:

(a) A student can receive any number of awards?
Solution: 30^5.

(b) Each student can receive at most 1 award?
Solution: $30 \times 29 \times 28 \times 27 \times 26 = 30^{(5)}$.

Example.

(a) In how many ways can 3 boys and 3 girls sit in a row?
Solution: $6! = 120$.

(b) In how many ways can 3 boys and 3 girls sit in a row if the boys and the girls are each to sit together?
Solution: $2! \times 3! \times 3! = 72$.

(c) In how many ways if only the boys must sit together?
Solution: $4! \times 3! = 144$.

(d) In how many ways if no two people of the same sex are allowed to sit together?
Solution: $3! \times 3! \times 2! = 72$.

3.3 Counting Subsets or Combinations

Definition (Combination)**.** A combination is an unordered selection of k objects chosen from n objects.

Remark (**Number of subsets of size** k). We use ${}^nC_k = \binom{n}{k}$ to denote the # of subsets of size k that can be selected from a set of n objects. Then, $m \times k! = n^{(k)}$ and we have

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

Properties of $\binom{n}{k}$**:**

1. $\binom{n}{k} = \frac{n^{(k)}}{k!} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}$ if $n \in \mathbb{R}$ and k is a non-negative integer
2. $\binom{n}{k} = \binom{n}{n-k}$ for all $k = 0, 1, \ldots, n$

3. $\binom{n}{0} = \binom{n}{n} = 1$
4. $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$
5. **Binomial Theorem:** $(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2} + \cdots + \binom{n}{n}x^n$

3.4 Number of Arrangements when Symbols are Repeated

Example. Suppose the letters of the word "STATISTICS" are arranged at random. Find the probability of the event G that the arrangement begins and ends with "S".

Solution: $P(G) = \frac{\text{\# of subsets in G}}{\text{\# of subsets in the sample space}}$.

The number of equally probable outcomes in the sample space S is:

$$\underbrace{\binom{10}{3}}_{\text{S}}\underbrace{\binom{7}{3}}_{\text{T}}\underbrace{\binom{4}{2}}_{\text{I}}\underbrace{\binom{2}{1}}_{\text{C}}\underbrace{\binom{1}{1}}_{\text{A}} = \frac{10!}{3!7!}\frac{7!}{3!4!}\frac{4!}{2!2!}\frac{2!}{1!1!}\frac{1!}{1!0!} = \frac{10!}{3!3!2!1!1!}.$$

The number of arrangements in G:

$$\underbrace{\binom{8}{1}}_{\text{S}}\underbrace{\binom{7}{3}}_{\text{T}}\underbrace{\binom{4}{2}}_{\text{I}}\underbrace{\binom{2}{1}}_{\text{C}}\underbrace{\binom{1}{1}}_{\text{A}} = \frac{8!}{1!3!2!1!1!}.$$

Thus, $P(G) = \dfrac{\frac{8!}{1!3!2!1!1!}}{\frac{10!}{3!3!2!1!1!}} = \dfrac{1}{15}$.

Remark (**Number of arrangements when symbols are repeated**).
If we have n_i symbols of type i, $i = 1, 2, \ldots, k$ with $n_1 + n_2 + \cdots + n_k = n$, then the # of arrangements using all of the symbols is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_k}{n_k} = \frac{n!}{n_1!n_2!\cdots n_k!}.$$

Example. Find the probability a bridge hand (13 cards picked at random from a standard deck of 52 cards without replacement) has

(a) at least 1 Ace.

(b) 6 spades, 4 hearts, 2 diamonds and 1 club.

(c) a 6-4-2-1 spilt between the 4 suits.

Solution (a):

$$\begin{aligned} P(\text{at least 1 Ace}) &= 1 - P(0 \text{ Aces}) \\ &= 1 - \frac{\binom{4}{0}\binom{52-4}{13}}{\binom{52}{13}} \end{aligned}$$

Solution (b): $\frac{\binom{13}{6}\binom{13}{4}\binom{13}{2}\binom{13}{1}}{\binom{52}{13}}$.

Solution (c): $4! \times \binom{13}{6}\binom{13}{4}\binom{13}{2}\binom{13}{1}$.

Example. If 12 people are to be divided into 3 committees of respective sizes 3, 4, and 5, how many divisions are possible?

Solution: $\binom{12}{3}\binom{9}{4}\binom{5}{5} = 27720$.

Example. A person has 8 friends, of whom 5 will be invited to a party.

(a) How many choices are there if 2 friends are feuding and will not attend together?

Solution: $\binom{8}{5} - \binom{2}{2}\binom{6}{3} = 36$ ways (sample space - both feuding friends come).

(b) How many choices are there if 2 of the friends will only attend together?

Solution: $\underbrace{\binom{2}{2}\binom{6}{3}}_{\text{Both are invited}} + \underbrace{\binom{2}{0}\binom{6}{5}}_{\text{Both don't go}} = 20 + 6 = 26.$

Example. There are 5 blue beads and 4 green beads to be arranged in a row on a string. The two ends of the string are not connected. Beads with the same colour are indistinguishable. Find the probability of the following events:

(a) A = "All 5 blue beads are adjacent to each other".

Solution: $P(A) = \frac{\binom{5}{1}\binom{4}{4}}{\binom{9}{5}\binom{4}{4}} = \frac{\frac{5!}{1!4!}}{\frac{9!}{1!4!}}$.

(b) B = "None of the green beads is adjacent to any other green beads".

Solution: $P(B) = \frac{\binom{6}{4}}{\binom{9}{5}\binom{4}{4}}$.

4 Probability Rules and Conditional Probability

4.1 General Methods

Theorem (Probability Rules).

1. **Rule 1:** $P(S) = 1$.

   Proof. $P(S) = \sum_{a \in S} P(a) = \sum_{\text{all } a} P(a) = 1.$ □

2. **Rule 2:** For any event A, $0 \leq P(A) \leq 1$.

   Proof. $P(A) = \sum_{a \in A} P(a) \leq \sum_{a \in S} P(a) = 1$ and since each $P(a) \geq 0$, we have $0 \leq P(A) \leq 1$. □

3. **Rule 3:** If A and B are two events with $A \subseteq B$ (that is, all of the points in A are also in B), then $P(A) \leq P(B)$.

   Proof. $P(A) = \sum_{a \in A} P(a) \leq \sum_{a \in B} P(a) = P(B)$, so $P(A) \leq P(B)$. □

Venn Diagrams

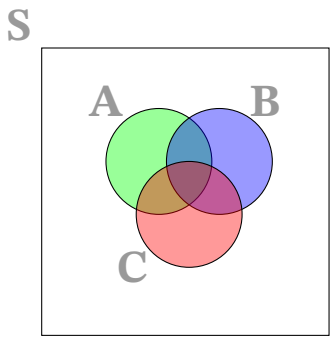


Figure 1: $A \cup B \cup C$.

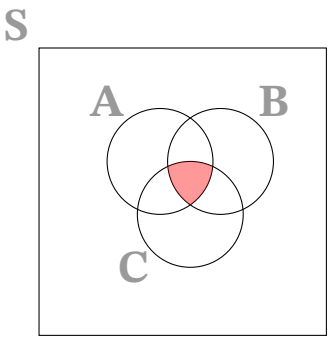


Figure 2: $A \cap B \cap C$.

Theorem (De Morgan's Laws).

(a) $\overline{A \cup B} = \overline{A} \cap \overline{B}$

(b) $\overline{A \cap B} = \overline{A} \cup \overline{B}$

Proof.

Use venn diagrams to prove it as an exercise. □

4.2 Rules for Unions of Events

Theorem (**Rule 4a: Addition Law of Probability or the Sum Rule**).
Let A and B be events (not necessarily mutually exclusive). Then

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

Proof.

$$\begin{aligned}
P(A) + P(B) &= \sum_{a \in A} P(a) + \sum_{a \in B} P(a) \\
&= \left(\sum_{a \in A \cap \overline{B}} P(a) + \sum_{a \in A \cap B} P(a) \right) + \left(\sum_{a \in \overline{A} \cap B} P(a) + \sum_{a \in A \cap B} P(a) \right) \\
&= \left(\sum_{a \in A \cap \overline{B}} P(a) + \sum_{a \in A \cap B} P(a) + \sum_{a \in \overline{A} \cap B} P(a) \right) + \sum_{a \in A \cap B} P(a) \\
&= \sum_{a \in A \cup B} P(a) + \sum_{a \in A \cap B} P(a) \\
&= P(A \cup B) + P(A \cap B)
\end{aligned}$$

Rearranging the equation, we obtain:

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

as desired. This can also be justified by using a Venn diagram. In the expression $P(A) + P(B)$, the points in $A \cap B$ have their probability counted twice, so they need to be subtracted once. □

Theorem (**Rule 4b: Probability of the Union of Three Events**).
Let A, B and C be events. then

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

Proof. Use venn diagrams to prove it as an exercise. □

Theorem (**Rule 4c: Probability of the Union of n Events**).
A generalization of the above rules to n events $A_1, A_2, \ldots, A_n$. This is often referred to as the *inclusion-exclusion principle*.

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \sum_{i<j<k<l} P(A_i \cap A_j \cap A_k \cap A_l) + \cdots$$

(where the subscripts are all distinct).

Proof. This can be proved using Rule 4a and induction. □

Definition (**Mutually Exclusive**). Events A and B are **mutually exclusive** if

$$A \cap B = \emptyset.$$

Remark. We can extend this definition to events $A_1, A_2, \ldots, A_n$.

Example. If a die is rolled twice, the events $A =$ "2 occurs on the 1st roll" and $B =$ "total is 10" are mutually exclusive events.

Theorem (**Rule 5a: Probability of the Union of Two Mutually Exclusive Events**).
Let A and B be mutually exclusive events. then

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

Theorem (**Rule 5b: Probability of the Union of n Mutually Exclusive Events**).
In general, let $A_1, A_2, \ldots, A_n$ be mutually exclusive events. then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i).$$

Proof. This can be proved from Rule 5a using induction or as an immediate consequence of Rule 4c. □

Theorem (**Probability of the Complement of an Event**).
For any event A, we have

$$P(A) = 1 - P(\overline{A}).$$

Proof. A and $\overline{A}$ are mutually exclusive and $A \cup \overline{A} = S$, so by Rule 5a,

$$\begin{aligned} P(A \cup \overline{A}) &= P(A) + P(\overline{A}) \\ 1 &= P(A) + P(\overline{A}) \qquad \left(\text{since } P(A \cup \overline{A}) = 1\right) \\ \implies P(A) &= 1 - P(\overline{A}). \end{aligned}$$

□

Example. Two fair dice are rolled. Find the probability that at least one of them turns up a six.
Solution 1: Defining appropriate events:
$A = \{\text{outcome from first die is a 6}\}$ and $B = \{\text{outcome from second die is a 6}\}$. Therefore, either first is a 6 and second is a six OR both give a 6.
$\implies P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{6} + \frac{1}{6} - \frac{1\times1}{36} = \frac{11}{36}$.
Solution 2: Using the complement:
$\implies P(\text{at least one 6}) = 1 - P(\text{zero 6s}) = 1 - \frac{5\times5}{36} = \frac{11}{36}$.

4.3 Intersections of Events and Independence

Definition (**Independent and Dependent Events**). Events A and B are **independent events** $\iff P(A \cap B) = P(A)P(B)$. Otherwise, we call the events **dependent**.

Independence means that knowing information about A will not affect the information about B.

Definition. The events $A_1, A_2, \ldots A_n$ are mutually independent if and only if

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$$

for all sets $(i_1, i_2, \ldots, i_k)$ of distinct subscripts chosen $(1, 2, \ldots, n)$.

4.4 Conditional Probability

Definition (**Conditional Probability**).

The **conditional probability** of event A, given event B, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) > 0.$$

Note. $P(A|B) = 1 - P(\overline{A}|B)$.

Theorem.

Suppose A and B are two events defined on a sample space S s.t. $P(A) > 0$ and $P(B) > 0$.

Then A and B are independent events $\iff$ either of the following statements is true;

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

Example (**exercise**). You ask your roommate to water a sickly plant while you are on vacation. Without water, the plant will die with probability 0.8 and with water it will die with probability 0.1. Your roommate will remember to water the plant with probability 0.85.

If the plant is alive when you return, what is the probability that your roommate remembered to water it?

Answer: The probability is 0.9623.

Example (**exercise**). The probability a randomly selected male is colour blind is 0.05, whereas the probability a female is colour blind is 0.0025. If the population is 50% male, what fraction of the population is colour blind?

Answer: The probability is 0.02625.

4.5 Product Rules, Law of Total Probability and Bayes' Theorem

Theorem (**Rule 7: Product Rules**).
Let $A, B, C, D, \ldots$ be arbitrary events in a sample space. Assume that $P(A) > 0$, $P(A \cap B) > 0$, and $P(A \cap B \cap C) > 0$. Then

$$\begin{aligned} P(A \cap B) &= P(A)P(B|A) \\ P(A \cap B \cap C) &= P(A)P(B|A)P(C|A \cap B) \\ P(A \cap B \cap C \cap D) &= P(A)P(B|A)P(C|A \cap B)P(D|A \cap B \cap C) \\ &\vdots \end{aligned}$$

Proof. Use the definition of the conditional probability $P(B|A)$. □

Note. Used when finding intersections given conditional probabilities.

Theorem (**Rule 8: Law of Total Probability**).
Let $A_1, A_2, \ldots, A_k$ be a partition of the sample space S into disjoint (mutually exclusive) events, that is

$$A_1 \cup A_2 \cup \cdots \cup A_k = S \quad \text{and} \quad A_i \cap A_j = \varnothing \quad \text{if } i \neq j.$$

Let B be an arbitrary event in S. Then

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_k) \\ &= \sum_{i=1}^{k} P(B|A_i)P(A_i). \end{aligned}$$

Example (**exercise**). At a police spot check, 10% of cars stopped have defective headlights and a faulty muffler. 15% have defective headlights and a muffler which is satisfactory. If a car which is stopped has defective headlights, what is the probability that the muffler is also faulty?
Answer: The probability is 0.4.

Theorem (**Bayes' Theorem**).

Suppose A and B are events defined on a sample space S. Suppose also that $P(B) > 0$. then

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

Proof.

$$\begin{aligned}
\frac{P(B|A)P(A)}{P(B|\overline{A})P(\overline{A}) + P(B|A)P(A)} &= \frac{P(A \cap B)}{P(B \cap \overline{A}) + P(B \cap A)} && \text{by the Product Rule} \\
&= \frac{P(A \cap B)}{P(B)} && \text{by the Law of Total Probability} \\
&= P(A|B).
\end{aligned}$$

□

Remark. Bayes' Theorem applies when given "opposite conditions", i.e. we want to find $A|B$ given $B|A$.

5 Discrete Random Variables

5.1 Random Variables and Probability Functions

Definition (**Random Variable**). A **random variable** is a function that assigns a real number to each point in a sample space S.

Example. Suppose an experiment consists of tossing a coin 3 times. The sample space

$$S = \{\text{HHH, THH, HTH, HHT, HTT, THT, TTH, TTT}\}.$$

Then X = number of heads that occur would be a random variable, associated with range $A = \{0, 1, 2, 3\}$.

Events	Outcomes from S
$X = 0$	{TTT}
$X = 1$	{HTT, THT, TTH}
$X = 2$	{HHT, HTH, THH}
$X = 3$	{HHH}

Ex. $P(X = 1) = P(\text{HTT} \cup \text{THT} \cup \text{TTH}) = P(\text{HTT}) + P(\text{THT}) + P(\text{TTH}) = \frac{3}{8}$.

Definition (**Discrete Random Variables**). **Discrete random variables** take integer values or, more generally, values in a countable set.

Note. A set is countable if its elements can be placed in a one-to-one correspondence with a subset of the positive integers.

Definition (**Continuous Random Variables**). **Continous random variables** take values in some interval of real numbers like $(0, 1)$ or $(0, \infty)$ or $(-\infty, \infty)$.

Note. Cardinality of the real numbers in an interval is NOT countable.

Definition (**Probability Function & Probability Distribution**).
Let X be a discrete random variable with range $(X) = A$. The **probability function** of X is

the function

$$f(x) = P(X = x), \quad \text{defined } \forall\, x \in A.$$

The set of pairs $\{(x, f(x)) : x \in A\}$ is called the **probability distribution** of X.

All probability functions must have two properties:

1. $f(x) \geq 0\ \forall x \in A$.
2. $\sum_{\text{all } x \in A} f(x) = 1$.

Remark. It follows that $0 \leq f(x) \leq 1\ \forall x \in A$.

Example. The random variable X has probability function given by

x	0	1	2	3	4
$f(x)$	$0.1c$	$0.2c$	$0.5c$	c	$0.2c$

(a) Find c.

Solution: Since $\sum_{\text{all } x \in A} f(x) = 1$. Then $0.1c + 0.2c + 0.5c + c + 0.2c = 1$. This gives $c = 0.5$.

(b) Find $P(X > 2)$.

Solution: $P(X > 2) = P(X \geq 3) = P(X = 3) + P(X = 4) = c + 0.2c = 1.2c = 0.6$.

Definition (**Cumulative Distribution Function**). The **cumulative distribution function** of the discrete random variable X is the function usually denoted by $F(X)$.

$$F(x) = P(X \leq x) \quad \text{defined } \forall\, x \in \mathbb{R}$$

In general, $F(x)$ can be obtained from $f(x)$ using:

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u).$$

Theorem (**Properties of a CDF** $F(x)$)**.**

1. $F(x)$ is a non-decreasing function of $x\ \forall x \in \mathbb{R}$. For example, $P(X \leq 8)$ cannot be less than $P(X \leq 7)$.
2. $0 \leq F(x) \leq 1\ \forall x \in \mathbb{R}$.
3. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

We can also obtain $f(x)$ from $F(x)$. If X takes on integer values then for values x s.t. $x \in A$ and $x-1 \in A$,

$$f(x) = F(x) - F(x-1).$$

In other words:

$$P(X = x) = P(X \leq x) - P(X \leq x-1).$$

We notice that $f(x)$ represents the size of the jump in $F(x)$ at the point x.

Plots of $f(x)$ and $F(x)$
For discrete random variables, the cdf $F(x)$ is represented as a step function, whereas the pf $f(x)$ is represented by a histogram.

5.2 Discrete Uniform Distribution

Physical Setup: Suppose the range of X is $\{a, a+1, \dots, b\}$ where a and b are integers and suppose all values are equally probable. Then X has a Discrete Uniform distribution on the set $\{a, a+1, \dots, b\}$. The variables a and b are called the parameters of the distribution.

Probability Function: There are $b-a+1$ values in the set $\{a, a+1, \dots, b\}$ so the probability of each value must be $\frac{1}{b-a+1}$ in order that $\sum_{x=a}^{b} f(x) = 1$. Therefore

$$f(x) = P(X = x) = \begin{cases} \frac{1}{b-a+1} & \text{for } x = a, a+1, \dots, b \\ 0 & \text{otherwise} \end{cases}$$

Example. Suppose a fair die is thrown once and let X be the number on the face. Find the cumulative distribution function of X.
Solution: $x = 1, 2, 3, 4, 5, 6$ and $X \sim \text{Unif}[1, 6]$.

$$f(x) = P(X = x) = \begin{cases} \frac{1}{6} & \text{for } x = 1, 2, \dots, 6 \\ 0 & \text{otherwise} \end{cases}$$

$F(1) = P(X \leq 1) = P(X = 1) = \frac{1}{6}$ and $F(2) = P(X \leq 2) = P(X = 1) + P(X = 2) = \frac{2}{6}$. So

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{for } x < 1 \\ \frac{[x]}{6} & \text{for } 1 \leq x < 6 \\ 1 & \text{for } x \geq 6 \end{cases}.$$

5.3 Hypergeometric Distribution

Physical Setup: We have ***N objects*** that can be classified into exactly ***two*** distinct types, "success" (S) vs. "failure" (F).

$r = \#$ of successes and $N - r = \#$ of failures.

We pick $\underline{n}$ objects at random **without** replacement. If X represents the number of successes, then X has a hypergeometric distribution.

Example. Of the 120 applicants for a job, only 80 are qualified. In total, 5 applicants are picked at random for an interview. If X represents the number of qualified applicants that are interviewed, then X has a hypergeometric distribution, where:

- $N = 120$
- $r = 80$ since $N - r = 40$
- $n = 5$
- Possible values of x are 0, 1, 2, 3, 4, 5.

Probability Function: Using counting techniques we note there are $\binom{N}{n}$ points in the sample space S if we don't consider order of selection. There are $\binom{r}{x}$ ways to choose the x success objects from the r available and $\binom{N-r}{n-x}$ ways to choose the remaining $(n - x)$ objects from the $(N - r)$ failures. Hence

$$f(x) = P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

where $x \geq \max(0, n - (N - r))$ and $x \leq \min(r, n)$.

Note. Do not need to memorize this formula because it will be easy to realize based on the given context.

Example (**Above Continued**)**.**

Find the probability that only two of the five selected will be qualified for the job.

Solution: $P(X=2)=\frac{\binom{80}{2}\binom{40}{3}}{\binom{120}{5}}=0.164.$

Example. In the game of Texas Hold'em, players are each dealt two private cards, and five community cards are dealt face-up on the table. Each player makes the best 5-card hand they can with their two private cards and the five community cards. What is the probability that a particular player can make a flush of spades (i.e. 5 spades or more)?

Solution: $X=\#$ of spades among 7 cards. $N=52$, $r=13$, $N-r=39$, $n=7$, and $x=0,1,2,\ldots,7$.

$$\begin{aligned}P(X\geq 5)&=P(X=5)+P(X=6)+P(X=7)\\&=\frac{\binom{13}{5}\binom{39}{2}+\binom{13}{6}\binom{39}{1}+\binom{13}{7}\binom{39}{0}}{\binom{52}{7}}\\&=0.0076.\end{aligned}$$

Example. A manufacturer of auto parts just shipped 25 auto parts to a dealer. Later, it found out that 5 of those parts were defective. By the time the company manager contacted the dealer, 4 auto parts from that shipment had been sold. What is the probability that 3 of those 4 parts were good parts and one was defective?

Solution: $X=\#$ of defectives selected, $N=25$, $r=5$, and $n=4$.

$$P(X=1)=\frac{\binom{5}{1}\binom{25-5}{4-1}}{\binom{25}{4}}=0.45.$$

5.4 Binomial Distribution

Physical Setup: Suppose an experiment has ***two*** possible outcomes, "success" and "failure". Let $P(\text{Success})=p$ and hence $P(\text{Failure})=1-p$. Repeat the experiment n ***independent*** times. Let X be the number of successes obtained, we say X has a ***Binomial distribution***.

We write $X\sim Bin(n,p)$ with n total trials and p probability to success.

Note. The ***n individual*** experiments are called "trials" or "Bernoulli trials" and the process is called a Bernoulli process or Binomial process.

Underlying assumptions for the Binomial Distribution: (Tims)

- T: Two outcomes
- I: Independent trials
- M: Multiple trials
- S: Same probability of success in each trial.

Example. A fair coin is tossed 12 times. Let X be the number of heads obtained. Then:

- Two outcomes: Head (success) vs. Tail (failure)
- Independent trials: flips are independent of each other
- Multiple trials: $n = 12$
- Same P(Success) in each trial $= p = 0.5$.

Thus, $X \sim Bin(12, 0.5)$ and $x = 0, 1, 2, \dots, 12$.

Probability Function: There are $\binom{n}{x}$ arrangements of x successes and $(n - x)$ failures over the n trials. Each arrangement has probability $p^x(1-p)^{n-x}$. Therefore,

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n \text{ and } 0 < p < 1.$$

Remark. Check the above function by using the property that $\sum_{\text{all } x} f(x) = 1$ for $0 < p < 1$.

Example. Seventy-five percent of students at a college with a large student population use Instagram. A sample of five students from this college is selected. What is the probability that at least 3 use Instagram?

Solution: $X = \#$ of students that use Instagram among 5 selected. Then $X \sim Bin(5, 0.75)$ for $x = 0, 1, \dots, 5$.

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) = 1 - P(X \leq 2).$$

Remark.
The probability of at least 'x': $P(X \geq x) = 1 - P(X \leq (x-1))$.
The probability of more than 'x': $P(X > x) = 1 - P(X \leq x)$.

Binomial vs. Hypergeometric

Similarities

- Both have two types of outcomes, success and failure.
- The experiment is repeated n times.
- X records the number of successes.

Differences

- Binomial reqires **independent** trials, where the probability of success is the same in each trial.
- In Hypergeometric, the draws are made from a fixed number of objects N **without** replacement. Hence, the trials are **not independent**.

Example. Suppose we have 20 cans of drinks placed in a big ice container such that the labels are not visible. It is known that 12 are coke and 8 are juice. We randomly pick 10 cans. Find the probability that 3 are coke.

Solution: Assume the draws are down without replacement (correct approach here). We have $N = 20$, $r = 12$, $N - r = 8$, and $n = 10$. So $X \sim$ Hypergeo$(20, 12, 10)$.

Thus, $P(X = 3) = \frac{\binom{12}{3}\binom{8}{7}}{\binom{20}{10}} = 0.00953$.

Remark. If N is large and n (number of objects being drawn) is relatively small, then binomial can be used as an approximation for the hypergeometric.

Rule of thumb: If the sample size (number of trials) n is at most 5% of the population size, the experiment can be analyzed as though it were exactly a Binomial experiment.

Example. Megan audits 130 clients during a year and finds irregularities for 26 of them.

(a) Give an expression for the probability that 2 clients will have irregularities when 6 of her clients are picked at random.

Solution: Let $X = \#$ of clients with irregularities among 6 selected, $N = 130$, $r = 26$, $n = 6$.

$$P(X = 2) = \frac{\binom{26}{2}\binom{130-26}{6-2}}{\binom{130}{6}} = 0.251.$$

(b) Evaluate your answer to (a) using a suitable approximation.

Solution: $X \sim \text{Bin}(6, \frac{26}{130})$. Then we have

$$P(X=2) = \binom{6}{2}\left(\frac{26}{130}\right)^2\left(1-\frac{26}{130}\right)^{6-2} = 0.246 \approx 0.25.$$

Notice that $\frac{n}{N} = \frac{6}{130} = 0.046 < 5\%$.

5.5 Negative Binomial Distribution

Physical Setup: similar to the Binomial Distribution. Do experiment until we obtain k successes. However now, X records the number of failures obtained before the kth success, $X \sim \text{NB}(k, p)$.

Example. Draw cards with replacement until you get 3 Aces. Let $X = \#$ of Non-Aces that we obtain before the third Ace. Then the distribution of X is $X \sim \text{NB}(k = 3, p = P(\text{Success}) = \frac{4}{52})$ with $x = 0, 1, ...$

Probability Function: In total, we have x failures and k successes, so there are $x + k$ trials, the last trial **MUST** be a success. Hence in the first $x+k-1$ trials, we observe x failures and $(k-1)$ successes. Therefore, there are $\binom{x+k-1}{x} = \binom{x+k-1}{k-1}$ arrangements, each arrangement has probability $p^k(1-p)^x$. Hence

$$f(x) = P(X=x) = \binom{x+k-1}{x}p^k(1-p)^x \quad \text{for } x = 0, 1, \ldots \text{ and } 0 < p < 1.$$

Note. An alternate version of Negative Binomial Distribution defined X to be the total number of trials needed to get the kth success. We'll have something like: $X - 4 \sim \text{NB}(k, p)$, where 4 is just an example here, representing the number of successes before the 5th success.

Binomial vs. Negative Binomial

- **Binomial:** We know that we have n trials in advance but do not know the # of successes we will obtain until after the experiment.
- **Negative Binomial:** We know the number k of successes in advance but do not know the # of trials that will be needed to obtain this # of success until after the experiment.

Example. Two athletic teams, A and B, play a best-of-three series of games (i.e the first team to win two games is the overall winner). Suppose team A is the stronger team and will win any game with probability 0.6, independently from other games. Let X be the number of games lost before Team A wins twice. Find the probability that Team A is the overall winner.

Solution: $X \sim \text{NB}(2, 0.6)$.

$$\begin{aligned}\text{Probability} &= P(X=0) + P(X=1) \\ &= \binom{0+2-1}{0}(0.6)^2(0.4)^0 + \binom{1+2-1}{1}(0.6)^2(0.4)^1\end{aligned}$$

Example. A startup is looking for 5 investors. Each investor will independently agree with probability 20%. A founder asks investors one at a time until they get 5 "yes". Let X be the total # of investors asked. Find $f(x)$ and $f(6)$.

Solution: Let $Y = \#$ of "No"s until the 5th yes is obtained. Then $Y \sim \text{NB}(5, 0.2)$, $y = 0, 1, 2, \ldots$. And $X = Y + 5$.

$$\begin{aligned}f_X(x) &= P(X = x) \\ &= P(X = x = Y + 5) \\ &= P(Y = x - 5) \\ &= f_Y(x - 5) \\ &= \binom{x-5+5-1}{x-5} p^5(1-p)^{x-5} \\ &= \binom{x-1}{x-5} p^5(1-p)^{x-5},\ x = 5, 6, 7, \ldots\end{aligned}$$

$f_x(6) = P(X = 6) = \binom{6-1}{6-5} 0.2^5(1-0.2)^{6-5} = 0.00128.$

5.6 Geometric Distribution

Physical Setup: Independent Bernoulli trials, each having two possible outcomes (Success vs. Failure). The probability, p, of success is the same each time. However, $X = \#$ of failures before the **FIRST** success (i.e. a Negative Binomial distribution with $k = 1$).

We write $X \sim \text{Geo}(p)$.

Probability Function: There is only the one way with x failures followed by 1 success.

$$f(x) = P(X = x) = (1-p)^x p \quad \text{for } x = 0, 1, \ldots \text{ and } 0 < p < 1.$$

Remark. Alternatively, substitute $k = 1$ in $f(x)$ for the Negative Binomial.

In summary, we notice that Binomial, Negative Binomial and Geometric all assume:

1. Two outcomes in each trial.
2. Independent Trails.
3. Each trial has the same probability of success.

Example. A company receives 60% of its orders over the internet.

(a) What is the probability that the fifth order received is the first internet order?
Solution: Let $X = \#$ of orders not over the internet untill the first internet order. Then $X \sim$ Geo(0.6), $x = 0, 1, 2, \ldots$. We have $P(X = 4) = (0.6)^1(1-0.6)^4$.

(b) What is the probability that the eighth order received is the fourth internet order?
Solution: Let $Y =$ total number of orders before the 4th internet order. $Y \sim$ NB(4, 0.6). Then $f_y(4) = P(Y = 4) = \binom{4+4-1}{4}(0.6)^4(1-0.6)^4$. Since we have 4 successes and 4 failures.
Alternatively, $Y =$ number of non-internet orders before 4th internet. Then $Y - 3 \sim$ NB(4, 0.6).

5.7 Poisson Distribution from Binomial

Physical Setup: $X = \#$ of events of some type. The events occur according to some rate μ, where $\mu > 0$. We write $X \sim$ Poisson(μ).

Probability Function:

$$f(x) = P(X = x) = \frac{e^{-\mu}\mu^x}{x!} \quad \text{for } x = 0, 1, \ldots$$

where $\mu > 0$.

Remark. Poisson arises from Binomial when $n \to \infty$ and $p \to 0, X \sim \text{Bin}(n, \frac{\mu}{n})$. ($\mu = np$)

Example. Consider the Tim Hortons event "Roll up the Rim". We are told that 1 in 9 cups are winners. Say you buy 100 cups, assuming they are independent, and use the Poisson approximation to solve for the probability that you get no more than 10 winning cups.

Solution: Find the exact probability using Binomial.

Let $X = \#$ of winning cups among 100. Then $X \sim \text{Bin}(100, \frac{1}{9})$. We have $P(X \leq 10) = 0.439$.

Now, use Poisson approximation. We have $\mu = np = 100 \times \frac{1}{9} = \frac{100}{9}$. Then

$$P(X \leq 10) = P(X = 0) + P(X = 1) + \cdots + P(X = 10) = e^{-\frac{100}{9}} \left[\frac{\mu^0}{0!} + \frac{\mu^1}{1!} + \cdots + \frac{\mu^{10}}{10!} \right] = 0.447.$$

Example. If you buy a lottery ticket in 50 lotteries, in each of which your chance of winning a prize is 1/100, what is the (approximate) probability that you will win a prize

(a) At least once,

(b) Exactly once,

(c) At least twice?

Solution: We have $\mu = np = 50 \times \frac{1}{100} = 0.5$.

(a) $P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{e^{-0.5}(0.5)^0}{0!} = 0.3935.$

(b) $P(X = 1) = \frac{e^{-0.5}(0.5)^1}{1!} = 0.3023.$

(c) $P(X \geq 2) = 1 - P(X \leq 1) = 1 - P(X = 0) - P(X = 1) = 0.09017.$

Note.

1. If p is close to 1, we can still use Poisson to approximate Binomial, simply by interchanging the labels "success" and "failure". Now, we can get P(success) is close to 0.

5.8 Poisson Distribution from Poisson Process

Definition (**Poisson Process**).

The following three conditions together define a **Poisson Process**:

1. **Independence:** the number of occurrences in non-overlapping intervals are independent.
2. **Individuality:** $P(2 \text{ or more events in } (t, t+\Delta t)) = o(\Delta t)$ (close to 0) as $\Delta t \to 0$.
3. **Homogeneity or Uniformity:** Events occur at a homogeneous (uniform) rate λ over time so that $P(\text{one event in } (t, t+\Delta t)) = \lambda \Delta t + o(\Delta t)$.

In a Poisson process with rate of occurrence λ, the number of event occurrences X in a time interval of length t has a Poisson distribution with $\mu = \lambda t$.

$$f(x) = P(X = x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, x = 0, 1, 2, \ldots$$

Interpreting μ and λ.

1. λ refers to the intensity or rate of occurrence. It represents the average rate of occurrence of events per unit of time.
2. $\mu = \lambda t$ represents the average number of occurrences in t units of time.

Example. Suppose earthquakes recorded in Ontario each year follow a Poisson process with an average of 6 per year. What is the probability that 7 will be recorded in a two year period?

Solution: Let $X = \#$ of earthquakes in the two year period ($t = 2$). And the intensity of earthquakes is $\lambda = 6$ per year. So $\mu = \lambda t = 6 \times 2 = 12$ in two years. Then, $f(7) = \frac{e^{-12}12^7}{7!} = 0.0437$.

The Poisson process also applies when "events" occur randomly in space. If X represents the number of events in a volume or area in space of size v, and if λ is the average number of events per unit voilume (or area), then X has a Poisson distribution with $\mu = \lambda v$.

The model is valid when we replace "time" by "volume" or "area".

Example. Coliform bacteria occur in a river water with an average intensity of 1 bacteria per 10 cubic centimeters (cc) of water. Find:

(a) The probability there are no bacteria in a 20cc sample of water which is tested.

Solution: Let $X = \#$ of Coliform bacteria observed in a specified volume. Then $X \sim \text{Poi}(\lambda = 1 \text{ per } 10\text{cc})$. We have $\mu = \lambda t = \frac{1}{10} \times 20 = 2$ per 20cc. So $P(X = 0) = \frac{e^{-2}2^0}{0!} = e^{-2}$.

(b) The probability there are 5 or more bacteria in a 50cc sample.

Solution: $\mu = \lambda t = \frac{1}{10} \times 50 = 5$ per 50cc. Then

$$\begin{aligned} P(X \geq 5) &= 1 - P(X \leq 4) \\ &= 1 - [P(X = 0) + \cdots + P(X = 4)] \\ &= 1 - e^{-5}\left[\frac{5^0}{0!} + \frac{5^1}{1!} + \cdots + \frac{5^4}{4!}\right]. \end{aligned}$$

Distinguishing Poisson from Binomial and Other Distributions

1. Can we specify in advance the maximum value which X can take?
   If we can, then the distribution is NOT Poisson. If there is no fixed upper limit, then might be Poisson.

2. Does it make sense to ask how often the event did not occur?
   If it does make sense, the distribution is NOT Poisson. If it does not make sence, then might be Poisson.

5.9 Combining Other Models with the Poisson Process

Example. A very large (essentially infinite) number of ladybugs is released in a large orchard. They scatter randomly so that on average a tree has 6 ladybugs on it. Trees are all the same size.

(a) Find the probability a tree has > 3 ladybugs on it.
Solution: Poisson distribution with $\lambda = 6$ and $v = 1$ (that is, any tree has a "volume" of one unit). So $\mu = \lambda v = 6$. Then $P(X > 3) = 1 - P(X \leq 3) = 1 - e^{-6}\left[\frac{6^0}{0!} + \frac{6^1}{1!} + \frac{6^2}{2!} + \frac{6^3}{3!}\right] = 0.8488$.

(b) When 10 trees are picked at random, what is the probability that 8 of these trees have > 3 ladybugs on them?
Solution: Binomial distribution where success means >3 ladybugs on a tree. We have $X \sim \text{Bin}(10, 0.8488)$. Then $P(X = 8) = \binom{10}{8}(0.8488)^8(1 - 0.8488)^2 = 0.2772$.

(c) Trees are checked until 5 with > 3 ladybugs are found. Let X be the total number of trees checked. Find the probability function, $f(x)$.

Solution: $X - 5 \sim \text{NB}(5, 0.8488)$ with $(x-5)$ failures. Then

$$\begin{aligned} f(x) &= \binom{x-5+5-1}{x-5}(0.8488)^5(1-0.8488)^{x-5} \\ &= \binom{x-1}{x-5}(0.8488)^5(0.1512)^{x-5} \\ &= \binom{x-1}{4}(0.8488)^5(0.1512)^{x-5}, x = 5, 6, 7, \ldots \end{aligned}$$

(d) Find the probability a tree with > 3 ladybugs on it has exactly 6.

Solution: Let $A = 6$ ladybugs and $B =$ more than 3 ladybugs.

Then, $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\frac{e^{-6}6^6}{6!}}{0.8488} = 0.1892$.

(e) On 2 trees there are a total of t ladybugs. Find the probability that x of these are on the first tree.

Solution:

$$\begin{aligned} P(x \text{ on first tree} | \text{total of } t) &= \frac{P(x \text{ on first tree} \cap \text{total of } t)}{P(\text{total of } t)} \\ &= \frac{P(x \text{ on first tree} \cap t - x \text{ on second tree})}{P(\text{total of } t)} \\ &= \frac{P(x \text{ on first tree}) \cdot P(t - x \text{ on second tree})}{P(\text{total of } t)} \end{aligned}$$

Use Poisson distribution with $\mu = 6 \times 2 = 12$ in the denominator since there are two trees.

$$\begin{aligned} &= \frac{\left(\frac{e^{-6}6^x}{x!}\right)\left(\frac{e^{-6}6^{t-x}}{(t-x)!}\right)}{\frac{e^{-12}12^t}{t!}} \\ &= \frac{t!}{x!(t-x)!}\left(\frac{6}{12}\right)^x\left(\frac{6}{12}\right)^{t-x} \\ &= \binom{t}{x}\left(\frac{1}{2}\right)^x\left(1-\frac{1}{2}\right)^{t-x} \text{ for } x = 0, 1, \ldots, t. \end{aligned}$$

Remark. We can also let $X = \#$ of ladybugs on the first tree, then $X \sim \text{Bin}(t, 0.5)$. Since the probability of a ladybug on the first tree is just 0.5 as there are two trees.

5.10 Summary of Probability Functions for Discrete Random Variables

Name	Probability Function
Discrete Uniform	$f(x) = \frac{1}{b-a+1}$ for $x = a, a+1, a+2, \ldots, b$ $(b > a)$
Hypergeometric	$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$ for $x = \max(0, n-(N-r)), \ldots, \min(n, r)$
Binomial	$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$ $(0 < p < 1)$
Negative Binomial	$f(x) = \binom{x+k-1}{x} p^k (1-p)^x$ for $x = 0, 1, 2, \ldots$ $(0 < p < 1)$
Geometric	$f(x) = p(1-p)^x$ for $x = 0, 1, 2, \ldots$ $(0 < p < 1)$
Poisson	$f(x) = \frac{e^{-\mu}\mu^x}{x!}$ for $x = 0, 1, 2, \ldots$ $(\mu > 0)$

6 Computational Methods and the Statistical Software R

Not covered material.

7 Expected Value and Variance

7.1 Summarizing Data on Random Variables

Frequency Histogram

- Symmetric: mean = median
- Right-skewed: mean > median
- Left-skewed: mean < median

Definition (**Sample**). A set of observed outcomes $x_1, \dots, x_n$ for a random variable X is a **sample**.

Definition (**Arithmetic Mean or Sample Mean**). The **mean** of n outcomes $x_1, \dots, x_n$ for a randome variable X is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}.$$

Definition (**Median**). The **median** of a sample is a value s.t. half the results are below it and half above it, when the results are arranged in numerical order.

Note. If there are even number of values. The median is the mean of the two middle values.

Definition (**Mode**). The **mode** of the sample is the value which occurs most often.

Note. There can be multiple modes.

7.2 Expectation of a Random Variable

Definition (**Expected Value (Expectation)**). Let X be a discrete random variable with range $(X) = A$ and probability function $f(x)$. The **expected value** of X is given by

$$\mu = \mathbb{E}[X] = \sum_{x \in A} x f(x).$$

Theorem (**Expected Value of** $g(X)$).
Let X be a discrete random variable with range $(X) = A$ and probability function $f(x)$. The **expected value** of some function $g(X)$ of X is given by

$$\mathbb{E}[g(X)] = \sum_{x \in A} g(x) f(x).$$

Note.

1. Interpret $\mathbb{E}[g(X)]$ as the average value of $g(X)$ in an infinite series of repetitions of the process where X is defined.
2. $\mathbb{E}[g(X)]$ may be a value that $g(X)$ never takes.
3. $\mathbb{E}[X]$ is NOT a random variable like X but a non-random constant.
4. Suppose X takes values from 1 to 10. Then $\mathbb{E}[X]$ cannot exceed 10 or smaller than 1.

Theorem (**Linearity Properties of Expectation**).

1. For constants a and b,
$$\mathbb{E}[ag(X) + b] = a\mathbb{E}[g(X)] + b.$$
2. For constants a, b and two functions g_1, g_2,
$$\mathbb{E}[ag_1(X) + bg_2(X)] = a\mathbb{E}[g_1(X)] + b\mathbb{E}[g_2(X)].$$

Note. For constant a, we have $\mathbb{E}[a] = a$.

Example. Suppose we have the random variable X s.t. $f_X(x) = \frac{x}{10}, x = 1, 2, 3, 4$. Find $\mathbb{E}[X(5-X)]$.

Solution:
$\mathbb{E}[X(5-X)] = \mathbb{E}\left[5X - X^2\right] = 5\mathbb{E}[X] - \mathbb{E}\left[X^2\right] = 5\left[(1 \times \frac{1}{10}) + \cdots + (4 \times \frac{4}{10})\right] - \left[(1^2 \times \frac{1}{10}) + \cdots + (4^2 \times \frac{4}{10})\right]$
$= 5 \times 3 - 10 = 5.$

7.3 Some Applications of Expectation

Example. A local television station sells 15sec, 30sec, and 60sec advertising spots. Let X denote the length of a randomly selected commercial appearing on this station, and suppose that the probability distribution of X is given by

x	15	30	60
$f(x)$	0.1	0.3	0.6

(a) Find $\mathbb{E}[X]$.
Solution: $\mathbb{E}[X] = \sum_{\text{all } x} xf(x) = (15 \times 0.1) + (30 \times 0.3) + (60 \times 0.6) = 46.5$ seconds.

(b) If a 15s spot sells for \$500, a 30s spot for \$800, and a 60s spot for \$1000, find the average amount paid for commercials appearing on this station.
Solution: $\mathbb{E}[Y] = (500 \times 0.1) + (800 \times 0.3) + (1000 \times 0.6) = \$890$.

7.4 Means and Variances of Distributions

Definition (Expectation for Probability Models).

1. Binomial: $\mathbb{E}[X] = np$.
2. Poisson: $\mathbb{E}[X] = \lambda t = \mu$.
3. Discrete Uniform: $\mathbb{E}[X] = \frac{a+b}{2}$.
4. Hypergeometric: $\mathbb{E}[X] = \frac{nr}{N}$.
5. Negative Binomial: $\mathbb{E}[X] = \frac{k(1-p)}{p}$.
6. Geometric: $\mathbb{E}[X] = \frac{1-p}{p}$.

Definition (Variance). The **variance** of a random variable X, denoted by $\mathrm{Var}(X)$ or σ^2, is

$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right].$$

Note. Variance is the average square of the distance from the mean (units^2).

The definition of variance is not efficient to use for calculation of $\mathrm{Var}(X)$, whereas the following two results are often useful:

Theorem.

1. $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \left[\mathbb{E}[X]\right]^2 = \mathbb{E}\left[X^2\right] - \mu^2$
2. $\mathrm{Var}(X) = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \left[\mathbb{E}[X]\right]^2 = \mathbb{E}[X(X-1)] + \mu - \mu^2$

Definition (Standard Deviation). The **standard deviation** of a random variable X is

$$\sigma = \mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\mathbb{E}[(X-\mu)^2]}$$

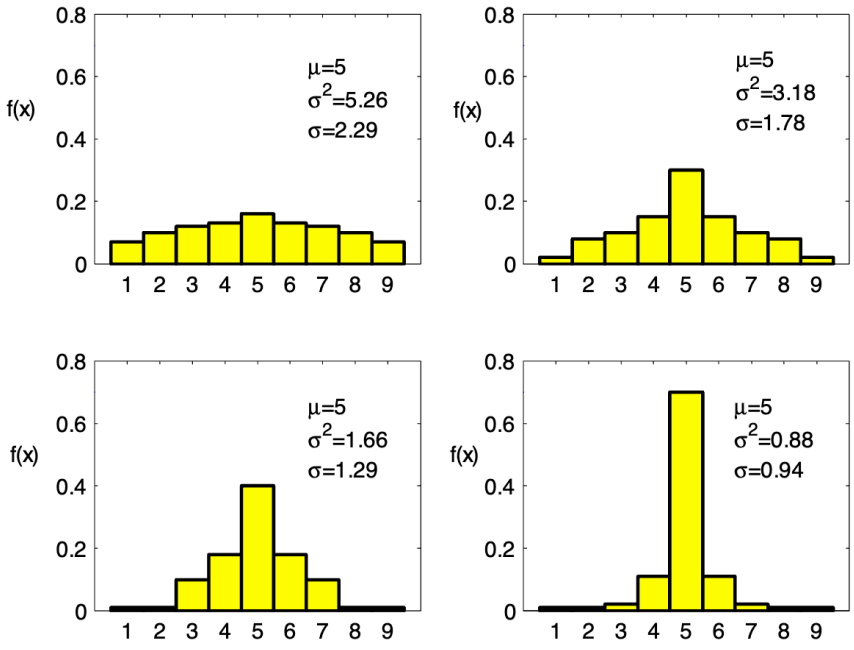


Figure 3: How $\mathrm{Var}(X)$ or $\mathrm{sd}(X)$ reflects the spread of a probability histogram.

Definition (**Variance for Probability Models**).

1. Binomial: $\operatorname{Var}(X) = np(1-p)$.
2. Poisson: $\operatorname{Var}(X) = \mu$.
3. Discrete Uniform: $\operatorname{Var}(X) = \dfrac{(b-a+1)^2 - 1}{12}$.
4. Hypergeometric: $\operatorname{Var}(X) = \dfrac{nr}{N}\left(1 - \dfrac{r}{N}\right)\dfrac{N-n}{N-1}$.
5. Negative Binomial: $\operatorname{Var}(X) = \dfrac{k(1-p)}{p^2}$.
6. Geometric: $\operatorname{Var}(X) = \dfrac{1-p}{p^2}$.

Theorem (**Properties of Mean and Variance**).

If a and b are constants and $Y = aX + b$, then

$$\mu_Y = \mathbb{E}[Y] = a\mathbb{E}[X] + b = a\mu_X + b$$

and

$$\sigma_Y^2 = \operatorname{Var}(Y) = a^2 \operatorname{Var}(X) = a^2\sigma_X^2,$$

where $\mu_X = \mathbb{E}[X]$, $\sigma_X^2 = \operatorname{Var}(X)$, $\mathbb{E}[Y] = \mu_Y$, and $\operatorname{Var}(Y) = \sigma_Y^2$.

Note. For constant a, we have $\operatorname{Var}(a) = 0$.

8 Continous Random Variables

8.1 General Terminology and Notation

Continuous random variables take values on some interval of real numbers. We have $P(X = x) = 0$ for each x, because $P(X = a) = \int_a^a f(x)\,dx = 0$.

We use the cumulative distribution function $F(x)$ and the probability density function $f(x)$ to describe a continuous random variable.

Definition (**Cumulative Distribution Function (c.d.f.)**).

$$F(x) = P(X \leq x).$$

Theorem (**Properties of a c.d.f.**).

1. $F(x)$ is defined $\forall x \in \mathbb{R}$.
2. $F(x)$ is a non-decreasing function of x.
3. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
4. $P(a \leq X \leq b) = F(b) - F(a)$.

Note. Since $P(X = x) = 0$ for each x, we have

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = F(b) - F(a).$$

Definition (**Probability Density Function (p.d.f.)**).

The **probability density function** $f(x)$ for a continous random variable X is the derivative

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x},$$

where $F(x)$ is the cumulative distribution function for X.

Note. From the way in which X was generated that $f(x)$ represents the **relative likelihood** of (small intervals around) different x-values (outcomes).

Theorem (**Properties of a p.d.f.**).

Assume that $f(x)$ is a continuous function of x at all points for which $0 < F(x) < 1$.

1. $P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)\,dx$.
2. $f(x) \geq 0$ (since $F(x)$ is non-decreasing, so its derivative should be non-negative).
3. $\int_{-\infty}^{\infty} f(x)\,dx = \int_{\text{all } x} f(x)\,dx = 1$.
4. $F(x) = \int_{-\infty}^{x} f(u)\,du$ (this is property 1 with $a = -\infty$).

Definition (**Quantiles and Percentiles**).

Suppose X is a continuous random variable with cumulative distribution function $F(x)$. The pth quantile of X (or of the distribution) is the value $q(p)$, s.t. $P[X \leq q(p)] = p$.

The value $q(p)$ is also called the 100th percentile of the distribution. If $p = 0.5$, then $m = q(0.5)$ is called the median of X or the median of the distribution.

Remark. The **Median** is the value m such that

$$F(m) = \int_{-\infty}^{m} f(x)\,dx = 0.5 = \int_{m}^{\infty} f(x)\,dx, \text{ which is the 50th percentile.}$$

Example. Let

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{x}{4} & \text{for } 0 < x \leq 4 \\ 1 & \text{for } x > 4 \end{cases}$$

(a) Solve for $f(x)$.

Solution: $f(x) = \frac{\mathrm{d}}{\mathrm{d}x}(F(x)) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x}{4}\right) = \frac{1}{4}$ for $0 < x \leq 4$. Note that outside this interval $f(x)$ is defined to be 0.

(b) Solve for the 90th percentile, i.e. solve for the value of x s.t. the area under the curve to its left is 0.9.

Solution: Note that the $f(x) = \frac{1}{4}$ is a constant function.

We need to solve for x: $F(x) = P(X \leq x) = 0.9$. So, $\frac{1}{4}x = 0.9 \implies x = 3.6$.

Example. Let X be a continuous random variable with p.d.f.

$$f(x) = \begin{cases} c(4x - 2x^2) & \text{for } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find

(a) the constant c.

(b) $F(x)$.

(c) $P(X > 1)$.

Solution:

(a)

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,dx &= 1 \\ \int_0^2 c(4x - 2x^2)\,dx &= 1 \\ c\left[2x^2 - \frac{2x^3}{3}\right]_0^2 &= 1 \\ &\vdots \\ c &= \frac{3}{8} \end{aligned}$$

(b) $F(X) = P(X \le x) = \displaystyle\int_{-\infty}^{x} f(u)\,du = \int_0^x \frac{3}{8}(4u - 2u^2)\,du = \frac{3}{8}\left[2u^2 - \frac{2u^3}{3}\right]_0^2 = \frac{3}{8}\left(2x^2 - \frac{2x^3}{3}\right)$.

Therefore, we have

$$F(x) = \begin{cases} 0 & x \le 0 \\ \frac{3}{8}\left(2x^2 - \frac{2x^3}{3}\right) & 0 < x < 2 \\ 1 & x \ge 2 \end{cases}.$$

(c) $P(X > 1) = \displaystyle\int_1^{\infty} f(x)\,dx = \int_1^2 f(x)\,dx = \cdots = 0.5.$

OR

$P(X > 1) = 1 - P(X \le 1) = 1 - F(1) = \cdots = 0.5.$

Defined Variables or Change of Variables:

Suppose we know the p.d.f. or c.d.f. for a continous random variable X, we sometimes want to find the p.d.f. or c.d.f. for some other random variable Y, a function of X. The procedure is summarized below.

1. Write the c.d.f. of Y as a function of X.
2. Use $F_X(x)$ to find $F_Y(x)$. Then if you want the p.d.f. $f_Y(y)$, you can differentiate $F_Y(x)$.
3. Find the range of values of y.

Example. X is a continuous random variable having

$$f(x) = \frac{1}{4} \text{ for } 0 < x \leq 4 \quad \text{and} \quad F(x) = \frac{x}{4} \text{ for } 0 < x \leq 4.$$

Let $Y = \frac{1}{X}$. Find the p.d.f. of Y.

Solution:

Step 1:

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(\frac{1}{X} \leq y) \\ &= P(X \geq \frac{1}{y}) \\ &= 1 - P(X \leq \frac{1}{y}) \\ &= 1 - F_X(\frac{1}{y}) \end{aligned}$$

Step 2:

$$F_Y(y) = 1 - F_X(\frac{1}{y}) = 1 - \frac{\frac{1}{y}}{4} = 1 - \frac{1}{4y}.$$

We can differentiate $F_Y(y)$ to obtain $f_Y(y)$.

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\left(1 - \frac{1}{4y}\right) = \frac{1}{4y^2} \text{ for } y \geq \frac{1}{4}.$$

Step 3: note that $x \in (0, 4]$, so $y = \frac{1}{x} \in [\frac{1}{4}, \infty)$.

Definition (**Mean, and Variance for Continous Random Variables**).

If X is a continous random variable, then we define

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

Remark (**Mean and Variance**).

- $\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\,dx$, which is the average of the distribution.
- $\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$.

Example. If X is a continuous random variable having p.d.f

$$f(x) = \frac{1}{4} \quad \text{for } 0 < x \leq 4.$$

Find $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

Solution: $\mathbb{E}[X] = 0 + \int_0^4 x\left(\frac{1}{4}\right)dx + 0 = 2$, and $\mathbb{E}\left[X^2\right] = 0 + \int_0^4 x^2\left(\frac{1}{4}\right)dx + 0 = \frac{16}{3}$.

Finally, we have $\mathrm{Var}(X) = \frac{16}{3} - 2^2 = \frac{4}{3}$.

Example (**Exercise!**). Let X be a random variable with p.d.f. given by

$$f(x) = \begin{cases} k\sqrt{x} & \text{for } 0 \leq x \leq 1 \\ \frac{k}{x^4} & \text{for } x > 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the constant k. **Ans:** $k = 1$

(b) Find the c.d.f. $F(x)$ for all values of x. **Ans:** $F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{2}{3}x^{3/2} & 0 \leq x \leq 1 \\ 1 - \frac{1}{3x^3} & x > 1 \end{cases}$

(c) Find $P(\frac{1}{3} < X < 3)$. **Ans:** 0.8594

(d) Calculate $\mathbb{E}[X]$ and $\mathrm{Var}(X)$. **Ans:** $\mathbb{E}[X] = 0.9$ and $\mathrm{Var}(X) = 0.48$

8.2 Continous Uniform Distribution

Physical setup: X is a continuous random variable takes values in $[a, b]$ (it doesn't matter whether the interval is open or closed) with all subintervals of a fixed length being equally likely.

Then X has a continuous uniform distribution. We write $X \sim \mathrm{U}(a, b)$, for $b > a$ and $a, b \in \mathbb{R}$.

P.D.F. and C.D.F: Since all points are equally likely (intervals contained in $[a, b]$ of a given length all have the same probability), it must be a constant function $f(x) = k$ for $a \leq x \leq b$ and for some constant k. Next, we need $\int_a^b f(x)\,dx = 1 \implies k(b-a) = 1$.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}.$$

It follows that the c.d.f. is

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

- $\mathbb{E}[X] = \int_a^b x\left(\frac{1}{b-a}\right) dx = \frac{a+b}{2}$.
- $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Example. Tansforming a random variable with a different continuous distribution to obtain a uniform distribution.

Suppose X has the continuous p.d.f. $f(x) = 0.1e^{-0.1x}$ for $x > 0$. We will show that the new random variable $Y = e^{-0.1X}$ has a uniform distribution, $\mathrm{U}(0, 1)$.

Solution:

$F_Y(y) = P(Y \leq y) = P(e^{-0.1X} \leq y) = P(X \geq -10 \ln y) = 1 - P(X < -10 \ln y) = 1 - F_X(-10 \ln y)$.

Next, since $x > 0$, we have $F_X(x) = \int_0^x 0.1e^{-0.1u}\,du = 1 - e^{-0.1x}$.

So $F_Y(y) = 1 - \left[1 - e^{-0.1(-10 \ln y)}\right] = y$ for $0 < y < 1$, since the range of X is $(0, \infty)$.

Thus, $f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}(F_Y(y)) = 1$ for $0 < y < 1$, which implies $Y \sim \mathrm{U}(0, 1)$.

8.3 Exponential Distribution

Physical sestup: In a Poisson Process for events in time let $X =$ length of time we wait until the first occurrence. Then X has an exponential distribution.

Note. Recall that the # of occurrences in a fixed time has a Poisson distribution.

Example. If phone calls to a fire station follows a Poisson process, then the length of time between phone calls follows an exponential distribution.

P.D.F. and C.D.F:

$$\begin{aligned} F(X) = P(X \leq x) &= P(\text{time to 1st occurrence} \leq x) \\ &= 1 - P(\text{time to 1st occurrence} > x) \\ &= 1 - P(0 \text{ occurrences in } (0, x)) \\ &= 1 - \frac{e^{-\lambda x}(\lambda x)^0}{0!} \\ &= 1 - e^{-\lambda x} \quad \text{for } x > 0. \end{aligned}$$

Note that the number of occurrences has a Poisson distribution with $\mu = \lambda x$, where λ is the average rate of occurrence.

Then,

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}(F(X)) = \lambda e^{-\lambda x} \quad \text{for } x > 0.$$

Alternate Form: Let $\theta = \frac{1}{\lambda}$, so $\theta =$ average waiting time to an occurrence. We have

$$F(x) = \begin{cases} 1 - e^{\frac{-x}{\theta}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

$$f(x) = \begin{cases} \frac{1}{\theta} e^{\frac{-x}{\theta}} & x > 0 \\ 0 & x \leq 0 \end{cases}.$$

where $\theta > 0$. We write $X \sim \text{Exp}(\theta)$.

Example. If $\lambda = 5$ occurrences per hour, then $\theta = \frac{1}{5}$ means: have to wait an average of $\frac{1}{5}$th of an hour to see an occurrence.

To find the mean and variance, we can use the properties of the Gamma Function.

Definition (**The Gamma Function**).

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$

is called the gamma function of α, where $\alpha > 0$.

Properties of the gamma function:

1. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for $\alpha > 1$.
2. $\Gamma(\alpha) = (\alpha - 1)!$ if α is a positive integer.
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
   (This can be proved using double integration.)

Goind back to the exponential distribution, we have

- $\mathbb{E}[X] = \theta = \dfrac{1}{\lambda}$.
- $\mathrm{Var}(X) = \theta^2 = \dfrac{1}{\lambda^2}$.

Remark. Leave proofs as exercises, use the Gamma function.

Example. The average amount of time in hours that a computer survives before breaking down is 100 hours. What is the probability that

(a) A computer will function between 50 and 150 hours before breaking down?

(b) It will function for fewer than 100 hours?

(c) If a computer survives more than 100 hours, what is the probability it survives an additional 50 hours?

Solution:

(a) Let $X =$ the length of time waited until 1st breakdown. Then $X \sim \text{Exp}(\theta = 100 \text{ hours})$.

$$\begin{aligned} P(50 \leq X \leq 150) = P(50 < x < 150) &= P(X \leq 150) - P(X \leq 50) \\ &= F(150) - F(50) \\ &= \left(1 - e^{\frac{-150}{100}}\right) - \left(1 - e^{\frac{-50}{100}}\right) \\ &= e^{\frac{-50}{100}} - e^{\frac{-150}{100}} \end{aligned}$$

(b) $P(X < 100) = F(100) = 1 - e^{\frac{-100}{100}} = 1 - \frac{1}{e} = 0.6321.$

(c)

$$\begin{aligned} P(X > 100 + 50 | X > 100) &= \frac{P(X > 150 \cap X > 100)}{P(X > 100)} \\ &= \frac{P(X > 150)}{P(X > 100)} \\ &= \frac{1 - P(X \leq 150)}{1 - P(X \leq 100)} \\ &= \frac{1 - \left(1 - e^{\frac{-150}{100}}\right)}{1 - \left(1 - e^{\frac{-100}{100}}\right)} \\ &= e^{\frac{-50}{100}} \; (= P(X > 50)) \end{aligned}$$

Remark. Part (c) illustrates the "**memoryless property**" of the Exponential distribution:

$$P(X > c + b | X > b) = P(X > c).$$

Given that you have waited b units of time for the next event, the probability you wait an additional c units of time **does not** depend on b but only depends on c.

Example (**Exercise**). Suppose that the length of a phone call in minutes is an exponential random variable with parameter $\lambda = \frac{1}{10}$. If someone arrives immediately ahead of you at a public telephone booth, find the probability that you will have to wait

(a) More than 10 mins. **Ans:** e^{-1}

(b) Between 10 and 20 mins. **Ans:** $e^{-1} - e^{-2} = 0.233$

(c) If you have been waiting for more than 10 mins, what is the probability you have to wait for more than an additional 5 mins? **Ans:** $e^{-\frac{1}{2}}$

8.4 A Method for Computer Generation of Random Variables

Not covered material.

8.5 Normal Distribution

Physical setup: A random variable X defined on $(-\infty, \infty)$ has a Normal distribution if it has p.d.f. of the form

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are parameters.

- $\mathbb{E}[X] = \mu$.
- $\text{Var}(X) = \sigma^2$.

We write $X \sim \text{N}(\mu, \sigma^2)$.

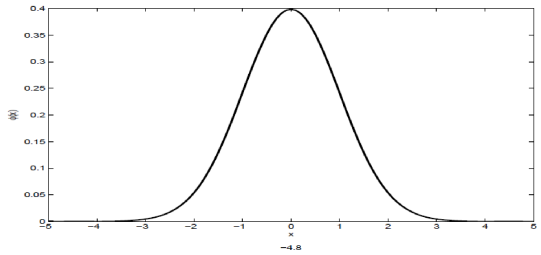


Figure 4: The Standard Normal $N(0,1)$ p.d.f.

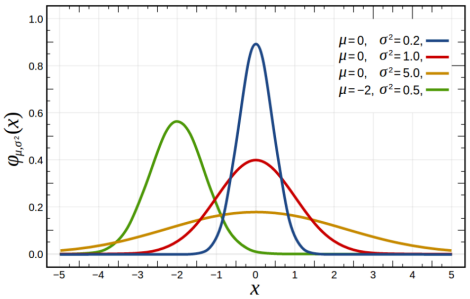


Figure 5: Some p.d.f.s

Note. For the first figure, we have $\mu = 0$ and $\sigma^2 = 1$. Normal distributions are symmetric about the mean (the line $x = \mu$), so 50% on each side.

Effects of the Mean and the Variance

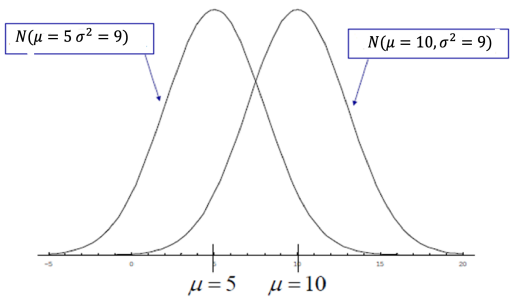


Figure 6: Effect of μ.

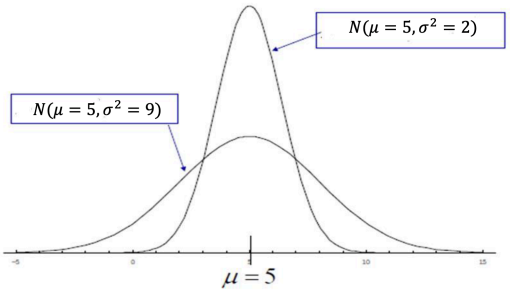


Figure 7: Effect of σ^2.

Note.

- The mean shifts the distribution horizontally (shifts to the right as μ increases).
- The variance stretches the distribution (gets thicker and lower as σ^2 increases).

Example. Heights or weights of males (or females) in large populations tend to follow a Normal distribution.

The C.D.F. of $N(\mu, \sigma^2)$ is

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \quad x \in \mathbb{R}.$$

Note. Numerical methods are used to compute its value.

$N(0,1)$ is called the "**standard**" Normal distribution with $\mu = 0$ and $\sigma = 1$. We have a "new" random variable $Z = \dfrac{X-\mu}{\sigma}$, which is distributed as $Z \sim \mathrm{N}(0,1)$.

Theorem. Let $X \sim \mathrm{N}(\mu, \sigma^2)$ and defined $Z = \dfrac{X-\mu}{\sigma}$. Then $Z \sim \mathrm{N}(0,1)$ and

$$P(X \le x) = P\left(Z \le \frac{X-\mu}{\sigma}\right).$$

Note. If we can compute the c.d.f. for $N(0,1)$, then we can compute it for other $N(\mu, \sigma^2)$ as well.

What is this transformation doing?

Consider the distribution of heights of young women aged 18 to 24. The distribution is approximately Normal with mean 64.5 inches and standard deviation 2.5 inches. We have $X \sim \mathrm{N}(64.5, 2.5^2)$.

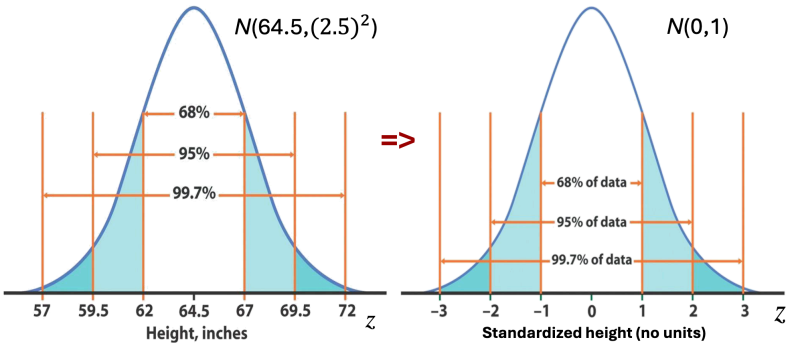


Figure 8: Transform to the standard Normal distribution.

Theorem (**Other Properties of Normal Distribution**).

If $X \sim \mathrm{N}(\mu, \sigma^2)$, then we have

- Symmetric about the mean:

$$P(X \leq \mu - t) = P(X \geq \mu + t)$$

or

$$P(X \leq \mu + t) = P(X \geq \mu - t).$$

- Density is unimodal: peak is at μ. Furthermore, mean and median also at $x = \mu$.

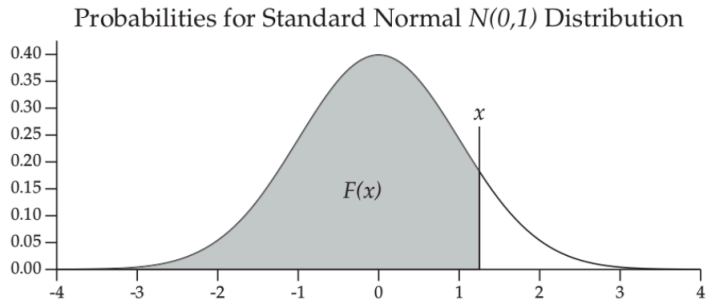
This table gives values of $F(x) = P(X \leq x)$ for $X \sim \mathrm{N}(0,1)$ and $x \geq 0$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

Figure 9: Standard Normal Distribution c.d.f. values.

How do we use the c.d.f. table?

$$F(x) = P(X \leq x), \text{ so e.g. } P(X \leq 0.53) = 0.70194:$$



This table gives the values of $F(x)$ for $x \geq 0$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56750	0.57142	0.57534
0.2	0.57926	0.58317	0.58706	0.59095	0.59484	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240

Figure 10: Example of using the c.d.f. table.

Example. Find a number b such that $P(|Z| \leq b) = 0.95$.

Solution: Note that $P(|Z| \leq b) = P(-b \leq Z \leq b) = 0.95$. By symmetry, the probability outside $(-b, b)$ must be 0.05.

$\implies P(Z \leq -b) + P(Z \geq b) = 0.05$. Furthermore, $P(Z \leq -b) = P(Z \geq b) = 0.025$ by symmetry again.

$\implies P(Z \leq b) = 0.025 + 0.95 = 0.975$. Looking for the value in the table, we get $b = 1.96$.

Remark. Drawing out the normal distribution helps a lot!!!!!

Gaussian Distribution: Another name for the Normal distribution. $X \sim \mathrm{G}(\mu, \sigma)$ means that X has Gaussian distribution with mean μ and standard deviation σ.

Example. We can write $X \sim \mathrm{N}(1, 4)$ as $X \sim \mathrm{G}(1, 2)$.

Theorem (**Standardization**).

$$P(X \le x) = P(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}) = P(Z \le \frac{x-\mu}{\sigma}),$$

where $Z \sim \mathrm{N}(0,1)$.

Remark. Standardization can also be used for calculating other types of probabilities:

- $P(X > x) = P(Z > \frac{x-\mu}{\sigma})$.
- $P(a < X < b) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) = P(Z < \frac{b-\mu}{\sigma}) - P(Z < \frac{a-\mu}{\sigma})$.

Example. Suppose $X \sim \mathrm{N}(10, 2)$, find $P(|X-10| \le 3)$.

Solution:

$$\begin{aligned}
P(|X-10| \le 3) &= P(-3 \le X-10 \le 3) \\
&= P(7 \le X \le 13) \\
&= P(\frac{7-10}{\sqrt{2}} \le \frac{X-\mu}{\sigma} \le \frac{13-10}{\sqrt{2}}) \quad \text{(standardize)} \\
&= P(-2.12 \le Z \le 2.12) \\
&= P(Z \le 2.12) - P(Z \le -2.12) \\
&= P(Z \le 2.12) - P(Z \ge 2.12) \\
&= P(Z \le 2.12) - [1 - P(Z \le 2.12)] \\
&= 2P(Z \le 2.12) - 1 \\
&= (2 \times 0.983) - 1 = 0.966.
\end{aligned}$$

Example (**Exercise**). Suppose $X \sim \mathrm{N}(-7, 14)$, find $P(|X+7| \ge 8)$.
Ans: $2 - 2P(Z \le 2.14) = 0.03236$.

Example. Suppose a certain mechanical component produced by a company has a width that is normally distributed with a mean $\mu = 2600$ and a standard deviation $\sigma = 0.6$.

(a) What proportion of the components have a width outside the range 2599 to 2601?

(b) If the company needs to be able to guarantee to its purchaser that no more than 1 in 1000 of the components have a width outside the range 2599 to 2601, by how much does the value of σ need to be reduced?

Solution:

(a) Let $X =$ the width of a component. Then $X \sim \mathrm{N}(2600, 0.6^2)$.

$$\begin{aligned} P(X > 2601) + P(X < 2599) &= P(Z > \frac{2601 - 2600}{0.6}) + P(Z < \frac{2599 - 2600}{0.6}) \\ &= P(Z > 1.67) + P(Z < -1.67) \quad \text{(standardize)} \\ &= [1 - P(Z \leq 1.67)] + [1 - P(Z \leq 1.67)] \\ &= 2 - 2P(Z \leq 1.67) \\ &= 2 - (2 \times 0.95254) = 0.095. \end{aligned}$$

(b)

$$\begin{aligned} P(X > 2601) + P(X < 2699) &\leq 0.001 \\ 1 - P(2599 < X < 2601) &\leq 0.001 \\ P(2599 < X < 2601) &\geq 0.999 \\ P(\frac{2599 - 2600}{\sigma} < Z < \frac{2601 - 2600}{\sigma}) &\geq 0.999 \\ P(-\frac{1}{\sigma} < Z < \frac{1}{\sigma}) &\geq 0.999 \end{aligned}$$

Using a sketch of the normal distribution, the probability within $(-\frac{1}{\sigma}, \frac{1}{\sigma}) \geq 0.999$. This means that we have probability ≤ 0.0005 at each side of the distribution.
$\implies P(Z \leq \frac{1}{\sigma}) \geq 0.995$ So, $\frac{1}{\sigma} = 3.29$.

Therefore, we must have $\sigma \leq 0.30395$, need to reduce σ by about 0.29605.

N(0,1) Quantiles: This table gives values of $F^{-1}(p)$ for $p \geq 0.5$

p	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.075	0.08	0.09	0.095
0.5	0.0000	0.0251	0.0502	0.0753	0.1004	0.1257	0.1510	0.1764	0.1891	0.2019	0.2275	0.2404
0.6	0.2533	0.2793	0.3055	0.3319	0.3585	0.3853	0.4125	0.4399	0.4538	0.4677	0.4959	0.5101
0.7	0.5244	0.5534	0.5828	0.6128	0.6433	0.6745	0.7063	0.7388	0.7554	0.7722	0.8064	0.8239
0.8	0.8416	0.8779	0.9154	0.9542	0.9945	1.0364	1.0803	1.1264	1.1503	1.1750	1.2265	1.2536
0.9	1.2816	1.3408	1.4051	1.4758	1.5548	1.6449	1.7507	1.8808	1.9600	2.0537	2.3263	2.5758

Figure 11: Standard Normal Distribution Quantiles.

Example. Let $X \sim \mathrm{N}(0.28, 0.05^2)$. Find the 90th quantile.

Solution: $F^{-1}(0.9) = 1.2816 = Z$. Since $Z = \dfrac{X - 0.28}{0.05}$, we have $X = 0.344$.

9 Multivariate Distributions

9.1 Basic Terminology and Techniques

Definition (**Joint Probability Function**).

Let X and Y be two discrete random variables, we define the **joint probability function** $f(x,y)$ of (X,Y) as

$$\begin{aligned} f(x,y) &= P(X = x \text{ and } Y = y) \\ &= P(X = x, Y = y) \end{aligned}$$

Remark. In general, if there are n discrete random variables $X_1, X_2, \dots, X_n$, we have

$$f(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n).$$

Theorem (**Properties of Joint Probability Function**).

- $f(x,y) > 0 \ \forall (x,y)$.
- $\sum_{\text{all } (x,y)} f(x,y) = 1$.

Example. Let X be the number of daily purchases of a luxury item from a factory outlet location and Y be the daily number of purchases made online. Let 1,2, and 3 denote the number of purchases less than five, at least five but less than 15, and 15 or more respectively. The joint pf of X and Y is

		x			
$f(x,y)$		1	2	3	
	1	0.09	0.12	0.13	
y	2	0.12	0.11	0.11	
	3	0.13	0.10	0.09	
					1

Find $f(1,2)$ and $f(2,2)$.

Solution: $f(1,2) = P(X = 1, Y = 2) = 0.12$ and $f(2,2) = P(X = 2, Y = 2) = 0.11$.

What if we are only interested in one of the variables?

Example (continued)**.** Say, we are only interested in X. Then,

$$\begin{aligned} P(X=1) = P(B) &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) \\ &= P(X=1, Y=1) + P(X=1, Y=2) + P(X=1, Y=3) \\ &= 0.34. \end{aligned}$$

Similarly, $P(X=2) = 0.33$ and $P(X=3) = 0.33$.

Note.

- $\sum\limits_{\text{all } x} f_X(x) = \sum\limits_{\text{all } y} f_Y(y) = \sum\limits_{\text{all } (x,y)} f_{X,Y}(x,y) = 1.$

Definition (**Marginal Distributions**)**.**

Given the joint probability function of X and Y, the **Marginal distributions** are give by

- $f_1(x) = f_X(x) = \sum\limits_{\text{all } y} f(x,y)$, with x fixed.
- $f_2(y) = f_Y(y) = \sum\limits_{\text{all } x} f(x,y)$, with y fixed.

Remark. This idea can be extended beyond two variables.

Definition (**Independent Random Variables**)**.**

X and Y are **independent** random variables $\iff f(x,y) = f_1(x)f_2(y), \forall\, (x,y)$.

In general, $X_1, X_2, \ldots, X_n$ are independent variables

$$\iff f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n) \quad \forall x_1, x_2, \ldots, x_n.$$

Remark. You can only conclude that X and Y are independent after checking ALL (x,y) combinations.

Definition (**Conditional Probability Function**).

The conditional probability function of X given $Y = y$ is

$$f_1(x|y) = \frac{f(x,y)}{f_2(y)} \quad \text{provided } f_2(y) > 0.$$

Similarly, the conditional probability function of Y given $X = x$ is

$$f_2(y|x) = \frac{f(x,y)}{f_1(x)} \quad \text{provided } f_1(x) > 0.$$

Note.

- $f_1(x) = f_X(x)$ and $f_2(y) = f_Y(y)$.
- $\sum_{\text{all } x} f(x|y) = 1$.

Example. Continuing with our example:

		x			
$f(x,y)$		1	2	3	$f_Y(y)$
	1	0.09	0.12	0.13	0.34
y	2	0.12	0.11	0.11	0.34
	3	0.13	0.10	0.09	0.32
$f_X(x)$		0.34	0.33	0.33	1

Find the conditional probability function of X given $Y = 1$, i.e find $f(x|1)$.

Solution: Since $f(x|1) = \frac{f(x,1)}{f_Y(1)}$, we have

x	1	2	3	Total
$f(x\|1)$	$\frac{0.09}{0.34} = 0.26$	$\frac{0.12}{0.34} = 0.35$	$\frac{0.13}{0.34} = 0.38$	1

Remark. If X and Y are independent, then $f(x|y) = \frac{f_1(x)f_2(y)}{f_2(y)} = f_1(x)$.

Functions of Random Variables

Example. Let $U = Y - X$, where X and Y have the joint probability function given below. We might now be interested in finding the probability function of U, which is a function of X and Y.

		x			
$f(x,y)$		1	2	3	$f_Y(y)$
	1	0.09	0.12	0.13	0.34
y	2	0.12	0.11	0.11	0.34
	3	0.13	0.10	0.09	0.32
$f_X(x)$		0.34	0.33	0.33	1

The possible values of U are seen by looking at the value of $u = y - x$ for each (x, y) in the range of (X, Y).

		x		
u		1	2	3
	1	0	−1	−2
y	2	1	0	−1
	3	2	1	0

$P(U = 0) = P(X = 1, Y = 1) + P(X = 2, Y = 2) + P(X = 3, Y = 3) = 0.09 + 0.11 + 0.09 = 0.29$. Similarly for other values of U.

Let $T = X + Y$. Notice that to find $P(T = t)$, we are simply adding the probabilities for all (x, y) combinations such that $x + y = t$. This could be written as:

$$f_T(t) = P(T = t) = \sum_{\substack{\text{all } (x,y) \\ \text{with } x+y=t}} f(x, y).$$

However, if $x + y = t$, then $y = t - x$, so we have:

$$f_T(t) = P(T = t) = \sum_{\text{all } x} f(x, y) = \sum_{\text{all } x} P(X = x, Y = t - x).$$

In general, to find the probability function for $U = g(X, Y)$ of two random variables X and Y, we have

$$f_U(u) = P(U = u) = \sum_{\substack{\text{all } (x,y) \\ \text{with } g(x,y)=u}} f(x, y).$$

This can also be extended to functions beyond two random variables.

Theorem.
If $X \sim \text{Poisson}(\mu_1)$ and $Y \sim \text{Poisson}(\mu_2)$ are independent, then $T = X+Y \sim \text{Poisson}(\mu_1+\mu_2)$.

Theorem.
If $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$ are independent, then
$T = X + Y \sim \text{Binomial}(n + m, p)$.

Example. In an auto parts company, an average of μ defective parts are produced per shift. The number, X, of defective parts produced has a Poisson distribution. An inspector checks all parts prior to shipping them, but there is a 10% chance that a defective part will slip by undetected.

Let Y be the number of defective parts the inspector finds on a shift. Find $f(x|y)$.

i.e. The company wants to know how many defective parts are produced, but can only know the number which are actually detected.

Solution: The author was also studying for Actsc 231, so did not have enough time to bother writing out the entire solution. But if you work really hard, you'll have a chance to get:

$$f(x|y) = \frac{e^{-\mu(1-p)}(\mu(1-p))^{x-y}}{(x-y)!} \quad \text{for x = y, y+1, ...}$$

9.2 Mutinomial Distribution

This is a generalization of the Binomial to the case where each trial has k possible outcomes.

Physical Setup: Similar to Binomial, except that now we have **k** types of outcomes. The experiment is repeated **independently n** times whith **k distinct outcomes** each time.

Let the probabilities of these **k** types be $\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_k}$ each time. Let $\mathbf{X_1}$ = # of times the 1st type occurs, ... , $\mathbf{X_k}$ = # of times the k^{th} type occurs. Then $(X_1, X_2, \ldots, X_k)$ has a Multinomial distribution.

Note.

1. $p_1 + p_2 + \cdots + p_k = 1$.
2. $X_1 + X_2 + \cdots + X_k = n$.
   If we wish to drop one of the variables (say X_k), we note that
   $$X_k = n - X_1 - X_2 - \cdots - X_{k-1}.$$

Example. A certain city has 3 television stations. During prime time on Saturday nights, Channel 12 has 50% of the viewing audience, Channel 10 has 30% and Channel 3 has 20%.

Let X_1 = # of families among n watching channel 12, X_2 = ... channel 10, and X_3 = ... channel 3. Then $X_1, X_2, X_3 \sim \text{Multi}(n, p_1 = 0.5, p_2 = 0.3, p_3 = 0.2)$.

Joint Probability Function:

There are $\binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{n-x_1-\cdots-x_{k-1}}{x_k} = \frac{n!}{x_1!x_2!\cdots x_k!}$ number of ways to arrange x_1 items

of the 1st type, x_2 items of the 2nd type, ..., x_k items of the k^{th} type with a total of n trials.

Each of these arrangements has probability $p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$ since p_1 is multiplied x_1 times in some order, etc, and trials are independent. Therefore,

$$f(x_1, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

with $x_i = 0, 1, \ldots, n$ and $\sum_{i=1}^{k} x_i = n$. Note that $\sum f(x_1, \ldots, x_k) = 1$.

Marginal and Joint Probability Functions:

If we are interested in finding the marginal distribution of one variable, X_2, in the multinomial distribution, we can:

1. Mathematical approach: fix the value x_2 and then sum over all the other variables:

$$f_2(x_2) = \sum_{\text{all } x_1, x_3, \ldots, x_k} f(x_1, \ldots, x_k)$$

   for each $x_2 = 0, 1, \ldots, n$. This is algebraically hard.

2. Intuitive and simple approach: if we are only interested in X_2 (i.e. the # of occurrences of the second type among n trials), we notice that

   - The experiment is repreated n times and $X_2 = 0, 1, \ldots, n$.
   - The probability that this second event occurs on each trial is p_2 and the probability it doesn't occur is $1 - p_2$.
   - Each trial is assumed independent.
   - Hence, $X_2 \sim \text{Binomial}(n, p_2)$ and $p_2 = 1 - (p_1 + p_3 + \cdots + p_k)$.

What about $T = X_1 + X_2$?
Using a similar argument, we could make our "success" = an occurrence in one of the first two types, whereas anything else is considered a "failure".

Let $T =$ # of type 1 or type 2 outcomes among n trials. Then $T \sim \text{Binomial}(n, p_1 + p_2)$.

- Two outcomes: type 1 & 2 vs. anything else.
- Independent trials.
- Multiple trials: n.
- Same $P(\text{success}) = P(\text{type 1} \cup \text{type 2}) = P(\text{type 1}) + P(\text{type 2}) = p_1 + p_2$.

What about the conditional distribution of X_1 given $T = t$?
$X_1|T = t \sim \text{Bin}(t, \frac{p_1}{p_1+p_2})$. So $f(x_1|t) = \binom{t}{x_1}\left(\frac{p_1}{p_1+p_2}\right)^{x_1}\left(1 - \frac{p_1}{p_1+p_2}\right)^{t-x_1}$.

Example. The probabilities that a certain electronic component will last less than 50 hours in continuous use, between 50 and 90 hours, or more than 90 hours, are $p_1 = 0.2$, $p_2 = 0.5$, and $p_3 = 0.3$, respectively.

The time to failure of eight such components is recorded.
$X_1 =$ # of components among 8 that fail in < 50 hours.
$X_2 =$ # of components among 8 that fail in $[50, 90]$ hours.
$X_3 =$ # of components among 8 that fail in > 90 hours.

(a) What is the probability that one will last less than 50 hours, five will last between 50 and 90 hours, and two will last more than 90 hours?

(b) What is the probability that at least 3 will last between 50 and 90 hours?

(c) What is the joint probability function of the number of components that last less than 50 hours and the number of components that last between 50 and 90 hours?

Solution:

(a) Mutinomial with $n = 8$. $f(1,5,2) = P(X_1 = 1, X_2 = 5, X_3 = 2) = \frac{8!}{1!5!2!}(0.2)^1(0.5)^5(0.3)^2$.

(b) $X_2 \sim \text{Bin}(8, 0.5)$.

$$\begin{aligned} P(X_2 \geq 3) &= 1 - P(X_2 \leq 2) \\ &= 1 - [P(X_2 = 0) + P(X_2 = 1) + P(X_2 = 2)] \\ &= 1 - \left[\binom{8}{0}(0.5)^0(0.5)^8 + \binom{8}{1}(0.5)^1(0.5)^7 + \binom{8}{2}(0.5)^2(0.5)^6\right] \end{aligned}$$

(c) $X_1, X_2 \sim \text{Multinomial}(n = 8, p_1 = 0.2, p_2 = 0.5, 1 - p_1 - p_2 = 0.3)$.
Therefore, $f(x_1, x_2) = \frac{8!}{x_1!x_2!(8-x_1-x_2)!}0.2^{x_1}0.5^{x_2x}0.3^{8-x_1-x_2}$.

9.3 Markov Chains

Not covered material.

9.4 Expectation for Multivariate Distributions: Covariance and Correlation

Extending the definition of expected value to multiple discrete random variables.

Definition.

$$\mathbb{E}[g(X,Y)] = \sum_{\text{all } (x,y)} g(x,y)f(x,y)$$

and

$$\mathbb{E}[g(X_1,X_2,\dots,X_n)] = \sum_{\text{all } (x_1,x_2,\dots,x_n)} g(x_1,x_2,\dots,x_n)f(x_1,x_2,\dots,x_n).$$

Example. Let the joint probability function $f(x,y)$ be given be the following table:

		x			
$f(x,y)$		0	1	2	$f_Y(y)$
y	1	0.1	0.2	0.3	0.6
	2	0.2	0.1	0.1	0.4
$f_X(x)$		0.3	0.3	0.4	1

Find $\mathbb{E}[XY]$ and $\mathbb{E}[X]$.

Solution: Here we have $g(X,Y) = XY$.

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{\text{all } (x,y)} xyf(x,y)\\ &= (0\times 1)(0.1) + (1\times 1)(0.2) + (2\times 1)(0.3) + (0\times 2)(0.2) + (1\times 2)(0.1) + (2\times 2)(0.1)\\ &= 1.4.\end{aligned}$$

For $\mathbb{E}[X]$, we keep x fixed and use $f_1(x)$:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=0}^{2} x f_1(x) \\ &= (0 \times 0.3) + (1 \times 0.3) + (2 \times 0.4) \\ &= 1.1.\end{aligned}$$

Theorem (**Property of Multivariate Expectation**).

$$\mathbb{E}[ag_1(X,Y) + bg_2(X,Y)] = a\mathbb{E}[g_1(X,Y)] + b\mathbb{E}[g_2(X,Y)].$$

(This can be extended beyond two functions and beyond two variables).

Relationships between Variables

Definition (**Covariance**). The **covariance** of X and Y, denoted by $\mathrm{Cov}(X,Y)$ or σ_{XY}, is

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)].$$

Note.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathbb{E}[XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mu_X\mathbb{E}[Y] - \mu_Y\mathbb{E}[X] + \mu_X\mu_Y \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

and we usually use $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ for calculations.

Example. Let the joint probability function $f(x,y)$ be given be the following table:

		x			
$f(x,y)$		0	1	2	$f_Y(y)$
y	1	0.1	0.2	0.3	0.6
	2	0.2	0.1	0.1	0.4
$f_X(x)$		0.3	0.3	0.4	1

Find $\mathrm{Cov}(X, Y)$.

Solution: From above example, we know that $\mathbb{E}[XY] = 1.4$ and $\mathbb{E}[X] = 1.1$.
And $\mathbb{E}[Y] = (1 \times 0.6) + (2 \times 0.4) = 1.4$. So $\mathrm{Cov}(X, Y) = 1.4 - (1.1 \times 1.4) = -0.14$ suggesting a negative relationship between X and Y.

Interpretation of Covariance

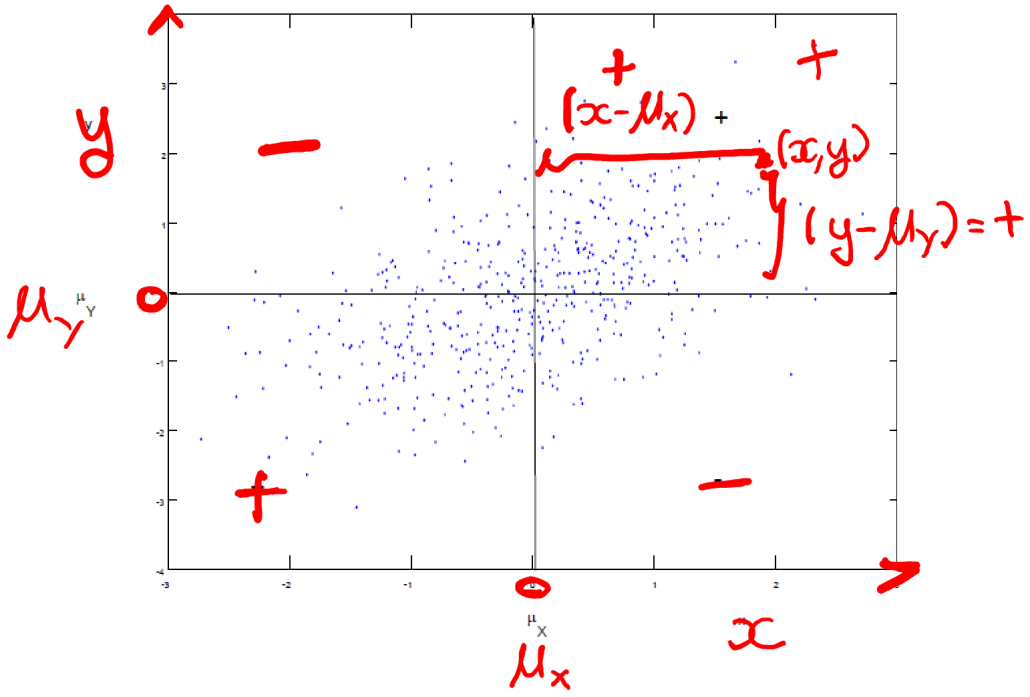


Figure 12: Random points (X, Y) with $\mathrm{Cov}(X, Y) = 0.5$, variances 1.

Theorem. If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$.

Note. The converse it NOT true.

Theorem. Suppose X and Y are independent random variables. Then, if $g_1(X)$ and $g_2(Y)$ are two functions, we have

$$\mathbb{E}[g_1(X)g_2(Y)] = \mathbb{E}[g_1(X)]\,\mathbb{E}[g_2(Y)].$$

Note. $\mathbb{E}[XY] = \sum_{\text{all } (x,y)} xyf(x, y) = \sum_{\text{all } x} \sum_{\text{all } y} xyf(x, y)$.

Definition (**Correlation Coefficient**). The **correlation coefficient** of X and Y is

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Note. Alternatively, $\rho = \dfrac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$.

The correlation coefficient measures the strength of the **linear** relationship between X and Y, it is essentially a rescaled version of the covariance, scaled to lie in the interval $[-1, 1]$.

Theorem (**Properties of** ρ).

1. Since $\sigma_X, \sigma_Y > 0$, ρ will have the same sign as $\text{Cov}(X, Y)$. Hence the interpretation of the sign of ρ is the same as for $\text{Cov}(X, Y)$, and $\rho = 0$ if X and Y are independent. When $\rho = 0$, we say that X and Y are uncorrelated.
2. $-1 \leq \rho \leq 1$ and as $\rho \to \pm 1$ the relation between X and Y becomes one-to-one and linear.

Example. The joint probability function of (X, Y) is:

		x			
$f(x, y)$		0	1	2	$f_Y(y)$
y	1	0.06	0.15	0.09	0.3
	2	0.14	0.35	0.21	0.7
$f_X(x)$		0.2	0.5	0.3	1

Calculate the correlation coefficient. What does this say about the relationship between X and Y?

Solution: $\mathbb{E}[Y] = (1 \times 0.7) = 0.7$, $\mathbb{E}[X] = (1 \times 0.5) + (2 \times 0.3) = 1.1$, $\mathbb{E}[XY] = (1 \times 1 \times 0.35) + (2 \times 1 \times 0.21) = 0.77$.

$\implies \text{Cov}(X, Y) = 0.77 - (0.7 \times 1.1) = 0$. So $\rho = \text{Corr}(X, Y) = \dfrac{0}{\sigma_X \sigma_Y} = 0$.

Therefore, X and Y have no linear association (and in fact, X and Y are independent, we can show this for all (x, y)).

9.5 Mean and Variance of a Linear Combination of Random Variables

Theorem (**Results for Expectations**).

1. $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.
2. Let a_i be constants and $\mathbb{E}[X_i] = \mu_i$, $i = 1, 2, \ldots, n$. Then $\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mu_i$. In particular, $\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$.
3. Let $X_1, X_2, \ldots, X_n$ be random variables which have mean μ. Then, the sample mean is $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ and $\mathbb{E}\left[\overline{X}\right] = \mu$.

Theorem (**Results for Covariance**).

1. $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$.
2. $\operatorname{Cov}(aX + bY, cU + dV) = ac\operatorname{Cov}(X, U) + ad\operatorname{Cov}(X, V) + bc\operatorname{Cov}(Y, U) + bd\operatorname{Cov}(Y, V)$.

Note. Covariance is symmetric: $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$. And for all $c \in \mathbb{R}$, $\operatorname{Cov}(c, X) = 0$.

Theorem (**Results for Variance**).

1. **Variance of a linear comobination:**

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y).$$

2. **Variance of a sum of independent random variables:**

   Let X and Y be independent. Since $\operatorname{Cov}(X, Y) = 0$, result 1 gives

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) = \operatorname{Var}(X - Y).$$

3. **Variance of a general linear combination of random variables:**

   Let a_i be constants. Then

$$\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}(X_i) + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j).$$

4. **Variance of a linear combination of independent random variables:**

   Special cases of result 3 are:

   (a) If $X_1, X_2, \ldots, X_n$ are independent, then $\text{Cov}(X_i, X_j) = 0$, so that

   $$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \,\text{Var}(X_i).$$

   (b) If $X_1, X_2, \ldots, X_n$ are independent and all have the same variance $\text{Var}(X) = \sigma^2$, then

   $$Var\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Remark. If $X_1, X_2, \ldots, X_n$ are independent random variables with the same mean μ and the same variance σ^2, then the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has

$$\mathbb{E}\left[\overline{X}\right] = \mu \quad \text{and} \quad \text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

What does this tell us about $\overline{X}$?

$\Longrightarrow$ $\overline{X}$ is less variable than X_i (the larger n is the less variability there is).

- This happens because as n increases, e are adding more information and so our sample mean. $\overline{X}$ is becoming **more precise** in sense that we have

  $$\text{Var}\left(\overline{X}\right) \to 0 \text{ as } n \to \infty.$$

- This implies that as $n \to \infty, \overline{X} \to \mu$.
- This is something called the "**law of averages**".

9.6 Linear Combinations of Independent Normal Random Variables

Theorem (**Linear Combinations of Independent Normal Random Variables**).

1. Let $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, where $a, b \in \mathbb{R}$. Then, $Y \sim N(a\mu + b, a^2\sigma^2)$.
2. Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent random variables, and let a and b be constants. Then, $aX + bY \sim \mathrm{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.

   In general, if $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$, are independent random variables and $a_1, a_2, \ldots a_n \in \mathbb{R}$, then $\displaystyle\sum_{i=1}^{n} a_i X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$.
3. Let $X_1, X_2, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Then $\displaystyle\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$ and $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Note. We have $\mathbb{E}[aX + b] = a\mu + b$ and $\mathrm{Var}(aX + b) = a^2\sigma^2$.

Example. Let $X \sim \mathrm{N}(3, 5)$ and $Y \sim \mathrm{N}(6, 14)$ be independent random variables. Find $P(X > Y)$.

Solution: Note that $P(X > Y) = P(X - Y > 0)$. Let $W = X - Y$. Then $\mathbb{E}[W] = \mathbb{E}[X] - \mathbb{E}[Y] = 3 - 6 = -3$ and $\mathrm{Var}(W) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 5 + 14 = 19$ since X and Y are independent. So $W \sim \mathrm{N}(-3, 19)$.

$$\begin{aligned} P(X > Y) &= P(W > 0) \\ &= P(Z > \frac{0 - (-3)}{\sqrt{19}}) \\ &= P(Z > 0.69) \\ &= 1 - P(Z \leq 0.69) = 0.245. \end{aligned}$$

Example. Let $X \sim \mathrm{N}(5, 4)$. An independent random variable Y is also normal with mean 7 and standard deviation of 3. Find:

(a) The probability $2X$ differs from Y by more than 4.

(b) The minimum number, n, of independent observations needed on X so that $P(|\overline{X} - 5| < 0.1) \geq 0.98$.

Solution:

(a) Let $W = 2X - Y$. Then $\mathbb{E}[W] = 2\mathbb{E}[X] - \mathbb{E}[Y] = 10 - 7 = 3$ and $\text{Var}(W) = 4\,\text{Var}(X) + \text{Var}(Y) = 16 + 3^2 = 25$. Next,

$$\begin{aligned} P(|2X - Y| > 4) = P(|W| > 4) &= P(W > 4) + P(W < -4) \\ &= P(Z > \frac{4-3}{5}) + P(Z < \frac{-4-3}{5}) \\ &= P(Z > 0.2) + P(Z < -1.4) \\ &= [1 - P(Z \leq 0.2)] + [1 - P(Z \leq 1.4)] \end{aligned}$$

(b) We have that $\overline{X} \sim \text{N}(\mu = 5, \frac{\sigma^2}{n} = \frac{4}{n})$.

$$\begin{aligned} P(|\overline{X} - 5| < 0.1) &\geq 0.98 \\ P(-0.1 < \overline{X} - 5 < 0.1) &\geq 0.98 \\ P(\frac{-0.1}{\sqrt{\frac{4}{n}}} < Z < \frac{0.1}{\sqrt{\frac{4}{n}}}) &\geq 0.98 \\ \implies P(Z \leq \frac{0.1}{\sqrt{\frac{4}{n}}}) &\geq 0.99 \quad \text{use a graph...} \\ \frac{0.1}{\sqrt{\frac{4}{n}}} \geq 2.33 &\implies n \geq 2171.56 \\ n = 2172. & \end{aligned}$$

More precisely, use R to get $n \geq 2165$.

9.7 Indicator Random Variables

An indicator variable is a binary variable (0 or 1) that indicates whether or not an event has occurred. It can allow us to take more complicated scenarios and break them into simpler ones.

Example. Let $X \sim \text{Bin}(n, p)$. Define new random variables X_i by

$$X_i = \begin{cases} 0 & \text{if } i^{\text{th}} \text{ trial was a failure} \\ 1 & \text{if } i^{\text{th}} \text{ trial was a success} \end{cases}.$$

The random variable X_i indicates whether 'success' occured on the i^{th} trial. The total number of successes is $X = \sum_{i=1}^{n} X_i$. We can find the mean and variance of X_i and then use our results for mean and variance to find μ and σ^2 of X. First,

$$\mathbb{E}[X_i] = \sum_{x_i=0}^{1} x_i f(x_i) = 0f(0) + 1f(1) = f(1).$$

But $f(1) = p$ since $P(\text{success}) = p$ on each trial. Therefore $\mathbb{E}[X_i] = p$. And since $X_i = 0$ or 1, so $X_i = X_i^2$, and therefore

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}[X_i] = p.$$

Thus,

$$\text{Var}(X_i) = \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 = p - p^2 = p(1-p).$$

In the Binomial distribution the trials are independent, so the X_i's are also independent. Thus

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np.$$
$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

Remark. Note that $X_i \sim \text{Binomial}(1, p)$ is actually a Binomial random variable. In some problems the X_i's are not independent, we will also need covariances.

Example. We have N letters to N different people, and N envelopes addressed to those N people. One letter is put in each envelope at random. Find the mean and variance of the number of letters placed in the right envelopes.

Solution: Let

$$X_i = \begin{cases} 0 & \text{if letter } i \text{ is not in envelope } i \\ 1 & \text{if letter } i \text{ is in envelope } i \end{cases}.$$

Then, the total number of correctly placed letters is $\sum_{i=1}^{N} X_i$. Note that X_i's are dependent since the experiment is done without replacenent!

$$\mathbb{E}[X_i] = \sum_{x_i=0}^{1} x_i f(x_i) = f(1) = \frac{1}{N} = \mathbb{E}\left[X_i^2\right]$$

since there is 1 chance in N that letter i will be put in envelope i and then,

$$\text{Var}(X_i) = \mathbb{E}[X_i] - (\mathbb{E}[X_i])^2 = \frac{1}{N} - \frac{1}{N^2} = \frac{1}{N}\left(1 - \frac{1}{N}\right).$$

Therefore,

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbb{E}[X_i] = \sum_{i=1}^{N} \frac{1}{N} = 1.$$

To find $\text{Var}\left(\sum_{i=1}^{N} X_i\right)$, we need to calculate the covariance terms since X_i's are dependent.
Next, $\text{Cov}(X_i, X_j) = \mathbb{E}\left[X_i X_j\right] - \mathbb{E}[X_i]\,\mathbb{E}\left[X_j\right]$ for $i \neq j$. So, we need

$$\mathbb{E}\left[X_i X_j\right] = \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} x_i x_j f(x_i, x_j) = f(1,1) = P(X_i = 1, X_j = 1) = P(A \cap B).$$

where $A =$ letter i is placed in envelop i, $B =$ letter j is placed in envelop j. Then,

$$\mathbb{E}\left[X_i X_j\right] = P(A)P(B|A) = \frac{1}{N} \cdot \frac{1}{N-1} = \frac{1}{N(N-1)}.$$

Now, we can calculate the covariance:

$$\begin{aligned}
\operatorname{Cov}(X_i, X_j) &= \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] \\
&= \frac{1}{N(N-1)} - \left(\frac{1}{N}\right)\left(\frac{1}{N}\right) \\
&= \frac{1}{N}\left(\frac{1}{N-1} - \frac{1}{N}\right) \\
&= \frac{1}{N^2(N-1)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{Var}(X) = \operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{N} \operatorname{Var}(X_i) + 2\sum_{i<j}^{N} \operatorname{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{N} \frac{1}{N}\left(1 - \frac{1}{N}\right) + 2\sum_{i<j}^{N} \frac{1}{N^2(N-1)} \\
&= \left[N \cdot \frac{1}{N}\left(1 - \frac{1}{N}\right)\right] + 2\binom{N}{2} \cdot \frac{1}{N^2(N-1)} \\
&= \left(1 - \frac{1}{N}\right) + 2\frac{N!}{2!(N-2)!} \cdot \frac{1}{N^2(N-1)} \\
&= 1 - \frac{1}{N} + \frac{1}{N} \\
&= 1.
\end{aligned}$$

For $\binom{N}{2}$, note that we are summing the covariance terms for all distinct pairs of X_i and X_j. We are wanting to decide how many combinations X_i, X_j are there for which $i < j$. Each i and j takes values from $1, 2, \ldots, N$ so there are $\binom{N}{2}$ different combinations of (i, j) values.

10 Central Limit Theorem and Moment Generating Functions

10.1 Central Limit Theorem

Under certain conditions, the normal distribution can be used to approximate probabilities for linear combinations of random variables having a non-normal distribution. This follows from the Central Limit Theorem (C.L.T.).

The normal distribution is commonly used because it tends to approximate the distribution of sums of random variables.

Theorem (**Central Limit Theorem**).
If $X_1, X_2, \ldots, X_n$ are **independent** random variables all having the **same distribution**, with mean μ and variance σ^2, then as $n \to \infty$, the cumulative distribution function (c.d.f.) of the random variable

$$Z \approx \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

approaches the $N(0,1)$ cumulative distribution function. Similarly, the cumulative distribution function of

$$Z \approx \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

approaches the $N(0,1)$ cumulative distribution function.

Remark. For large n, we have

- $S_n = \sum_{i=1}^{n} X_i \sim \mathrm{N}(n\mu, n\sigma^2)$.
- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$.

If X_i's themselves have normal distributions, then S_n and $\overline{X}$ have **exactly** normal distributions $\forall n$. Otherwise, S_n and $\overline{X}$ have **approximately** normal distributions.

Note.

- Although this theorem is about limits, we will use it when n is large, but finite.
- This theorem works for all distributions except those whose μ and σ^2 do not exist.
- The accuracy of the approximation depends on n (bigger is better) and on the actual distribution X_i's. It works well for small n when X_i's p.d.f. is close to symmetric.

Theorem (**The Law of Large Numbers**).
Draw simple random samples of size n at random from a large population with mean μ. As the number of observations drawn increases (i.e. as $n \to \infty$), then the sample mean $\overline{X}$ approaches the population mean μ.

General idea of the C.L.T. is that it takes any distribution and makes it normal.

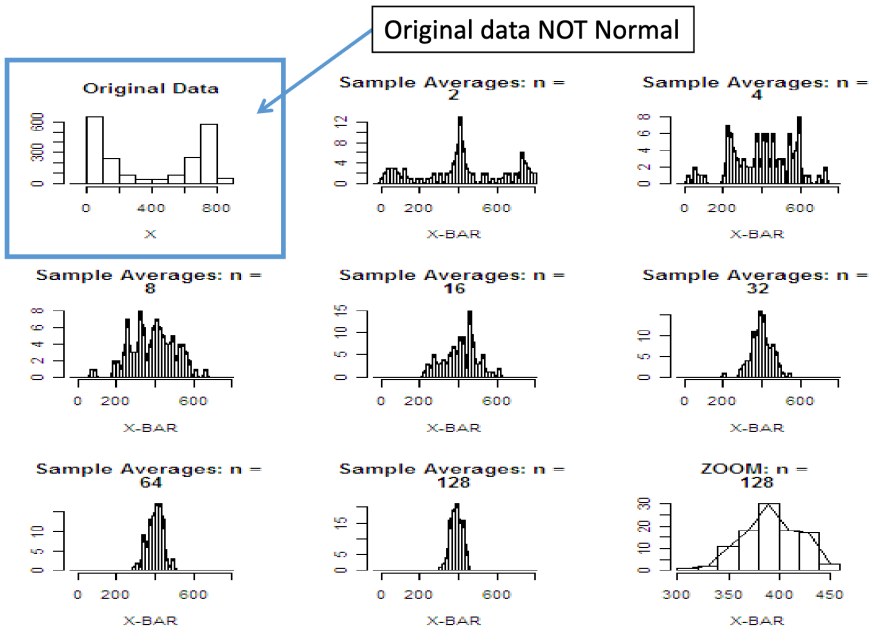
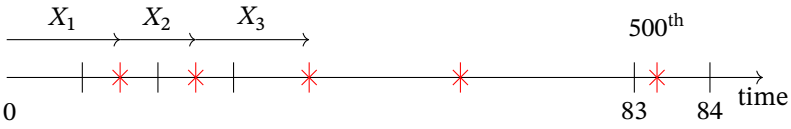


Figure 13: Visualizing the idea of CLT.

Note. The distribution of sample means becomes normal as the sample size increases. Also, the sample mean converges to the population mean when n becomes larger.

Example. Suppose fires reported to a fire station satisfy the conditions for a Poisson process, with a mean of 1 fire every 4 hours. Find the probability the 500^{th} fire of the year is reported on the 84^{th} day of the year.

Solution: Let X_i = time beteen $(i-1)^{\text{st}}$ and the i^{th} fires (X_1 = time to the first fire). Then, these X_i's are independent and identically distributed as: $X_i \sim \text{Exp}(\theta = 4 \text{ hrs} = \frac{1}{6} \text{ day})$, since $\lambda = \frac{1}{4}$ fires per hour.



We want to find $P\left(83 < \underbrace{\sum_{i=1}^{500} X_i}_{S_{500}} \leq 84\right)$. By the Central Limit Theorem, $S_{500} = \sum_{i=1}^{500} X_i$ has approximately a $\text{N}(500\mu, 500\sigma^2) = \text{N}(\frac{500}{6}, \frac{500}{36})$ distribution.

$$\begin{aligned} P(83 < S_{500} \leq 84) &\approx P\left(\frac{83 - \frac{500}{6}}{\sqrt{\frac{500}{36}}} < Z \leq \frac{84 - \frac{500}{6}}{\sqrt{\frac{500}{36}}}\right) \\ &= P(-0.09 < Z \leq 0.19) \\ &\vdots \\ &= 0.57142 + 0.53586 - 1 \\ &= 0.10728. \end{aligned}$$

Note. In this example, we used the normal distribution to approximate a continuous random variable (i.e. the exponential distribution). When approximating discrete random variables, we have to make a small adjustment, see next page!

Remark (**Continuity Correction**).
When approximating a discrete random variable, a slight adjustment is required to improve the approximation.

For example, we are in a '100-cup challenge' where $P(\text{winning cup}) = \frac{1}{6}$, we bought 100 cups to see how many times we win. We want to find the probability of having between 15 to 20 winning cups.

Let X_i = whether the i^{th} cup is a winning cup. Then $X_i \sim \text{Binomial}(n = 1, p = \frac{1}{6})$ for $i = 1, 2, \dots, 100$. And $\mathbb{E}[X_i] = \frac{1}{6}$, $\text{Var}(X_i) = \frac{5}{36}$, with $S_{100} = \sum_{i=1}^{100} X_i$ = total number of winning cups. By the CLT, we have S_{100} approximately have $\text{N}(\frac{100}{6}, \frac{500}{36})$, we'll see a theorem below about this. Then,

$$P(15 \leq S_{100} \leq 20) = P(-0.447 \leq Z \leq 0.894) = 0.487.$$

If we compute the exact probability using $S_{100} \sim \text{Binomial}(100, \frac{1}{6})$, we get $P(15 \leq S_{100} \leq 20) = 0.561$. This is quite off!

Since X_i are discrete, so S_{100} is also discrete and cannot take non-integer values. In this case, the approximation is underestimate as seen in Figure 14 (part of $X = 15$ and $X = 20$ are not included by the normal approximation in blue), so subtract 0.5 to get $P(14.5 \leq S_{100} \leq 20.5)$, which gives a better approximation.

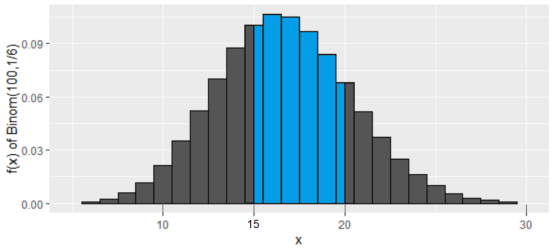


Figure 14: Without adjustment (underestimate).

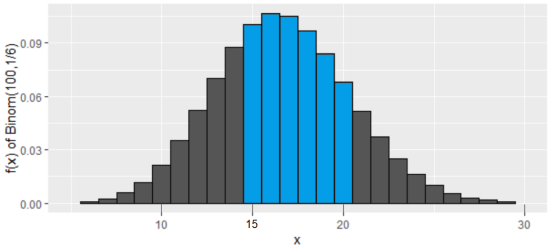


Figure 15: With adjustment (better).

$P(14.5 \leq S_{100} \leq 20.5) = 0.568$, which is way better!

This adjustment is called the "**Continuity Correction**".

Note.

- It is only applied when using a **continous** distribution to approximate a **discrete** one.
- Quick sketch to see whether to add or subtracct 0.5 (see below proposition).

Proposition (**Continuity Correction Rules**).

- $P(X > n)$, use $P(X > n + 0.5)$.
- $P(X \geq n)$, use $P(X > n - 0.5)$.
- $P(X < n)$, use $P(X < n - 0.5)$.
- $P(X \leq n)$, use $P(X < n + 0.5)$.
- $P(a < X < b)$, use $P(a + 0.5 \leq X \leq b - 0.5)$.
- $P(a \leq X \leq b)$, use $P(a - 0.5 \leq X \leq b + 0.5)$.
- $P(X = n)$, use $P(n - 0.5 \leq X \leq n + 0.5)$.

Theorem (**Normal Approximation to Poisson**).

Suppose $X \sim Poisson(\mu)$. Then the cumulative distribution function of the standarized random variable

$$Z = \frac{X - \mu}{\sqrt{\mu}}$$

approaches that of a standard Normal random variable as $\mu \to \infty$.

Remark. We have $X \sim \mathrm{N}(\mu, \mu)$, where $\mu = \lambda t$. Approximation will be good when $\mu > 5$.

Theorem (**Normal Approximation to Binomial**).

Suppose $X \sim Binomial(n, p)$. Then for large n, the random variable

$$W = \frac{X - np}{\sqrt{np(1-p)}}$$

has approximately a $N(0, 1)$ distribution.

Remark. We write $\frac{X-np}{\sqrt{np(1-p)}} \sim \mathrm{N}(0, 1)$ or $X \sim \mathrm{N}(np, np(1-p))$.

Example. Suppose $X \sim$ Poisson(9). Use the Normal approximation to approximate $P(X > 9)$, and compare with the true value.

By Normal approximation to Poisson, we have $Z = \frac{X-\mu}{\sqrt{\mu}}$. Using continuity correction,

$$P(X > 9) \approx P(X > 9 + 0.5) = P\left(Z > \frac{9.5 - 9}{\sqrt{9}}\right) = P(Z > 0.17) = 0.4324.$$

Note: the true value is 0.4126.

Example. Suppose $X \sim \text{Binomial}(20, 0.4)$. Find the approximate probability $P(4 \leq X \leq 12)$ and compare it to the true value.

Solution: Note that $\mathbb{E}[X] = np = 8$ and $\text{Var}(X) = np(1-p) = 4.8$. By the Normal approximation to the Binomial, we have $X \sim \text{N}(8, 4.8)$ approximately.

$$\begin{aligned} P(4 \leq X \leq 12) &\approx P(4 - 0.5 \leq X \leq 12 + 0.5) \\ &= P\left(\frac{3.5-8}{\sqrt{4.8}} \leq \frac{X-8}{\sqrt{4.8}} \leq \frac{12.5-8}{\sqrt{4.8}}\right) \\ &= P(-2.054 \leq Z \leq 2.054) \\ &\vdots \\ &= 0.96. \end{aligned}$$

Note: the true value is $\sum_{x=4}^{12} \binom{20}{x} (0.4)^x (0.6)^{20-x} = 0.963$.

Example. Let p be the proportion of Canadians who think Canada should adopt the US dollar.

(a) Suppose 400 Canadians are randomly chosen and asked their opinion. Let X be the number who say yes. Find the probability that the proportion, $X/400$, of people who say yes is within 0.02 of p, if p is 0.20. **Ans:** $2P(Z < 1.06) - 1 = 0.711$.

(b) Find the number, n, who must be surveyed so there is a 95% chance that X/n lies within 0.02 of p, when p is unknown.

Ans: $Z = 1.96 = \frac{0.02}{\sqrt{\frac{p(1-p)}{n}}}$. Set $(p(1-p))' = 0$ to find the max $p = 0.5$. Sub in $p = 0.5$ to get $n = 2401$, and this works for all p since it works for the max p.

10.2 Moment Generating Functions

So far, we have seen two functions which characterize a distribution of a random variable:

- p.d.f.
- c.d.f.

However, there is a third function which also **uniquely** determines a distribution: the **moment generating function**.

Definition (**Moment Generating Functions for Discrete**).
Consider a discrete random variable X with probability function $f(x)$. The moment generating function (m.g.f.) of X is defined as

$$M(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{\text{all } x} e^{tx} f(x).$$

We will assume that the moment generating function is defined and finite for values of t in an interval around 0 (i.e. for some $a > 0$, $\sum_x e^{tx} f(x) < \infty \ \forall t \in [-a, a]$).

The m.g.f. of X can be used to evaluate the moments of the random variable X, where the **moments** of X are defined as the **expectations** of the functions X^k for $k = 1, 2, \ldots$.

$$\mathbb{E}\left[X^k\right] \text{ is the } k^{\text{th}} \text{ moment of } X.$$

Example.

- The mean $\mu = \mathbb{E}[X]$ is the first moment of X.
- $\mathbb{E}\left[X^2\right]$ is the second moment of X and so on.

Theorem.
Suppose the random variable X has moment generating function $M(t)$ defined $\forall t \in [-a, a]$ for some $a > 0$. Then

$$\mathbb{E}\left[X^k\right] = M^{(k)}(0) \quad \text{for } k = 1, 2, \ldots$$

where $M^{(k)}(0) = \dfrac{\mathrm{d}^k}{\mathrm{d}t^k} M(t)\,|_{t=0} \quad \text{for } k = 1, 2, \ldots$

Example (**MGF for Binomial**).

Suppose $X \sim \text{Binomial}(n, p)$. Then its moment generating function is

$$\begin{aligned} M(t) &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= (pe^t + 1 - p)^n \quad \text{by the Binomial Theorem } \forall t. \end{aligned}$$

Therefore,

$$\begin{aligned} M'(t) &= npe^t \left(pe^t + 1 - p\right)^{n-1} \\ M''(t) &= npe^t \left(pe^t + 1 - p\right)^{n-1} + n(n-1)p^2 e^{2t} \left(pe^t + 1 - p\right)^{n-2} \end{aligned}$$

and so

$$\begin{aligned} \mathbb{E}[X] &= m'(0) = np \\ \mathbb{E}\left[X^2\right] &= M''(0) = np + n(n-1)p^2 \\ \text{Var}(X) &= \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2 = np(1-p). \end{aligned}$$

Example (exercise)**.** Find the m.g.f. for a Poisson distrubition $X \sim \text{Poisson}(\mu)$.

Theorem (**Uniqueness Theorem for MGFs**).

Suppose that random variables X and Y have moment generating functions $M_X(t)$ and $M_Y(t)$, respectively. If $M_X(t) = M_Y(t)$ $\forall t$, then X and Y have the same distribution.

The m.g.f. uniquely identifies a distribution. For example, if X has m.g.f.: $e^{2(e^t-1)}$. Then, you can immediately tell that $X \sim \text{Poisson}(2)$.

MGFs can also be used to determine that a sequence of distributions gets closer and closer to some limiting distribution. To show this, suppose that a sequence of probability functions $f_n(x)$ have corresponding moment generating functions:

$$M_n(t) = \sum_{\text{all } x} e^{tx} f_n(x).$$

And suppose that the probability functions $f_n(x)$ converge to another probability function $f(x)$ pointwise in x as $n \to \infty$. Then since

$$f_n(x) \to f(x) \text{ as } n \to \infty \text{ for each } x,$$

we have

$$\sum_{\text{all } x} e^{tx} f_n(x) \to \sum_{\text{all } x} e^{tx} f(x) \text{ as } n \to \infty \text{ for each } t.$$

In other words, $M_n(t)$ converges to $M(t)$, the moment generating function of the limiting distribution.

Note: The converse also holds. Suppose that X_n has m.g.f. $M_n(t)$ and $M_n(t) \to M(t)$ for each t, such that $M(t) < \infty$, then

$$f_n(x) \to f(x) \text{ as } n \to \infty \text{ for each } x.$$

Example. The Poisson approximation of the Binomial distribution when n is very large and p is close to 0.

$\mu = np \implies p = \frac{\mu}{n}$. Then, the m.g.f. of such a Binomial random variable:

$$\begin{aligned} M(t) &= (pe^t + 1 - 1)^n \\ &= \left[1 + p(e^t - 1)\right]^n \\ &= \left[1 + \frac{\mu}{n}(e^t - 1)\right]^n. \end{aligned}$$

Now, take the limit of this expression as $n \to \infty$. Since $\lim_{n\to\infty} \left(1 + \frac{c}{n}\right)^n \to e^c$, we have

$$\lim_{n\to\infty} \left[1 + \frac{\mu}{n}(e^t - 1)\right]^n = e^{\mu(e^t - 1)},$$

and this is the m.g.f. of a Poisson distribution with parameter μ. This shows a little more formally that the Binomial distribution with small p and large n approaches the Poisson distribution with $\mu = np$ as $n \to \infty$.

Moment Generating Function of a Continuous Random Variable

Definition (**Moment Generating Functions for Continous**).
Consider a continuous random variable X with probability density function $f(x)$. Then moment generating function of X is defined as

$$M(t) = \mathbb{E}\left[e^{tx}\right] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

We will assume that the m.g.f. is defined and finite for values of t in an interval around 0. That is, for some $a > 0$, $\int_{-\infty}^{\infty} e^{tx} f(x)\, dx < \infty \ \forall t \in [-a, a]$.

Example. Suppose that $X \sim$ Exponential(θ). Show that the m.g.f. is given by

$$M(t) = \frac{1}{1-\theta t} \quad \text{for } t < \frac{1}{\theta}.$$

Solution: Recall that $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\, dx$. By definition, we have

$$\begin{aligned}
M(t) &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x}\, dx \\
&= \lambda \int_0^{\infty} x^{1-1} e^{-x(\lambda - t)}\, dx \quad \text{let } u = x(\lambda - t), dx = \frac{du}{\lambda - t} \\
&= \frac{\lambda}{\lambda - t} \int_0^{\infty} \left(\frac{u}{\lambda - t}\right)^{1-1} e^{-u}\, dx \\
&= \frac{\lambda}{\lambda - t} \int_0^{\infty} u^{1-1} e^{-u}\, dx \\
&= \frac{\lambda}{\lambda - t} \cdot \Gamma(1) \\
&= \frac{\lambda}{\lambda - t} \\
&= \frac{1}{1-\theta t} \quad \text{since } \theta = \frac{1}{\lambda}.
\end{aligned}$$

Notice that $\lambda - t > 0$ is necessary for the integral to converge, so we have $t < \lambda = \frac{1}{\theta}$.

Proposition (**Summary of MGFs**).

For discrete random variables:

- If $X \sim \text{Binomial}(n, p)$, then $M(t) = (pe^t + 1 - p)^n$, $t \in \mathbb{R}$.
- If $X \sim \text{Poisson}(\mu)$, then $M(t) = e^{\mu(e^t - 1)}$, $t \in \mathbb{R}$.
- If $X \sim \text{NB}(k, p)$, then $M(t) = \left(\frac{p}{1-(1-p)e^t}\right)^k$, $t < -\ln(1-p)$.
- If $X \sim \text{Geometric}(p)$, then $M(t) = \frac{p}{1-(1-p)e^t}$, $t < -\ln(1-p)$.

For continuous random variables:

- If $X \sim \text{U}(k, p)$, then $M(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$, $t \neq 0$.
- If $X \sim \text{Exponential}(\theta)$, then $M(t) = \frac{1}{1-\theta t}$, $t < -\frac{1}{\theta}$.
- If $X \sim \text{NB}(k, p)$, then $M(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$, $t \in \mathbb{R}$.

10.3 Motivariate Moment Generating Functions

Did-not-have-time-to-cover material.

END OF STAT 230!