STAT 230 Notes

Zhilin (Catherine) Zhou

April 6, 2020

Contents

Chapter 1

Intro to Probability

1.1 Definitions

Uncertainty of Randomness

- variability in populations consisting of animate or inanimate objects
- variability in processes or phenomena

Definition of Probability

Definition 1.1.1. ***Probability***

- *Classical Definition:* $\dfrac{\textit{number of ways the } \textbf{\textit{event}} \textit{ can occur}}{\textit{number of outcomes in } \textbf{\textit{S}}}$. *All points in the sample space* S *are equally likely.*
- *Relative Frequency: The (limiting)* ***proportion*** *(or fraction)* ***of times*** *the event occurs in a very long series of repetitions of an experiment or process.*
- *Subjective Probability: A measure of* ***how sure*** *the person making the statment is that the event will happen.*

Note: All these definitions have serious **limitations**:

- Classical Definition: hard to define **"equally likely"**
- Relative Frequency: we can never repeat an experiment **indefinitely**, or obtain a long series of **repetitions** due to time, cost, or other limitations
- Subjective Probability: it gives no rational basis for people to agree on a right answer

1.1.1 Probability Model

Definition 1.1.2. ***Probability Model***

- *a sample space of all possible outcomes of a random experiment is defined*
- *a set of events, subsets of the sample space to which we can assign probabilities, is defined*

- *a mechanism of assigning probabilities (numbers between 0 and 1) to events is specified*

Definition 1.1.3. ***Discrete***
S is discrete if it consists of a finite or countably infinite set of simple events. A countably infinite sequence is one that can be put into a one-to-one correspondence with the positive integers.

Definition 1.1.4. ***Simple event***
An event in discrete sample space is a subset $A \subset S$, if the event is indivisible so it contains only one point, we call it simple event.

Definition 1.1.5. ***Compound event***
An event A made up of two or more simple events such as is called a compound event.

Definition 1.1.6. ***Probability distribution***
Let $S = \{a_1, a_2, \ldots\}$ be a discrete sample space, assign probabilities $P(a_i), i = 1, 2, 3, \ldots$ to the a_i's such that the following two conditions hold:

1. $0 \leq P(a_i) \leq 1$
2. $\sum_{\text{all } a_i} P(a_i) = 1$

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

Chapter 2

Mathematical Probability Models

2.1 Sample Spaceas and Probability

Definition 2.1.1. ***Experiment***: *the phenomenon or process*

Definition 2.1.2. ***Trial***: *a single repetition of the experiment*

2.1.1 Sample Space

Definition 2.1.3. ***Sample Space***: *The set of all possible* <u>*distinct*</u> *outcomes to a random experiment, with the property that in a single trial, one and only one of these outcomes occurs.*

Experiment is made up of trials. The outcomes of an experiment is the sample space.
Sample space may be **discrete** (finite or countably infinite set of simple events) or **non-discrete**. We only consider discrete sample spaces.

Chapter 3

Probability and Counting Techniques

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} \tag{3.1}$$

3.1 Addition and Multiplication Rules

Theorem 3.1.1. ***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.*

Theorem 3.1.2. ***Multiplication Rule***
Suppose we can do job 1 in p ways, ***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.*

3.2 Counting Arranegements or Permutations

Example:
Suppose the letters a, b, c, d, e, f are arranged at random to form a six-letter word (an arrangement) – we must use each letter once only.

We generalize the problem in several ways. In each case, we count the number of arragements by counting the number of ways we can fill the positions in the arrangement. Suppose we start with n symbols, we have

- $n \times (n-1) \times \cdots \times 1$ arrangements of length n using each symbol once and only once. This product is denoted by $n!$
- $n \times (n-1) \times \cdots \times (n-k+1)$ arragements of length k using each symbol at most once. This product is denoted by $n^{(k)}$
- $n \times n \times \cdots \times n = n^k$ arrangements of length k using each symbol as often as we wish

Theorem 3.2.1. ***Stirling's Approximation****:*
For large n there is an approximation to $n!$ called Stirling's approximation. Note that the sequence $\{a_n\}$ is asymptotically equivalent to the sequence $\{b_n\}$ if

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 1$$

3.3 Counting Subsets or Combinations

Definition 3.3.1. ***Number of subsets of size*** k
We use the combinatorial symbol $\binom{n}{k}$ *to denote the number of subsets of size* k *that can be selected from a set of* n *objects.*

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

3.3.1 Properties of Combination

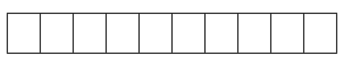
You should be able to prove the following for n and k non-negative integers:

1. $n^{(k)} = \frac{n!}{(n-k)!} = n(n-1)^{(k-1)}$ for $k \geq 1$
2. $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n^{(k)}}{k!}$
3. $\binom{n}{k} = \binom{n}{n-k}$ for all $k = 0, 1, \ldots, n$
4. If we define $0! = 1$, then the formulas hold with $\binom{n}{0} = \binom{n}{n} = 1$
5. $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$
6. **Binomial Theorem**: $(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots + \binom{n}{n}x^n$

3.4 Number of Arragements 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 arragement begins and ends with S

Solution. We construct the arragements by filling ten boxes corresponding to the positions in the arragement



We can choose the position for the three S's in $\binom{10}{3}$ ways. For each of these choices, we can choose the positions for the three T's in $\binom{7}{3}$ ways, then we can place the two I's in $\binom{4}{2}$ ways, then the C in $\binom{2}{1}$ ways and finally the A in $\binom{1}{1}$ ways. The number of equally probable outcomes in the sample space is

$$\binom{10}{3}\binom{7}{3}\binom{4}{2}\binom{2}{1}\binom{1}{1} = \frac{10!7!4!2!1!}{3!7! \cdot 3!4! \cdot 2!2! \cdot 1!1! \cdot 1!0!} = \frac{10!}{3!3!2!1!1!}$$

■

Chapter 4

Probability Rules and Conditional Probability

4.1 General Methods

Rules:

1. $P(S) = 1$
2. For any event A, $0 \leq P(A) \leq 1$
3. If A and B are two events with $A \subseteq B$, then $P(A) \leq P(B)$

Theorem 4.1.1. ***De Morgan's Laws***

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

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

4.1.1 Rules of Union of Events

Rules:

5. (a) **Addition Law of Probability or the Sum Rule**

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

(b) **Probability of the Union of Three Events**

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

(c) **Probability of the Union of n Events**

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_n) = & \sum_i P(A_i) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) \\ & - \sum_{i<j<k<l} P(A_i A_j A_k A_l) + \cdots \end{aligned}$$

where the subscripts are all distince, for example $i < j < k < l$

6. (a) **Probability of the Union of Two Mutually Exclusive Events**
   Let A and B be mutually exclusive events. Then

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

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

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

7. **Probability of the Complement of an Event**

$$P(A) = 1 - P(\bar{A})$$

4.2 Intersections of Events and Independence

Definition 4.2.1. ***Independent and Dependent Events***
Events A and B are independent events iff

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

If the events are not independent, they are dependent

4.3 Conditional Probability

Definition 4.3.1. $P(A|B)$
*The probability that event A occurs, when we know that B occurs. We call this the **conditional probability** of A given B.*

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

Theorem 4.3.1. *A and B are two events defined on a sample space S s.t.* $P(A) > 0,\ P(B) > 0$*. A and B are **independent** events iff*

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

Example 4.3.1. *The probability of a randomly selected male is colour-blind is 0.05, whereas the probability a female is colour-blind is only 0.0025. If the population is* 50% *male, what is the fraction that is colour-blind?*

Solution. We denote that
C – the person selected is colour-blind
M – the person selected is male

F = $\overline{M}$ – the person selected is female.
We need to find P(C). Given that

$$P(C|M) = 0.05$$
$$P(C|F) = 0.0025$$
$$P(M) = 0.5 = P(F)$$

We need to know what is P(C). From Definition 4.1, we know that

$$P(M) = \frac{P(C \cap M)}{P(C|M)}$$
$$P(F) = \frac{P(C \cap F)}{P(C|F)}$$
$$C = (C \cap M) \cup (C \cap \overline{M})$$

Therefore, we have

$$P(C \cap M) = 0.05 \times 0.5 = 0.025$$
$$P(C \cap F) = 0.0025 \times 0.5 = 0.00125$$
$$P(C) = 0.025 + 0.00125 = 0.02625$$

■

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

Theorem 4.4.1. ***Rule 7: Product Rules***
Let A, B, C, D... 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

$$P(A|B) = P(A)P(B|A)$$
$$P(ABC) = P(A)P(B|A)P(C|AB)$$
$$P(ABCD) = P(A)P(B|A)P(C|AB)P(D|ABC)$$
$$\cdots$$

Theorem 4.4.2. ***Rule 8: Law of Total Probability***
Let $A_1,\ A_2, \cdots A_k$ be a partition of the sample space S into disjoint events, that is

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

Let B be an arbitrary event in S. Then

$$P(B) = P(BA_1) + P(BA_2) + \cdots + P(BA_k)$$
$$= \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

Example 4.4.1. *In an insurance portfolio 10% of the policy holders are in class A_1, 40% are in Class A_2, 10% are in Class A_3. The probability there is a claim on a Class A_1 policy in a given year is 0.10; similar probabilities for Classes A_2 and A_3 are 0.05 and 0.02. Find the probability that if a claim is made, it is made on a Class A_1 policy.*

Solution. We denote
B – policy has a claim
A_i – policy is of Class A_i, i = 1, 2, 3

$$\begin{aligned} P(A_1) &= 0.1 & P(B|A_1) &= 0.10 \\ P(A_2) &= 0.4 & P(B|A_2) &= 0.05 \\ P(A_3) &= 0.5 & P(B|A_3) &= 0.02 \end{aligned}$$

We need to find P(A_1|B). We know that

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1 \cap B)}{P(B)} \\ P(B) &= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) \\ &= P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + P(B|A_3) \cdot P(A_3) \\ &= 0.10 \times 0.1 + 0.05 \times 0.4 + 0.02 \times 0.5 = 0.01 + 0.02 + 0.01 = 0.04 \\ P(A_1 \cap B) &= P(B|A_1) \cdot P(A_1) = 0.10 \times 0.1 = 0.01 \\ \Longrightarrow \quad P(A_1|B) &= \frac{0.01}{0.04} = 0.25 \end{aligned}$$

■

Example 4.4.2. ***Testing for HIV***
Tests used to diagnose medical conditions are often imperfect, and give false positive or false negative results, as described in Problem 2.6 of Chapter 2. A fairly cheap blood test for the Human Immunodeficiency Virus (HIV) that causes AIDS (Acquired Immune Deficiency Syndrome) has the following characteristics: the false negative rate is 2$\%$ and the false positive rate is 0.5$\%$. It is assumed that around 0.04$\%$ of Canadian males are infected with HIV. Find the probability that if a male tests positive for HIV, he actually has HIV.

Solution. We can define

$$\begin{aligned} A &= \text{male has HIV} \\ B &= \text{blood test is positive} \end{aligned}$$

We need to find the value of $P(A|B)$ (when we know that the blood test is positive, it is a male). We know that

$$\begin{aligned} P(A) &= 0.0004 & P(\overline{A}) &= 0.9996 \\ P(B|A) &= 0.98 & P(B|\overline{A}) &= 0.05 \end{aligned}$$

Since by Theorem 4.2, $P(AB) = P(A)P(B|A)$, we can conclude that

$$\begin{aligned} P(AB) &= 0.0004 \times 0.98 \approx 0.000392 \\ P(\overline{A}B) &= P(\overline{A})P(B|\overline{A}) = 0.9996 \times 0.05 \approx 0.004998 \\ P(B) &= P(AB) + P(\overline{A}B) \approx 0.00539 \\ P(A|B) &= \frac{P(AB)}{P(B)} \approx 0.0727 \end{aligned}$$

■

Theorem 4.4.3. ***Bayes' Theorem***
Suppose A and B are events defined on a sample space S with $P(B) > 0$, we have

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

4.5 Useful Series and Sums

4.5.1 Geometric Series

$$\sum_{i=0}^{n-1} t^i = 1 + t + t^2 + \cdots + t^{n-1} = \frac{1-t^n}{1-t} \text{ for } t \neq 1$$

If $|t| < 1$, then

$$\sum_{x=0}^{\infty} t^x = 1 + t + t^2 + \cdots = \frac{1}{1-t}$$

Note: other identities can be obtained from this one by differentiation.

4.5.2 Binomial Theorem

$$(1+t)^n = 1 + \binom{n}{1}t^1 + \binom{n}{2}t^2 + \cdots + \binom{n}{n}t^n = \sum_{x=0}^{n}\binom{n}{x}t^x$$

n is a positive integer and t is any real number.
If $|t| < 1$, then

$$(1+t)^n = \sum_{x=0}^{n}\binom{n}{x}t^x$$

4.5.3 Multinomial Theorem

A generalization of the Binomial Theorem:

$$(t_1 + t_2 + \cdots + t_k)^n = \sum \frac{n!}{x_1!x_2!\dots x_k!} t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}$$

where x_i are all non-negative integers s.t. $\sum_{i=1}^{k} x_1 = n$ where n is a positive integer.

4.5.4 Hypergeometric Identity

$$\sum_{x=0}^{\infty}\binom{a}{x}\binom{b}{n-x} = \binom{a+b}{n}$$

Note: There will not be an infinite number of terms if a and b are positive integers since the terms become 0 eventually.

$$\binom{4}{5} = \frac{4^{(5)}}{5!} = \frac{4 \times 3 \times 2 \times 1 \times 0}{5!} = 0$$

4.5.5 Exponential Series

Let $f(x) = e^x$, then $f^{(k)}(0) = 1$ for $k = 1,\ 2, \ldots$, therefore,

$$e^t = \frac{t^0}{0!} + \frac{t^1}{1!} + \frac{t^2}{2!} + \cdots = \sum_{n=0}^{\infty} \frac{t^n}{n!} \ \forall\, t \in \mathbb{R}$$

We can use the limit definition of the exponential function:

$$e^t = \lim_{n\to\infty} (1 + \frac{t}{n})^n$$

4.5.6 Special series involving integers

$$\begin{aligned} 1 + 2 + \cdots + n &= \frac{n(n+1)}{2} \\ 1^2 + 2^2 + \cdots + n^2 &= \frac{n(n+1)(2n+1)}{6} \\ 1^3 + 2^3 + \cdots + n^3 &= \left[\frac{n(n+1)}{2}\right]^2 \end{aligned}$$

Example 4.5.1. *Find*

$$\sum_{x=0}^{\infty} x(x-1)\binom{a}{x}\binom{b}{n-x}$$

Solution. Since for $x = 0$ or $x = 1$ the term becomes 0, we can start the sumation with $x = 2$.

$$\begin{aligned} \sum_{x=0}^{\infty} x(x-1)\binom{a}{x}\binom{b}{n-x} &= \sum_{x=2}^{\infty} x(x-1)\frac{a!}{x(x-1)(x-2)!(a-x)!}\binom{b}{n-x} \\ &= \sum_{x=2}^{\infty} \frac{a!}{(x-2)!(a-x)!}\binom{b}{n-x} \\ &= \sum_{x=2}^{\infty} \frac{a(a-1)(a-2)!}{(x-2)![(a-2)-(x-2)]!}\binom{b}{n-x} \\ &= a(a-1)\sum_{x=2}^{\infty}\binom{a-2}{x-2}\binom{b}{(n-2)-(x-2)} \\ &= a(a-1)\binom{a+b-2}{n-2} \qquad \text{(by Hypergeometric Identity)} \end{aligned}$$

■

Chapter 5

Discrete Random Variables

5.1 Random Variables and Probability Functions

Introduce numerical-valued variables X, Y, . . . to describe outcomes.

Definition 5.1.1. ***Range****: the set of possible values for the variable.*
e.g. the random variable X = number of heads that occur, it has range A = $\{0,\ 1,\ 2,\ 3\}$

Definition 5.1.2. ***Random Variable****: a function that assigns a real number to each point in a sample space S.*

We denote random variables by capital letters (X, Y, etc.) and denote the actual numbers taken by random variables by small letters (x, y, etc.). There's a difference between **functions** (f(x) or X(a)) and **value of a function** (f(2) or X(a) = 2).

- **Discrete random variables** take integer values or, more generally, values in a countable set. **(finite number, only has one answer)**
  Recall: a set is countable if its elements can be placed in a one-one correspondence with a subset of the positive integers.
- **Continuous random variables** take values in some interval of real numbers like (0, 1), or (0, ∞), or (-∞, ∞). (**infinite number, usually used in measuring. Different person might give a different answer.**)

We want to set up general models to describe how the probability is distributed among the possible values in the range of a random variable X.

Definition 5.1.3. ***Probability Function***
Let X be a discrete random variable with range(X) = A. The ***probability function*** *(p.f.) of X is*

$$f(x) = P(X = x) \text{ defined for all } x \in A$$

The set of pairs $\{(x,\ f(x)) :\ x \in A\}$ *is the* ***probability distribution*** *of X. All probability functions must have two properties:*

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

Definition 5.1.4. ***Comulative Distribution Function***
c.d.f. of X is usually denoted by F(X)

$$F(X) = P(X \leq x) \text{ defined for all } x \in \mathbb{R}$$

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

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

1. *F(x) is a non-decreasing function of x for all* $x \in \mathbb{R}$
2. $0 \leq F(X) \leq 1$ *for all* $x \in \mathbb{R}$
3. $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$
4. *If* $x \in A$ *and* $x - 1 \in A$, $f(x) = F(x) - F(x-1) = P(X = x)$

Remark: State the domain of the probability function. That is, the possible values of the random variable, or the values x for which f(x) is defined. This is the **essential part** of the function's definition.
In general in a probability histogram, probabilities are depicted by areas.

5.1.1 Model Distributions

Many processes or problems have the same structure.

Example 5.1.1. ***The following three problems are essentially the same***

1. *A fair coin is tossed 10 times and the "number of heads obtained" (X) is recorded.*
   (a) *Type of outcome: heads/tails*
   (b) *Repeated times: 10*
2. *20 seeds are planted in separate pots and the "number of seeds germinating" (X) is recorded.*
   (a) *Type of outcome: germinate/don't germinate*
   (b) *Repeated times: 20*
3. *12 items are picked at random from a factory's production line and examined for defects. The number of items having no defects (X) is recorded.*
   (a) *Type of outcome: no defects/defects*
   (b) *Repeated times: 12*

5.1.2 Statistical Computing

Software systems developed for Probability and Statistics, e.g. R.

5.2 Discrete Uniform Distribution

Physical Setup:
Suppose X takes values $a,\ a+1,\ a+2, \ldots b$ with all values being equally likely. Then X has a discrete Uniform Distribution, on the set $\{a,\ a+1,\ a+2, \ldots,\ b\}$.

Probability Function

There are $b - a + 1$ values X can take so the probability at each of these values must be $\frac{1}{b-a+1}$ in order that $\sum_{x=a}^{b} f(x) = 1$. Therefore,

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

Example 5.2.1. *Suppose a fair die is thrown once and let X be the number on the face. First find the cumulative distribution function, $F(x)$ or X.*

Solution. It's a discrete Uniform distribution on the set $\{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ having $a = 1$, $b = 6$, and the probability function,

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

The cumulative distributtion function is $F(x) = P(X \leq x)$,

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

■

5.3 Hypergeometric Distribution

Physical Setup:
We have a collection of N objects which can be classified into two distinct types. Call one type "success" (S) and the other type "failure" (F). There are r sucesses and $N - r$ failures. Pick n objects at random **without replacement**. Let X be the number of successes obtained. Then X has a Hypergeometric distribution.

5.3.1 Probability Function

If we don't consider the order of selection, then there are $\binom{N}{n}$ points in the sample space S. There are $\binom{r}{x}$ ways to choose hte 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,

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

We know that $x \geq 0$. If the number n picked exceeds the number $N - r$ of failures, the difference $n - (N - r)$ must be successes. So, $\max(0, n - N + r) \leq x \leq \min(r, n)$ since we can't get more successes than the number available or the number of objects chosen.

Example 5.3.1. *In Lotto $6/49$ a player selects a set of six numbers (with no repeats) from the set $\{1, 2, \dots, 40\}$. In the lottery draw six numbers are selected at random. Find the probability function for X, the number from your set which are drawn.*

Solution. Think of your numbers as S (success) objects and the remainder as F (failure) objects. Then X has a Hypergeometric distribution with $N = 49, r = 6, n = 6$, so we have

$$f(x) = P(X = x) = \frac{\binom{6}{x}\binom{43}{6-x}}{\binom{49}{6}} \text{ for } x = 0, 1, \ldots, 6$$

■

5.4 Binomial Distribution

Physical Setup:
Suppose an "experiment" has two types of distinct outcomes – "success" (S) and "failure" (F), and let their probabilities be p (for S) and $1 - p$ (for F). Repeat the experiment n **independent** times. Let X be the number of successes obtained. Then X has a **Binomial distribution**. We write $X \sim Binomial(n, p)$ as a shorthand for "X is distributed according to a Binomial Distribution with n repetitions and probability p of success". The n individual experiments are often called "trials" or "Bernoulli trials" and the process is called a "Bernoulli process" or a "Bionomial Process".

5.4.1 Probabiity Function

There are $\dfrac{n!}{x!(n-x)!} = \binom{n}{x}$ different arrangements of x S's and $(n-x)$ F's over the n trials. the probability of each has p multiplied together x times and $(1-p)$ multiplied $(n-x)$ times in some order, since the trials are independent. Therefore,

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n$$

5.4.2 Computation

Many software packages and some calculators give Binomial probabilities. In R we use the function $dbinom(x, n, p)$ to compute $f(x)$ and $pbinom(x, n, p)$ to compute the corresponding cumulative distribution function $F(x) = P(X \leq x)$.

5.4.3 Remark – Comparison of Bionomial and Hypergeometric Distribution

These two are similar. The key difference is that the Binomial requires **independent repetitions with the same probability of S**, whereas the draws in Hypergeometric are made from **a fixed collection of objects without replacement**. The trials (draws) are therefore not independent.

For example, if there are $r = 10$ S objects and $N - r = 10$ F objects, then the probability of getting an S on draw two depends on what was obtained in draw one. If these draws had been made **with** replacement, however, they would be independent and we would use the Bionomial rather than Hypergeometric.

If N is large and the number n being drawn is relatively small in Hypergeometric setup, then we are unlikely to get the same object more than once even if we do replace it. So it makes little practical difference whether we draw with or without replacement, meaning that the Binomial and the Hypergeometric models should produce similar probabilities.

Example 5.4.1. *Suppose we have 15 cans of soup with no labels, but 6 are tomato and 9 are pea soup. We randomly pick 8 cans and open them. Find the probability three of them are tomato.*

Solution. The correct solution uses the Hypergeometric distribution. and is (with X = number of tomato soup cans picked)

$$P(X=3) = \frac{\binom{6}{3}\binom{9}{5}}{\binom{15}{8}} \approx 0.3916$$

If we incorrectly used the Binomial distribution, we would obtain

$$\binom{8}{3}(\frac{6}{15})^3(\frac{9}{15})^5 \approx 0.2787$$

However, if we had 1500 cans: 600 tomato and 900 pea, we are not likely to get the same can again even if we did replace each of the 8 cans after opening it. (Put another way, to probability we get a tomato soup on each pick is very close to 0.4, regardless of what the other picks give). The Hypergeometric probability gives

$$\frac{\binom{600}{3}\binom{900}{5}}{\binom{1500}{8}} \approx 0.2974$$

The Binomial probability,

$$\binom{8}{3}(\frac{600}{1500})^3(\frac{900}{1500})^5 \approx 0.2787$$

which is a very good approximation. ■

5.5 Negative Bionomial Distribution

5.5.1 Physical Setup

It is almost the same as Bionomial: an experiment (trial) has two distinct types of outcome (S and F) and is **repeated independently** with the same probability p, of success at each time. Continue doing the experiment until a specified number, k, of success have been obtained. Let X be the number of failures obtained before the k'th success. Then X has a negative Binomial Distribution. We can denote it as $X \sim NegativeBionmial(k,p)$.

5.5.2 Probability Function

In all there will be $x+k$ trials (x F's and k S's) and the last trial must be S. In the first $x+k-1$ trials we then need x failures and $(k-1)$ successes in any order. There are

$$\frac{(x+k-1)!}{x!(k-1)!} = \binom{x+k-1}{x}$$

different orders. Each order will have probability $p^k(1-p)^x$ since there must be x trials which are failures and k which are successes.

Note: You need to be careful to read how X is defined in a problem rather than mechanically "plugging in" numbers in the above formula for $f(x)$.

5.5.3 Comparison btw Bionomial and Negative Bionomial Distributions

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

Example 5.5.1. *A specific blood type T is 0.08 (8%). For blood donation purposes it is necessary to find 5 people with type T blood. If randomly selected individuals form the population are tested one after another, then*

1. What is the probability y persons have to be teseted to get 5 type T persons?

   Solution. Let a type T person as a success (S) and a non-type T as an F. Let Y = number of persons who have to be tested and let X = number of non-type T persons in order to get 5 S's. Then X has a Negative Binomial distribution with $k = 5$ and $p = 0.8$ and

   $$P(X = x) = f(x) = \binom{x+4}{x}(0.08)^5(0.92)^x \text{ for } x = 0, 1, 2, \ldots$$

   We are actually asked here about $Y = X + 5$. Thus,

   $$\begin{aligned} P(Y = y) &= P(X = y - 5) \\ &= f(y - 5) \\ &= \binom{y-1}{y-5}(0.08)^5(0.92)^{y-5} \text{ for } y = 5, 6, 7, \ldots \end{aligned}$$

   ■

2. What is the probability that over 80 people have to be tested?

   Solution.

   $$\begin{aligned} P(Y > 80) &= P(X > 75) \\ &= 1 - P(X \le 75) \\ &= 1 - \sum_{x=0}^{75} f(x) \\ &= 0.2235 \end{aligned}$$

   ■

5.6 Geometric Distribution

5.6.1 Physical Setup

Consider the Negative Binomial Distribution with $k = 1$. In this case we repeat independent Bernoulli trials with two types of outcomes (S and F) each time, and the same probability, p, of success each time until we obtain the first success. Let X be the number of failures obtained before the first sucess. We write $X \sim Geometric(p)$

5.6.2 Probability Function

There is only one arrangement with x failures followed by 1 success. This arrangement has probability

$$f(x) = P(X = x) = (1-p)^x p \text{ for } x = 0, 1, 2, \dots$$

Alternatively if we substitute $k = 1$ in the probability function for the Negative Binomial, we obtain

$$f(x) = \binom{x+1-1}{x} p^1 (1-p)^x = p(1-p)^x \text{ for } x = 0, 1, 2, \dots$$

which is the same.

Note:

- The Geometric Distribution involved a Geometric series.
- The Hypergeometric Distribution used the Hypergeometric Identity.
- Both the Binomial and Negative Binomial Distributions used the Bionmial Theorm.

Definition 5.6.1. ***Bernoulli Trials***

the Binomial, Negative Binomial and Geometric models all involve trials (experiments) which:

1. *are independent*
2. *have* 2 *distinct types of outcome* (S *and* F)
3. *have the same probability* p *of "success"* (S) *each time.*

5.7 Poisson Distribution from Binomial

It has the probability function of the form:

$$f(x) = P(X = x) = e^{-\mu} \frac{\mu^x}{x!} \text{ for } x = 0, 1, 2, \dots$$

where $\mu > 0$ is a **parameter** whose value depends on the setting for the model.

5.7.1 Physical Setup

It is as a limiting case of the Binomial distribution as $n \to \infty$ and $p \to 0$. In particular, we keep the product np fixed at some constant value, μ, while letting $n \to \infty$. This automacially makes $p \to 0$. Let us see what the limit of the Binomial probability function $f(x)$ in this case.

5.7.2 Probability Function

Since $np = \mu$, $p = \dfrac{\mu}{n}$ and for x

$$\begin{aligned}
f(x) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n^{(x)}}{x!}(\frac{\mu}{n})^x(1-\frac{\mu}{n})^{n-x} \\
&= \frac{\mu^x}{x!}\frac{\overbrace{n(n-1)(n-2)\cdots(n-x+1)}^{x \text{ terms}}}{n^x}(1-\frac{\mu}{n})^{n-x} \\
&= \frac{\mu^x}{x!}(\frac{n-1}{n})(\frac{n-2}{n})\cdots(\frac{n-x+1}{n})(1-\frac{\mu}{n})^n(1-\frac{\mu}{n})^{-x} \\
\lim_{n\to\infty} f(x) &= \frac{\mu^x}{x!}\underbrace{(1)(1)\cdots(1)}_{x \text{ terms}}e^{-\mu}(1)^{-x} \qquad \text{(since } e^k = \lim_{n\to\infty}(1+\tfrac{k}{n})^n\text{)} \\
&= \frac{\mu^x e^{-\mu}}{x!} \text{ for } x = 0, 1, 2, \ldots
\end{aligned}$$

Example 5.7.1. *There are* 200 *people at party. What is the probability that* 2 *of them were born on January* 1*?*

Solution. Assuming all days of the year are equally likely for a birthday (and ignore February 29) and that the birthdays are independent (e.g. no twins!) We can use Bionmial distribution with $n = 200$ and $p = \frac{1}{365}$ for X = number born on January 1, giving

$$f(2) = \binom{200}{2}\left(\frac{1}{365}\right)^2\left(1-\frac{1}{365}\right)^{198} = 0.0876767$$

Since n is large and p is close to 0, we can use the Poisson distribution to approximate this Binomial probability, with

$$\mu = np = \frac{200}{365}$$

giving

$$f(2) = \frac{(\frac{200}{365})^2 e^{-\frac{200}{365}}}{2!} = 0.086791$$

■

5.8 Poisson Distribution from Poisson Process

We derive the Poisson distribution as a model for the number of a certain kind of event or occurrence that occur at points in time or in space. To this end, we use the **"order" notation**

$$g(\Delta t) = o(\Delta t)$$

as $\Delta t \to 0$ to mean that the function g approaches 0 **faster** than Δt as Δt approaches 0, or that

$$\frac{g(\Delta t)}{\Delta t} \to 0 \text{ as } \Delta t \to 0$$

For example, $g(\Delta t) = (\Delta t)^2 = o(\Delta t)$, but $(\Delta t)^{\frac{1}{2}}$ is not $o(\Delta t)$.

5.8.1 Physical Setup

Consider a situation in which a certain type of event occurs at random points in time (or space) according to the following conditions:

1. **Independence**: the number of occurences in non-overlapping intervals are independent.
2. **Individuality**: for sufficiently short time periods of length Δt, the probability of 2 or more events occuring in the interval is close to zero, i.e. events occur singly not in clusters. More precisely, as $\Delta t \to 0$, the probability of two or more events in the interval of length Δt must go to zero faster than $\Delta t \to 0$ or
$$P(\text{2 or more events in } (t, t+\Delta t)) = o(\Delta t) \text{ as } \Delta t \to 0$$
3. **Homogeneity or Uniformity**: events occur at a uniform or homogeneous rate λ over time so that the probability of one occurrence in an interval $(t, t+\Delta t)$ is approximately $\lambda \Delta t$ for small Δt for any value t. More precisely,
$$P(\text{one event in } (t, t+\Delta t)) = \lambda \Delta t + o(\Delta t)$$

Theorem 5.8.1. *Suppose a process satisfies the three conditions above, then assume events occur t the average rate of λ per unit time. Let X be the number of times an event occur in a time period of t units, then $X \sim Poisson(\mu = \lambda t)$*

These three conditions together define a **Poisson Process**.
Let X be the number of event occurrences in a time period of length t. Then it can be shown that X has a Poisson distribution with $\mu = \lambda \cdot t$.

5.8.2 Probability Function

We are interested in time intervals of arbitrary length t, so as a temporary notation, let $f_t(x)$ be the probability of x occurences in a time interval of length t. We can determine what $f_t(x)$ is by relating $f_t(x)$ and $f_{t+\Delta t}(x)$.

To find $f_{t+\Delta t}(x)$ we note that for Δt small there are only two ways to get a total of x event occurrences:

1. there are x events by time t and no more from t to $t + \Delta t$
2. there are $x-1$ by time t and 1 more from t to $t + \Delta t$

$$f_{t+\Delta t}(x) \approx f_t(x)(1-\lambda \Delta t) + f_t(x-1)(\lambda \Delta t) + o(\Delta t)^2$$

Re-arranging we have:
$$\frac{f_{t+\Delta t}(x) - f_t(x)}{\Delta t} \approx \lambda[f_t(x-1) - f_t(x)] + o(\Delta t)$$

Taking the limit as $\Delta t \to 0$ we get

$$\frac{d}{dt} f_t(x) = \lambda[f_t(x-1) - f_t(x)]$$

We need to find $f_t(x)$. We can approach the problem by using Bionomial approximation. Suppose that the interval $(0, t)$ is divided into $n = \frac{t}{\Delta t}$ small subintervals of length Δt. The probability that an event falls into a subinterval is approximately $p = \lambda \Delta t$ provided that the length of interval is small. We can ignore the probability of two or more events fall in the same subinterval since $n \times o(\Delta t) \to 0$ as $\Delta t \to 0$. The "successes" are independent on the n different subintervals, so the total number of successes recorded as X is approximately $Binomial(n, p)$.

$$P(X = x) \approx \binom{n}{x} p^x (1-p)^{n-x} = \frac{n^{(x)} p^x}{x!} (1-p)^n (\frac{1}{1-p})^x$$

We can see that for fixed t, as $\Delta t \to 0$, $p = \lambda \Delta t \to 0$ and $n = \frac{t}{\Delta t} \to 0$, $p = \frac{\lambda t}{n}$. Therefore,

$$\lim_{\Delta t \to 0} (1-p)^n = \lim_{\Delta t \to 0} (1 - \frac{\lambda t}{n})^n = e^{-\lambda t}$$

For fixed x,

$$n^{(x)} p^x = n^{(x)} (\frac{\lambda t}{n})^x = (\lambda t)^x$$

This yields the approximation:

$$P(X = x) \approx \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

We can then confirm that

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

provides a solutionto the system with required initial conditions. If we let $\mu = \lambda t$, we cna re-write $f(x)$ as $f(x) = \dfrac{\mu^x e^{-\mu}}{x!}$, which is the Poisson distribution from 5.7.

Interpretation of μ and λ

1. μ refers to the **intensity** or **rate of occurrence** parameter for the events.
2. $\lambda t = \mu$ represents the **average number of occurrences** in t units of time.
3. IMPORTANT: the value of λ **depends on** the units used to measure time.

Example 5.8.1. *If phone calls arrive at store at an average rate of 20 per hour, then $\lambda = 20$ when time $t =$ an hour, and the average in 3 hours will be $3 \times 20 = 60$. However, if time is measured in minutes then $\lambda = \frac{20}{60} = \frac{1}{3}$; the average in 180 minutes is still $\frac{1}{3}(180) = 60$*

Example 5.8.2. *At a nuclear power station an average of 8 leaks of heavy water are reported per year. Find the probability of 2 or more leaks in 1 month, if leaks follow a Poisson process.*

Solution. A month is $\frac{1}{12}$ of a year. Let X be the number of leaks in one month. Then X has the Poisson distribution with $\lambda = 8$ and $t = \frac{1}{12}$, so $\mu = \lambda t = \frac{8}{12}$. Thus,

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - [f(0) + f(1)] \\ &= 1 - [\frac{(\frac{8}{12})^0 e^{-\frac{8}{12}}}{0!} + \frac{(\frac{8}{12})^1 e^{-\frac{8}{12}}}{1!}] \\ &= 0.1443 \end{aligned}$$

■

5.8.3 Random Occurrence of Events in Space

If X is 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 volume (or area), then X has a Poisson distribution with $\mu = \lambda v$.

For this model to be valid, it is assumed that the Poisson process conditions given previously apply here, with "time" replaced by "volume" or "area". Once again, note that teh value of λ **depends on** the units used to measure volume or area.

5.8.4 Distinguishing Poisson from Binomial and Other Distributions

To konw when to use the Poisson distsribution and when not to use it, we need to check the three conditions of a Poisson process. However, a quick decision can be made by asking ourselves the following questions:

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, the distribution might be Poisson, but is certainly not Binomial or Hypergeometric. e.g. the number of seeds which germinate out of a package of 25 does not have a Poisson distribution since we know in advance that $X \leq 25$.

2. Does it make sense to ask how often the event did not occur?
   If it does make sense, then the distribution is not Poisson. If it makes sence, then it might be Poisson. For instance, it does not make sense to ask how often a person did not hiccup during an hour. So the number of hiccups in an hour might have a Poisson distribution. It would certainly not be Binomial, Negative Binomial, or Hypergeometric.

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:

1. Find the probability a tree has > 3 ladybugs on it.

   Since the ladybugs are **randomly scattered**, we can use the Poisson distirbution to solve this. In this case, $\lambda = 6$ and $v = 1$ (i.e. any tree has a "volume" of one unit), so $\mu = 6$ and

$$\begin{aligned}P(X > 3) = 1 - P(X \leq 3) &= 1 - [f(0) + f(1) + f(2) + f(3)] \\ &= [\frac{6^0 e^{-6}}{0!} + \frac{6^1 e^{-6}}{1!} + \frac{6^2 e^{-6}}{2!} + \frac{6^3 e^{-6}}{3!}] \\ &= 0.8488\end{aligned}$$

2. When 10 trees are picked at random, what is the probability 8 of these trees have > 3 ladybugs on them?

   We can use the Binomial Distribution where "success" means to have > 3 ladybugs on a tree. We have $n = 10$, and

$$f(8) = \binom{10}{8} (0.8488)^8 (1 - 0.8488)^{10-8} = 0.2772$$

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

   We can use the Negative Binomial Distribution. We need the number of success, k, to be 5, and the number of failures to be $(x-5)$. Then,

$$f(x)=\binom{x-1}{4}(0.8488)^5(1-0.8488)^{x-5} \text{ for } x=5,6,7,\ldots$$

4. Find the probability a tree with >3 ladybugs on it has exactly 6.

   This is a conditional probability. Let A = {6 ladybugs} and B = {a tree with >3 ladybugs}. Then

$$P(A|B)=\frac{P(AB)}{P(B)}=\frac{\frac{6^6e^{-6}}{6!}}{0.8488}=0.1892$$

5. On 2 trees there are a total of t ladybugs. Find the probability that x of these are on the first of these 2 trees.

   This is also a conditional probability. Let A ={x on the first tree}, B ={$t-x$ on second tree}, and C ={2 trees has a total of t ladybugs}. Then we have

$$\begin{aligned}P(X=x)=P(A|C)&=\frac{P(AC)}{P(C)}=\frac{P(AB)}{P(C)}\\&=\frac{P(A)P(B)}{P(C)}\end{aligned}$$

   We can use Poisson Distribution at this point to calculate each, with $\mu=6\times2=12$in the denominator since there are 2 trees.

$$\begin{aligned}P(A|C)&=\frac{\left(\frac{6^xe^{-6}}{x!}\right)\left(\frac{6^{t-x}e^{-6}}{(t-x)!}\right)}{\frac{12^te^{-12}}{t!}}\\&=\frac{t!}{x!(t-x)!}\left(\frac{6}{12}\right)^x\left(\frac{6}{12}\right)^{t-x}\\&=\binom{t}{x}\left(\frac{1}{2}\right)^x\left(1-\frac{1}{2}\right)^{t-x}\quad \text{for } x=0,1,\ldots,\ t\end{aligned}$$

5.10 Summary of Probability Functions for Discrete Random Variables

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

Chapter 6

Computational Methods and the Statistical Software R

Not covered material.

Chapter 7

Expected Value and Variance

7.1 Summarizing Data on Random Variables

"Summary" materials are often more helpful than giving full details of every outcomes.

Definition 7.1.1. ***Frequency Distribution***

X	*Frequency*
1	*6*
2	*8*
3	*5*
4	*3*
5	*2*
6	*1*

Definition 7.1.2. ***Sample***
A set of observed outcomes $x_1, \ldots, x_n$ *for a random variable* X

Definition 7.1.3. ***Sample Mean***
We refer to the fact that this is the average for a particular sample.

Unless somebody deliberately "cooked" the study, we would not expect to get precisely the sample mean if we repeated it another time.

Definition 7.1.4. ***Median***
A value such that half the results are below it and half above it, when the results are arranged in numerical order.

Definition 7.1.5. ***Mode***
The value which occurs most often. There is no guarentee there will only be a single mode.

7.2 Expectation of a Random Variable

Example 7.2.1. *Referencing to the data we have for Frequency Distribution, we have*

$$\begin{aligned}\bar{x} &= \frac{(1\times 6)+(2\times 8)+(3\times 5)+(4\times 3)+(5\times 2)+(6\times 1)}{25} \\ &= (1)(\frac{6}{25})+(2)(\frac{8}{25})+(3)(\frac{5}{25})+(4)(\frac{3}{25})+(5)(\frac{2}{25})+(6)(\frac{1}{25}) \\ &= \sum_{x=1}^{6} x \times \textit{ fraction of times } x \textit{ occurs} \\ &= 2.60\end{aligned}$$

Now, suppose we know that the probability function of X is given by:

x	1	2	3	4	5	6
$f(x)$	0.30	0.25	0.20	0.15	0.09	0.01

*Using the frequency "definition" of probability, **in theory**, we would expect the mean to be*

$$(1)(0.30)+(2)(0.25)+(3)(0.20)+(4)(0.15)+(5)(0.09)+(6)(0.01) = 2.51$$

This "theoratical" mean is usually denoted by μ or $E(x)$

Definition 7.2.1. ***Expected Value***
Let X be a discrete random variable with $\text{range}(X)$ = *A and probability function $f(x)$. The **expected value** (also called the mean or the expectation) of X is given by*

$$E(X) = \sum_{x\in A} x f(x)$$

Example 7.2.2. *If you are playing a casino game in which X represents the amount you win in a single play, then $E(X)$ represents your average winnings (or losses!) per play.*

Theorem 7.2.1. *Let X be a discrete random variable with* $\text{range}(X)$ = *A and probability function $f(x)$. The **expected value** of some function $g(X)$ of X is given by*

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

Proof. To use Definition 7.2.1, we need to determine the expected value of the random variable $Y = g(X)$ by first finding the probability function of Y, say $f_Y(y) = P(Y = y)$ and then computing

$$E[g(X)] = E(Y) = \sum_{y\in B} y f_Y(y) \tag{7.1}$$

where $\text{range}(Y) = B$. Let $D_y = \{x : g(x) = y\}$ be the set of x values with a given value y for $g(x)$, then

$$f_Y(y) = P[g(X) = y] = \sum_{x\in D_y} f(x)$$

Substituting this in (7.1) we obtain

$$\begin{aligned}E[g(X)] &= \sum_{y\in B} y f_Y(y) \\ &= \sum_{y\in B} y \sum_{x\in D_y} f(x) \\ &= \sum_{y\in B}\sum_{x\in D_y} g(x)f(x) \\ &= \sum_{x\in A} g(x)f(x)\end{aligned}$$

□

Notes:

1. You can interpret $E[g(X)]$ as the average value of $g(X)$ in an infinite series of repetitions of the process where X is defined.
2. $E[g(x)]$ is also known as the "expected value" of $g(X)$. However, this value can be a value $g(X)$ never takes.
3. When calculating expectations, look at your answer to be sure it makes sence. Suppose for example that X takes values from 1 to 10. Then since

$$1 = \sum_{x=1}^{10}(1)P(X=x) \leq \sum_{x=1}^{10} xP(X=x) = E(X) \leq \sum_{x=1}^{10}(10)P(X=x) = 10(1) = 10$$

you should know you've made an arror if you get $E(X) > 10$ or $E(X) < 1$. In physical terms, $E(X)$ is the balance point for the probability histogram of $f(x)$

7.2.1 Linearity Properties of Expectations

1. For constants a and b,

$$E[ag(X)+b] = aE[g(x)] + b$$

Proof.

$$\begin{aligned}E[ag(X)+b] &= \sum_{\text{all } x}[ag(x)+b]f(x) \\ &= \sum_{\text{all } x}[ag(x)f(x)+bf(x)] \\ &= a\sum_{\text{all } x} g(x)f(x) + b\sum_{\text{all } x} f(x) \\ &= aEg(X) + b \quad \text{since } \sum_{\text{all } x} f(x) = 1\end{aligned}$$

□

2. Similarly for constants a and b and two functions g_1 and g_2, it is also easy to show

$$E[ag_1(X) + bg_2(X)] = aE[g_1(X)] + bE[g_2(X)]$$

7.3 Some Applications of Expectation

Be cautious that the expected value does NOT tell the whole story about a distribution. One investment could have a higher expected value than another but much much larger probability of large losses.

Example 7.3.1. ***Expected Winnings in a Lottery***
A small lottery sells 1000 tickets numbered 000, 001,…,999; the tickets cost $10 each. When all the tickets have been sold, the draw takes place. This consists of a single ticket from 000 to 999 being chosen at random. For ticket holders the prize structure as follows:

- *Your ticket is drawn – win $5000*
- *Your ticket has the same first two numbers as the winning ticket – win $100*
- *Your ticket has the same first number as the winning ticket – win $10*
- *All other cases – win nothing.*

Let the random variable X be the winnings from a given ticket. Find $E(X)$

Solution. The possible values for X are 0, 10, 100, 500 (dollars). First, we need to find the probability function for X. We find (make sure you can do this) that $f(x) = P(X = x)$ has values

$$f(0) = 0.9, \qquad f(10) = 0.09, \qquad f(100) = 0.009, \qquad f(5000) = 0.001$$

The expected winnings are thus the expected value of X, or

$$E(X) = \sum_{\text{all } x} x f(x) = 10 \times 0.09 + 100 \times 0.009 + 5000 \times 0.001 = \$6.80$$

Thus, the gross expected winnings per ticket are $6.80. However, since a ticket costs $10 your expected net winings are negative, -$3.20 (that is, an expected loss of $3.20) ■

Remark: For any lottery or game of chance the expected net winnings per play is a key value. A fair game is one for which this value is 0. Needless to say, casino games and lotteries are never fair: the expected net winnings for a player are always negative.

Example 7.3.2. ***Diagnostic Medical Tests***
Often there are cheaper, less accurate tests for diagnosing the presence of some conditions in a person, along with more expensive, accurate tests. Suppose we have two cheap tests and one expensive test, with the following characteristics. All three tests are positive if a person has the condition (there are no "false negatives"), but the cheap tests give "false positives". Let a person be chosen at random, and let D = {person has the condition}. For the three tests the probability of a false positive and cost are:

Test	*P(positive test|$\bar{D}$)*	*Cost (in dollars)*
1	*0.05*	*5*
2	*0.03*	*8*
3	*0*	*40*

We want to check a large number of people for the condition, and have to choose among three testing strategies:

1. *Use Test 1, followed by Test 3 if Test 1 is positive*
2. *Use Test 2, followed by Test 3 if Test 2 is positive*

3. *Use Test 3*

Determine the expeceted cost per person under each of strategies 1, 2, and 3. We will then choose the strategy with the lowest expecetd cost. It is known that about 0.001 of the population have the condition

$$P(D) = 0.001, \qquad P(\bar{D}) = 0.999$$

Assume that given D or $\bar{D}$, tests are independent of one another.

Solution. For a person tested chosen at random and tested, define the random variable X as follows:

$$\begin{aligned} X &= 1 \qquad \text{if the initial test is negative} \\ X &= 2 \qquad \text{if the initial test is positive} \end{aligned}$$

Let $g(x)$ be the total cost of testing the person, the expected cost per person is then

$$E[g(X)] = \sum_{x=1}^{2} g(x)f(x)$$

The probability function $f(x)$ for X and the function $g(x)$ differ for strategies 1, 2, 3. Consider for example strategy 1. Then

$$\begin{aligned} P(X = 2) &= P(\text{initial test positive}) \\ &= P(D) + P(\text{positive}|\bar{D})P(\bar{D}) \\ &= 0.001 + (0.005)(0.999) \\ &= 0.0510 \end{aligned}$$

The rest of the probabilities, associated with the values of $g(X)$ and $E[g(X)]$ are obtained below.

1. Strategy 1

$$\begin{aligned} f(2) &= 0.0510 \text{ obtained above} \\ f(1) &= P(X = 1) = 1 - f(2) = 1 - 0.0510 = 0.949 \\ g(1) &= 5 \qquad g(2) = 45 \\ E[g(X)] &= 5(0.949) + 45(0.0510) = \$7.04 \end{aligned}$$

2. Strategy 2

$$\begin{aligned} f(2) &= 0.001 + (0.03)(0.999) = 0.03097 \\ f(1) &= 1 - f(2) = 0.96903 \\ g(1) &= 8 \qquad g(2) = 48 \\ E[g(X)] &= 8(0.96903) + 48(0.03097) = \$9.2388 \end{aligned}$$

3. Strategy 3

$$\begin{aligned} f(2) &= 0.001, \qquad f(1) = 0.999 \\ g(2) &= g(1) = 40 \\ E[g(X)] &= \$40.00 \end{aligned}$$

Therefore, the cheapest strategy is strategy 1. ■

7.4 Means and Variance of Distributions

7.4.1 Expected value of a Binomial random variable

Let $X \sim Binomial(n, p)$. Find $E(X)$.

Solution.

$$\begin{aligned}\mu = E(X) &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}\end{aligned}$$

When $x = 0$ the value of the expression is 0. We can therefore begin our sum at $x = 1$. Provided $x \neq 0$, we can expand $x!$ as $x(x-1)!$. Therefore

$$\begin{aligned}\mu &= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)![(n-1)-(x-1)]!} p p^{x-1} (1-p)^{(n-1)-(x-1)} \\ &= np(1-p)^{n-1} \sum_{x=1}^{n} \binom{n-1}{x-1} \left(\frac{p}{1-p}\right)^{x-1}\end{aligned}$$

Let $y = x - 1$ in the sum to get

$$\begin{aligned}\mu &= np(1-p)^{n-1} \sum_{y=0}^{n-1} \binom{n-1}{y} \left(\frac{p}{1-p}\right)^{y} \\ &= np(1-p)^{n-1} \left(1 + \frac{p}{1-p}\right)^{n-1} \quad \text{by the Binomial Theorem} \\ &= np(1-p)^{n-1} \frac{(1-p+p)^{n-1}}{(1-p)^{n-1}} \\ &= np\end{aligned}$$

■

7.4.2 Expected Value of the Poisson random variable

Let X have a Poisson distribution where λ is the average rate of occurance and the time interval is of length t. Find $\mu = E(X)$

Solution. Since the probability function of X is

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

then

$$\mu = E(X) = \sum_{x=0}^{\infty} x \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

As in the Binomial example, we can eliminate the term when $x = 0$ and expand $x!$ to $x(x-1)!$ for $x = 1, 2, \ldots$ to obtain

$$\begin{aligned}
\mu &= \sum_{x=1}^{\infty} x \frac{(\lambda t)^x e^{-\lambda t}}{x(x-1)!} \\
&= \sum_{x=1}^{\infty} (\lambda t) e^{-\lambda t} \frac{(\lambda t)^{x-1}}{(x-1)!} \\
&= (\lambda t) e^{-\lambda t} \sum_{x=1}^{\infty} \frac{(\lambda t)^{x-1}}{(x-1)!} \\
&= (\lambda t) e^{-\lambda t} \sum_{y=0}^{\infty} \frac{(\lambda t)^y}{y!} \text{ letting } y = x-1 \text{ in the sum} \\
&= (\lambda t) e^{-\lambda t} e^{\lambda t} \text{ since } e^x = \sum_{y=0}^{\infty} \frac{x^y}{y!} \\
&= \lambda t
\end{aligned}$$

Note that we used the symbol $\mu = \lambda t$ earlier in connection with the Poisson model: this was because we knew (but couldn't show until now) that $E(X) = \mu$. ■

Note:
In Chapter 9 we will give a simpler method of finding the means of Hypergeometric and Negative Binomial distributions, in which

$$E(X) = \frac{nr}{N} \quad \text{Hypergeometric}$$

$$E(X) = \frac{k(1-p)}{p} \quad \text{Negative Binomial}$$

7.4.3 Variability

Definition 7.4.1. *Variance*
*The **variance** of a random variable X, denoted by $\mathrm{Var}(X)$ or by σ^2, is*

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

In other words, the variance is the average square of the distance from the mean. This turns out to be a very useful measure of the variability of X.

The basic definition of variance is often awkward to use for mathematical calculation of $\mathrm{Var}(X)$, whereas the following two results are often useful.

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

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

Proof. 1. Using properties of expected value,

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(X) = E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E(X^2) - 2\mu E(X) + \mu^2 \text{ since } \mu \text{ is constant} \\ &= E(X^2) - 2\mu^2 + \mu^2 \text{ since } E(X) = \mu \\ &= E(X^2) - \mu^2\end{aligned}$$

2. Since $X^2 = X(X-1) + X$,

$$\begin{aligned}E(X^2) - \mu^2 &= E[X(X-1) + X] - \mu^2 \\ &= E[X(X-1)] + E(X) - \mu^2 \\ &= E[X(X-1)] + \mu - \mu^2\end{aligned}$$

□

Formula (2) is most often used when there is an $x!$ term in the denominator of $f(x)$. Otherwise, formula (1) is generally easier to use.

Suppose the random variable X is the number of dollars that a person wins if they play a certain game. We notice that the units of measurement of $E(X)$ will also be dollars but units of measurement for $\mathrm{Var}(X)$ will be $(\text{dollars})^2$. We can regain the original units by taking the square root of $\mathrm{Var}(X)$. This is called the *standard deviation* of X, and is denoted by σ, or by $\mathrm{sd}(X)$.

Definition 7.4.2. ***Standard Deviation***
The standard deviation of a random variable X is

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

Both variance and standard deviation are commonly used to measure variability.

Example 7.4.1. *Suppose X is a random variable with probability function given by*

x	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*
$f(x)$	*0.07*	*0.10*	*0.12*	*0.13*	*0.16*	*0.13*	*0.12*	*0.10*	*0.07*	*1*

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

Solution.

$$\begin{aligned}\mu &= E(X) \\ &= 1(0.07) + 2(0.10) + 3(0.12) + 4(0.13) + 5(0.16) \\ &+ 6(0.13) + 7(0.12) + 8(0.10) + 9(0.07) \\ &= 5\end{aligned}$$

$E(X) = 5$ should be obvious by looking at the histogram. If a probability historgram is symmetric about the line $x = \mu$ then $E(X) = \mu$ with any calculation.

Without doing any calculations we also know that $\text{Var}(X) = \sigma^2 \leq 16$. This is because the possible values of X are {1,2,...,9} and so the maximum possible value for $(X-\mu)^2$ is $(9-5)^2$ or $(1-5)^2 = 16$. Therefore,

$$\begin{aligned}\text{Var}(X) = E[(X-5)^2] &= \sum_{x=1}^{9}(x-5)^2 P(X=x)\\ &\leq \sum_{x=1}^{9}(9-5)^2 P(X=x) = 16\sum_{x=1}^{9} P(X=x) = 16(1) = 16\end{aligned}$$

An expected value of a funciton, say $E[g(X)]$ is always somewhere between the minimum and the maximum value of the function $g(x)$ so in this case $0 \leq \text{Var}(X) \leq 16$. Since

$$\begin{aligned}E(X^2) &= (1)^2(0.07) + (2)^2(0.10) + (3)^2(0.12) + (4)^2(0.13) + (5)^2(0.16)\\ &+ (6)^2(0.13) + (7)^2(0.12) + (8)^2(0.10) + (9)^2(0.07)\\ &= 30.26\end{aligned}$$

Therefore,

$$\begin{aligned}\sigma^2 = \text{Var}(X) &= E(X^2) - \mu^2\\ &= 30.26 - (5)^2\\ &= 5.26\end{aligned}$$

and

$$\begin{aligned}\sigma &= \sqrt{\text{Var}(X)}\\ &= \sqrt{5.26}\\ &= 2.2935\end{aligned}$$

■

Example 7.4.2. ***Variance of Binomial random variable***
Let $X \sim Binomial(n,p)$. *Find* $\text{Var}(X)$.

Solution. The probability function for X is

$$f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \text{ for } x = 0,1,\ldots,n$$

so we'll use formula (2) above,

$$E[X(X-1)] = \sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

If $x = 0$ or $x = 1$ the value of the term is 0, so we can begin summing at $x = 2$. For $x \neq 0$ or 1, we can expand the $x!$ as $x(x-1)(x-2)!$ Therefore,

$$E[X(X-1)] = \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!}p^x(1-p)^{n-x}$$

Now re-group to fit the Binomial Theorem, since that was the summation technique used to show $\sum f(x) = 1$ and to derive $\mu = np$

$$
\begin{aligned}
E[X(X-1)] &= \sum_{x=2}^{n} \frac{n(n-1)(n-2)!}{(x-2)![(n-2)-(x-2)]!} p^2 p^{x-2}(1-p)^{(n-2)-(x-2)} \\
&= n(n-1)p^2(1-p)^{n-2} \sum_{x=2}^{n} \binom{n-2}{x-2} \left(\frac{p}{1-p}\right)^{x-2}
\end{aligned}
$$

Let $y = x - 2$ in the sum, giving

$$
\begin{aligned}
E[X(X-1)] &= n(n-1)p^2(1-p)^{n-2} \sum_{y=0}^{n-2} \binom{n-2}{y} \left(\frac{p}{1-p}\right)^{y} \\
&= n(n-1)p^2(1-p)^{n-2} \left(1 + \frac{p}{1-p}\right)^{n-2} \\
&= n(n-1)p^2(1-p)^{n-2} \frac{(1-p+p)^{n-2}}{(1-p)^{n-2}} \\
&= n(n-1)p^2
\end{aligned}
$$

Then

$$
\begin{aligned}
\sigma^2 &= E[X(X-1)] + \mu - \mu^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= n^2p^2 - np^2 + np - n^2p^2 \\
&= np(1-p)
\end{aligned}
$$

■

Example 7.4.3. ***Variance of Poisson random variable***
Suppose X has a Poisson(μ) distribution. Find $\mathrm{Var}(X)$.

Solution. The probability function for X is

$$
f(x) = \frac{\mu^x e^{-\mu}}{x!} \text{ for } x = 0, 1, 2, \ldots
$$

from which we obtain

$$
\begin{aligned}
E[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1) \frac{\mu^x e^{-\mu}}{x!} \\
&= \sum_{x=2}^{\infty} x(x-1) \frac{\mu^x e^{-\mu}}{x(x-1)(x-2)!} \text{ setting the lower limit to 2 and expanding } x! \\
&= \mu^2 e^{-\mu} \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!}
\end{aligned}
$$

Let $y = x - 2$ we get

$$\begin{aligned} E[X(X-1)] &= \mu^2 e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^y}{y!} = \mu^2 e^{-\mu} e^{\mu} = \mu^2 \\ \sigma^2 &= E[X(X-1)] + \mu - \mu^2 \\ &= \mu^2 + \mu - \mu^2 = \mu \end{aligned}$$

■

For the Poisson distribution, the variance equals the mean

7.4.4 Properties of Mean and Variance

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

$$\mu_Y = E(Y) = aE(X) + b = a\mu_X + b$$

and

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

where $\mu_X = E(X),\ \sigma_X^2 = \mathrm{Var}(X),\ E(Y) = \mu_Y,\ \mathrm{Var}(Y) = \sigma_Y^2$.

Proof. We already showed that $E(Y) = E(aX + b) = a\mu_X + b = \mu_Y$. Then

$$\begin{aligned} \sigma_Y^2 &= E\left[(Y - \mu_Y)^2\right] = E\{[(aX + b) - (a\mu_X + b)^2]\} \\ &= E\left[(aX - a\mu_X)^2\right] = E\left[a^2 (X - \mu_X)^2\right] \\ &= a^2 E\left[(X - \mu_X)^2\right] = a^2 \sigma_X^2 \end{aligned}$$

□

Adding a constant, b, to all values of X has no effect on the amount of variability. So it makes sence that $\mathrm{Var}(aX + b)$ doesn't depend on the value of b. Also since variance is in squared units, multiplication by a constant results in multiplying the variance by the constant squared.

Chapter 8

Continuous Random Variables

8.1 General Terminology and Notation

For **continuous** random variables teh range (set of possible values) is an interval (or a collection of intervals) on a real line.

8.1.1 Cumulative Distribution Function

Properties of a cumulative distribution function are the same for continuous variables as for discrete variables.

1. $F(x)$ is defined for all real x.
2. $F(x)$ is a non-decreasing function of x for all real x.
3. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$
4. $P(a < X \leq b) = F(b) - F(a)$

Note that, as indicated before, for a continuous random variable, we have

$$0 = P(X = a) = \lim_{\epsilon\to 0} P(a - \epsilon < X \leq a) = \lim_{\epsilon\to 0} F(a) - F(a - \epsilon)$$

This means that $\lim_{\epsilon\to 0} F(a - \epsilon) = F(a)$ or that the distribution function F is a continuous function (in the sense of continuity in calculus). Also, since the probability is 0 at each point:

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

For a discrete random variable, each of these 4 probabilities could be different. For the continuous distributions in this chapter, we do not worry about whether intervals are open, closed, or half-open since the probability of these intervals are the same.

8.1.2 Probability Density Function

To develop an intuitive picture of which values of x are more likely, and which are less likely, we have the probability X lies in the interval

$$P(x \leq X \leq x + \Delta x) = F(x + \Delta x) - F(x)$$

Definition 8.1.1. *The* ***probability density function*** *(p.d.f) $f(x)$ for a continuous random variable X is the derivative*

$$f(x) = \frac{dF(x)}{dx}$$

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

If the derivative of F does not exist at $x = a$ we usually define $f(a) = 0$ for convenience. Assume $f(x)$ is a continuous function of x at all points for which $0 < F(x) < 1$

Properties of a probability density function

1. $P(a \le X \le b) = F(b) - F(a) = \int_a^b f(x)dx$ (This follows from the definition of $f(x)$)
2. $f(x) \ge 0$ (since $F(x)$ is non-decreasing, its derivative is non-negative)
3. $\int_{-\infty}^{\infty} f(x)dx = \int_{\text{all } x} f(x)dx = 1$ (This is because $P(-\infty \le X \le \infty) = 1$)
4. $F(x) = \int_{-\infty}^{x} f(u)du$ (This is just property 1 with $a = -\infty$)

To see that $f(x)$ represents the relative likelihood of different outcomes, we note that for Δx small,

$$P\left(x - \frac{\Delta x}{2} \le X \le x + \frac{\Delta x}{2}\right) = F\left(x + \frac{\Delta x}{2}\right) - F\left(x - \frac{\Delta x}{2}\right) \approx f(x)\Delta x$$

Example 8.1.1. *Consider the following spinner example, where*

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

Thus, the probability density function is $f(x) = F'(x)$, or

$$f(x) = \frac{1}{4} \text{ if } 0 < x < 4$$

and outside this interval the probability density function is defined to be 0.

Remark: Continuous probability distributions are, like discrete distributions, mathematical **models**. Thus, the Uniform distribution assumed for the spinner above is a model, and it sesems likely it would be a good model for many real spinners.

Remakr: It may seem paradoxical that $P(X = x) = 0$ for a continuous random variable and yet we record the outcomes $X = x$ in real "experiments" with continuous variables. The catch is that all measurements have finite precision. They are in effect discrete. In measurements we are actually observing something like

$$P(x - 0.5\Delta x \le X \le x + 0.5\Delta x)$$

where Δx may be very small, but not zero. The probability of this outcome is **NOT** zero, it is (approximately) $f(x)\Delta$

Example 8.1.2. *Let X be a continuous random variable with probability density function*

$$f(x) = \begin{cases} kx^2 & 0 < x \le 1 \\ k(2 - x) & 1 < x < 2 \\ 0 & \textit{otherwise} \end{cases}$$

Find:

1. *the constant k*
2. *the cumulative distribution function* $F(x) = P(X \leq x)$
3. $P(0.5 < X < 1.5)$

Solution. 1. When finding teh area of region bounded by different functions we split the integral into pieces.

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f(x)dx \\
&= \int_{-\infty}^{0} 0dx + \int_{0}^{1} kx^2dx + \int_{1}^{2} k(2-x)dx + \int_{2}^{\infty} 0dx \\
&= 0 + k\int_{0}^{1} x^2dx + k\int_{1}^{2}(2-x)dx + 0 \\
&= k\frac{x^3}{3}|_0^1 + k\left(2x - \frac{x^2}{2}|_1^2\right) \\
&= \frac{5k}{6} \text{ and therefore } k = \frac{6}{5}
\end{aligned}$$

2. let us start with the easy pieces (which are unfortunately often left out) first:

$F(x) = P(X \leq x) = 0$ if $x \leq 0$
$F(x) = P(X \leq x) = 1$ if $x \geq 2$ since the probability density function equals 0 for all $x \geq 2$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(z)dz = 0 + \int_{-\infty}^{x} \frac{6}{5}z^2dz = \frac{6}{5}\frac{x^3}{3}|_0^x = \frac{2x^3}{5} \text{ if } 0 < x < 1$$

$$\begin{aligned}
F(x) = P(X \leq x) &= 0 + \int_{0}^{1} \frac{6}{5}z^2dz + \int_{1}^{x} \frac{6}{5}(2-z)dz = \frac{6}{5}\frac{x^3}{3}|_0^1 + \frac{6}{5}(2x - \frac{x^2}{2})|_1^x \\
&= \frac{12x - 3x^2 - 7}{5} \text{ if } 1 < x < 2
\end{aligned}$$

Therefore,

$$F(x) = P(X \leq x) = \begin{cases} 0 & x \leq 0 \\ \frac{2x^3}{5} & 0 < x \leq 1 \\ \frac{12x-3x^2-7}{5} & 1 < x < 2 \\ 1 & x \geq 2 \end{cases}$$

As a rough check, $F(x)$ should have the same value as we approach each boudary point from above and from below.
For example,

$$\begin{aligned}
&\text{as } x \to 0^+, \frac{2x^3}{5} \to 0 \\
&\text{as } x \to 1^-, \frac{2x^3}{5} \to \frac{2}{5} \\
&\text{as } x \to 1^+, \frac{12x - 3x^2 - 7}{5} \to \frac{2}{5} \\
&\text{as } x \to 2^-, \frac{12x - 3x^2 - 7}{5} \to 1
\end{aligned}$$

This quick check won't prove your answer is right, but will detect many careless errors.

3.
$$
\begin{aligned}
P(0.5 < X < 1.5) &= \int_{0.5}^{1.5} f(x)dx = F(1.5) - F(0.5) \\
&= \frac{12(1.5) - 3(1.5)^2 - 7}{5} - \frac{2(0.5^3)}{5} = 0.8
\end{aligned}
$$

■

Definition 8.1.2. ***Quantiles and Percentiles***
Suppose X is a continuous random variable with cumulative distribution fucntion $F(x)$. The pth quantile of X (or the pth quantile of the distribution) is the value $q(p)$, such that

$$P[X \leq q(p)] = p$$

or

$$F(q(p)) = p$$

. 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.

Example 8.1.3. *For the exmaple above, find*

1. *the 0.4 quantile (40th percentile) of the distribution*
2. *the median of the distribution*

Solution. 1. Since $F(1) = 0.4$, the 0.4 quantile is equal to 1.

2. The median is the solution to

$$F(x) = \frac{12x - 3x^2 - 7}{5} = 0.5$$

or

$$24x - 6x^2 - 19 = 0$$

which has two solutions. Since $F(1) = 0.4$ we know that the mexian lie betwene 1 and 2 and we choose the solution $x \approx 1.087$. The median is approximately equal to 1.087.

■

8.1.3 Defined Variables or Change of Variable

Sometimes we want to find the probability density function or comulative distribution function for some other random variable Y which is a function of X. It is based on the fact that the cumulative distribution function $F_Y(y)$ for Y equals $P(Y \leq y)$, and this can be rewritten in terms of X since Y is a function of X. Thus:

1. Write the comulative distribution function of Y as a function of X
2. Use $F_X(x)$ to find $F_Y(y)$. Then if you want the probability density function $f_Y(y)$, you can differentiate the expression for $F_Y(y)$.
3. Find the range of values of y.

Example 8.1.4. *In the earlier spinner example,*

$$f(x) = \begin{cases} \frac{1}{4} & 0 < x \leq 4 \\ 0 & \textit{otherwise} \end{cases}$$

and

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

Find the probability density function of $Y = X^{-1}$

Solution. Step 1 from above becomes:

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

For step (2), we can substitute $\frac{1}{y}$ in place of x in $F_X(x)$) giving:

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

and then differentiate to obtain the probability density function

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{4y^2} \text{ for } y \geq \frac{1}{4}$$

(Note that as x goes from 0 to 4, $y = \frac{1}{x}$ goes between ∞ and $\frac{1}{4}$).

Alternatively, and a little more generally, we can use the chain rule:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy}[1 - F_X(y^{-1})] \\ &= -f_X(y^{-1}) \frac{d}{dy}(y^{-1}) \text{ since } \frac{d}{dx} F_X(x) = f_X(x) \\ &= -f_X(y^{-1})(-y^{-2}) = \frac{1}{4}(-y^{-2}) \\ &= \frac{1}{4y^2} \text{ for } y \geq \frac{1}{4} \end{aligned}$$

■

8.1.4 Expectation, Mean, and Variance for Continuous Random Variables

Definition 8.1.3. *When* X *is a continuous random variable we define*

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

Example 8.1.5. *For the earlier spinner exmaple,*

$$f(x) = \begin{cases} \frac{1}{4} & 0 < x \leq 4 \\ 0 & \textit{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned}
\mu = E(X) &= \int_{-\infty}^{\infty} x f(x) dx = 0 + \int_0^4 x \frac{1}{4} dx + 0 = \frac{1}{4}\left(\frac{x^2}{2}\right)|_0^4 = 2 \\
E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = 0 + \int_0^4 x^2 \frac{1}{4} dx + 0 = \frac{1}{4}\left(\frac{x^3}{3}\right)|_0^4 = \frac{16}{3} \\
\sigma^2 = \text{Var}(X) &= E(X^2) - \mu^2 = \frac{16}{3} - (2)^2 = \frac{4}{3}
\end{aligned}$$

Example 8.1.6. *Let X have probability density funciton*

$$f(x) = \begin{cases} \frac{6x^2}{5} & 0 < x \leq 1 \\ \frac{6}{5}(2-x) & 1 < x < 2 \\ 0 & \textit{otherwise} \end{cases}$$

Then,

$$\begin{aligned}
\mu = E(X) &= \int_{-\infty}^{\infty} x f(x) dx = 0 + \int_0^1 x \frac{6}{5} x^2 dx + \int_1^2 x \frac{6}{5}(2-x) dx + 0 \\
&= \frac{6}{5}\left[\frac{x^4}{4}|_0^1 + \left(x^2 - \frac{x^3}{3}|_1^2\right)\right] = \frac{11}{10} = 1.1 \\
E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = 0 + \int_0^1 x^2 \frac{6}{5} x^2 dx + \int_1^2 x^2 \frac{6}{5}(2-x) dx + 0 \\
&= \frac{6}{5}\left[\frac{x^5}{5}|_0^1 + 2\left(\frac{x^3}{3}\right)|_1^2 - \frac{x^4}{4}|_1^2\right] = \frac{67}{50} \\
\sigma^2 = \text{Var}(X) &= E(X^2) - \mu^2 = \frac{67}{50} - \left(\frac{11}{10}\right)^2 \\
&= \frac{13}{100} \\
&= 0.13
\end{aligned}$$

8.2 Continuous Uniform Distribution

Physical Setup:

Supopse X takes values in some interval $[a, b]$ (it doesn't actually matter whether interval is open or closed) with all subintervals of a fixed length being equally likely. Then X has a **continuous Uniform distribution**, and we denote $X \sim U(a, b)$

Illustrations:

(1) In the spinner example $X \sim U(0, 4)$

(2) Computers can generate a random number X which appears as though it is drawn from the distribution $U(0, 1)$. This is the starting point for many computer simulations for random processors

8.2.1 P.D.F and C.D.F

Since all points are equally likely, the probability density function must be a constant $f(x) = k$ for all $a \le x \le b$ for some constant k. To make $\int_a^b f(x)dx = 1$, we require $k = \frac{1}{b-a}$. Therefore, the probability density function is

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

It is easy to varify that $\int_{-\infty}^{\infty} f(x)dx = 1$ since $(b-a)(\frac{1}{b-a}) = 1$

The comulative distribution function is

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

8.2.2 Mean and Variance

The mean can easily be determined since the graph of the pdf is symmetric about the line $x = \frac{a+b}{2}$. Since the integral $\int_a^b xdx$ exists therefore $E(X)$ exists and by symmetry $E(X) = \frac{a+b}{2}$

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)dx = \int_a^b x^2 \frac{1}{b-a}dx = \frac{1}{b-a}(\frac{x^3}{3}|_a^b) = \frac{b^3-a^3}{3(b-a)} \\ &= \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3} \end{aligned}$$

and therefore

$$\begin{aligned} \sigma^2 = \mathrm{Var}(X) = E(X^2) - \mu^2 &= \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{4b^2+4ab+4a^2-3b^2-6ab-3a^2}{12} = \frac{b^2-2ab+a^2}{12} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

In summary,

$$\text{If } X \sim U(a,b) \text{ then } E(X) = \frac{a+b}{2},\ \sigma^2 = \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Example 8.2.1. *Suppose X has the continuous probability density function*

$$f(x) = 0.1e^{-0.1x},\ x > 0$$

and zero otherwise (This is called the Exponential distribution and is dicussed in the next section. It is used in areas such as queueing theory and reliability.) We will show that hte new random variable

$$Y = e^{-0.1X}$$

has a $U(0,1)$ distribution.

Solution.

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

Since for $x > 0$

$$F_X(x) = \int_0^x 0.1e^{-0.1u}du = 1 - e^{-0.1x}$$

we have

$$F_Y(y) = 1 - [1 - e^{-0.1(-10\ln(y))}] = y, \ 0 < y < 1$$

The range of Y is $(0, 1)$ since $X > 0$. Thus,

$$f_Y(y) = \begin{cases} \frac{d}{dy}F_Y(y) = 1 & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

which implies $Y \sim U(0, 1)$ ■

8.2.3 Pseudo-random number generator

Many computer software systems use this way to generate random numbers because they are based on deterministic algorithms. In addition they give obervations Y that have finite precision so they cannot be **exactly** like continuous $U(0, 1)$ random variables. However, good generators give Y's that appear indistinguishable in most ways from $U(0, 1)$ random variables. Given such a generator, we can also simulate random variables X with the Exponential distribution above by the following algorithm:

1. Generate $Y \sim U(0, 1)$ using the computer random number generator
2. Compute $X = -10\ln(Y)$

Then X has the desired distribution. This is a particular case of a method decribed in Section 8.4 for generating random variables from a general distribution.

8.3 Exponential Distribution

The continuous random variable X is said to have an **Exponential distribution** if its probability density function is of the form

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda > 0$ is a real parameter value.

Physical setup:

In a Poisson process for events in time let X be the length of time we wait for the first event occurrence. We'll show that X has a exponential distribution (Recall that the **number** of occurrences in a fixed time has a Poisson distribution. The difference between the Poisson andn Exponential distribution lies in what is being measured).

Illustrations:

(1) The length of time X we wait with a Geiger counter until the emission of a radioactive particle is recorded following an Exponential distribution

(2) The length of time between phone calls to a fire station (assuming calls follow a Poisson distribution) follows an Exponential Distribution

8.3.1 Derivation of the PDF and the CDF

For $x > 0$

$$\begin{aligned} F(x) = P(X \le x) &= P(\text{time to } 1^{st} \text{ occurrence } \le x) \\ &= 1 - P(\text{time to } 1^{st} \text{ occurrence } > x) \\ &= 1 - P(\text{no occurrence in the interval } (0, x)) \end{aligned}$$

We have now expressed $F(x)$ in terms of number of occurrences in a Poisson process by time x, but the number of occurrences has a Poisson distribution with mean $\mu = \lambda x$ where λ is the average rate of occurrence. Therefore,

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

Also since $\frac{d}{dx}(1 - e^{-\lambda x}) = \lambda e^{-\lambda x}$ we have

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

which is the formula we gave above.

Alternate Form:

It is common to use the parameter $\theta = \frac{1}{\lambda}$ in the Exponential distribution (We'll see below that $\theta = E(X)$) This gives

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

and

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

We write $X \sim Exponential(\theta)$

From the graph of the Exponential distribution, we can see that it is positively skewed (skewed to the right) or have a long right tail.

8.3.2 Mean and Variance

We use **gamma function** to solve for μ and σ^2, which extends the notion of factorials beyond the integers to the positive real numbers.

Definition 8.3.1. ***The Gamma Function***

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$$

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

Note that α is 1 more than the power of y in the integrand. For instance,

$$\Gamma(5) = \int_0^\infty y^4 e^{-y}dy$$

There are three properties of gamma functions we can use:

1. $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ for $\alpha > 1$

   Proof. Using integration by parts:

   $$\int_{-\infty}^\infty y^{\alpha-1}e^{-y}dy = -\lim_{y\to\infty} y^{\alpha-1}e^{-y} + (\alpha-1)\int_0^\infty y^{\alpha-2}e^{-y}dy$$

   and provided that $\alpha > 1$, $\lim_{y\to\infty} y^{\alpha-1}e^{-y} = 0$, therefore,

   $$\int_0^\infty y^{\alpha-1}e^{-y}dy = (\alpha-1)\int_0^\infty y^{\alpha-2}e^{-y}dy = (\alpha-1)\Gamma(\alpha-1)$$

   □

2. $\Gamma(\alpha) = (\alpha-1)!$ if α is a positive integer

   Proof. It is easy to show that $\Gamma(1) = 1$. Using property 1 repeatedly, we obtain

   $$\begin{aligned}\Gamma(2) &= 1\Gamma(1) = 1\\ \Gamma(3) &= 2\Gamma(2) = 2!\\ \Gamma(4) &= 3\Gamma(2) = 3!\\ &\vdots\end{aligned}$$

   In general,

   $$\Gamma(n+1) = n! \text{ for } n = 0, 1, \ldots$$

   □

3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

   Proof. (This can be proved using double integration) □

Returning to the Exponential distribution we have:

$$\begin{aligned}\mu &= \int_{-\infty}^\infty xf(x)dx = \int_0^\infty x\frac{1}{\theta}e^{-\frac{x}{\theta}}dx \text{ let } y = \frac{x}{\theta} \text{ with } dx = \theta dy\\ &= \int_0^\infty ye^{-y}\theta dy = \theta\int_0^\infty y^1 e^{-y}dy = \theta\Gamma(2) = \theta(1!)\\ &= \theta\end{aligned}$$

Note: read questions carefully. If you are given the average **rate** of occurrences is a Poisson process, then this is the parameter λ. If you are given the average waiting **time** for an occurrence, then this is the parameter θ.

To get $\sigma^2 = \mathrm{Var}(X)$, we first find

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} x^2 \frac{1}{\theta} e^{-\frac{x}{\theta}} dx \text{ let } y = \frac{x}{\theta} \\ &= \int_0^{\infty} \theta^2 y^2 \frac{1}{\theta} e^{-y} \theta dy = \theta^2 \int_0^{\infty} y^2 e^{-y} dy = \theta^2 \Gamma(3) = \theta^2 (2!) \\ &= 2\theta^2 \end{aligned}$$

Then,

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

In summary,

$$\text{If } X \sim Exponential(\theta) \text{ then } E(X) = \theta \text{ and } \mathrm{Var}(X) = \theta^2$$

Example 8.3.1. *Suppose buses arrive at a bus stop according to a Poisson process with an average of 5 buses per hour. ($\lambda = 5/hr$ so $\theta = \frac{1}{5}hr$ or 12 min)*

Find the probability:

(1) you have to wait longer than 15 minutes for a bus

(2) you have to wait more than 15 minutes longer, having already waited for 6 minutes

Solution. (1)

$$P(X > 15) = 1 - P(X \leq 15) = 1 - F(15) = 1 - (1 - e^{-\frac{15}{12}}) = e^{-1.25} = 0.2865$$

(2) If X is the total waiting time, the question asks for the conditional probability

$$\begin{aligned} P(X > 21 | X > 6) &= \frac{P(X > 21 \cap X > 6)}{P(X > 6)} \\ &= \frac{P(X > 21)}{P > 6} \\ &= \frac{1 - (1 - e^{-\frac{21}{12}})}{1 - (1 - e^{-\frac{6}{12}})} \\ &= \frac{e^{-\frac{21}{12}}}{e^{-\frac{6}{12}}} \\ &= e^{-1/25} = 0.2865 \end{aligned}$$

Surprisingly, the fact that you've already waited for 6 minutes doesn't seem to matter.

■

8.3.3 Memoryless Property of the Exponential Distribution

The example above illustrates the "memoryless property" of the Exponential distribution:

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

In other words for a Poisson process, given that you've 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

8.4 A Method for Computer Generation of Random Variables

Most computer software has a built-in "pseudo-random number" generator that will simulate observations U from a $U(0, 1)$ distribution, or at least a reasonable approximation to this Uniform distribution.

If we wish a random variable with a non-Uniform distribution, the standard approach is to take a suitable function of U. We can do this through the inverse cumulative distribution function.

Definition 8.4.1. ***Inverse Cumulative Distribution Function***
For an arbitrary cumulative distribution function $F(x)$, we define

$$F^{-1}(y) = \min\{x : F(x) \geq y\}$$

This is a real inverse, in other words,

$$F(F^{-1}(y)) = F^{-1}(F(y)) = y$$

when the cumulative distribution is continuous and strictly increasing.

However, in the more general case of a ***discontinuous non-decreasing*** *cumulative distribution function (such as the cdf of a discrete distribution), the function still has at least some of the properties of an inverse.*

F^{-1} is useful for generating random variables having cumulative distribution function $F(x)$ from U, a Uniform random variable on the interval $[0, 1]$

Theorem 8.4.1. *If F is an arbitrary cumulative distribution function and U is Uniform on $[0, 1]$ then the random variable defined by*

$$X = F^{-1}(U)$$

has the cumulative distribution function $F(x)$

Proof. The proof is a consequence of the fact that

$$[U < F(x)] \subset [X \leq x] \subset [U \leq F(x)] \text{ for all } x$$

Taking probabilities on all sides of this, and using the fact that

$$P[U \leq F(x)] = P[U < F(x)] = F(x)$$

we discover that $P(X \leq x) = F(x)$ □

The relation $X = U^{-1}$ implies that $F(X) \geq U$ and for any point $z < X, F(z) < U$.

Example 8.4.1. ***A Geometric random number generator***

For the Geometric distribution, the cumulative distribution function is given by

$$F(x) = 1 - (1-p)^{x+1} \text{ for } x = 0, 1, 2, \ldots$$

Then if U is a Uniform random number in the interval $[0, 1]$, we see an integer X such that

$$F(X-1) < U \leq F(x)$$

You should confirm that this is the value of X at which the above horizontal line strikes the graph of the cdf, and solving these inequalities give

$$\begin{aligned}
1-(1-p)^X &< U \leq 1-(1-p)^{X+1} \\
(1-p)^X &> 1-U \geq (1-p)^{X+1} \\
X\ln(1-p) &> \ln(1-U) \geq (X+1)\ln(1-p) \\
X &< \frac{\ln(1-U)}{\ln(1-p)} \leq X+1
\end{aligned}$$

so we compute the value of $\frac{\ln(1-U)}{\ln(1-p)}$ and round down to the next lower integer.

8.5 Normal Distribution

Physical Setup:

A random variable X has a Normal distribution if it has probability density function of the form

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

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

$$E(X) = \mu$$

and

$$\text{Var}(X) = \sigma^2$$

for this distribution. We write $X \sim N(\mu, \sigma^2)$ to denote that X has a normal distribution with mean μ and variance σ^2 (standard deviation σ) The Normal distribution is the most widely used distribution in probability and statistics.

Illustrations:

- Heights or weights of males (or of females) in large populations tend to follow a Normal distribution
- The logarithms of stock prices are often assumed to have a Normal distribution

The graph of pdf $f(x)$ is symmetric about the line $x = \mu$. The shape of the pdf is often termed a **"bell shape" or "bell curve"**.

We can show that $f(x)$ integrates to 1:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx && (\text{let } z = \tfrac{x-\mu}{\sigma}) \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz && (\text{let } y = \tfrac{1}{2}z^2 \text{ and } dz = \tfrac{dy}{\sqrt{2}y^{\frac{1}{2}}}) \\
&= 2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y} \frac{dy}{\sqrt{2}y^{\frac{1}{2}}} \\
&= \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} y^{-\frac{1}{2}} e^{-y} dy \\
&= \frac{1}{\sqrt{\pi}} \Gamma(\frac{1}{2}) && (\text{where } \Gamma \text{ is the gamma function}) \\
&= 1 && (\text{since } \Gamma(\tfrac{1}{2}) = \sqrt{\pi})
\end{aligned}$$

8.5.1 Cumulative Distribution Function

The cdf of the Normal distribution $N(\mu, \sigma^2)$ is

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

In the statistical packages R we get $F(x)$ using the function $pnorm(x, \mu, \sigma)$. Before computers, people needed to produce tables of probabilities $F(x)$ by numerical integration, using mechanical calculators. Fortunately it is necessary to do this only for a single Normal distribution: the one with $\mu = 0$ and $\sigma = 1$. This is called the **"standard" Normal distribution** and is denoted $N(0, 1)$

It is easy to see that if $X \sim N(\mu, \sigma^2)$ then the "new" random variable $Z = \frac{X-\mu}{\sigma}$ is distributed as $Z \sim N(0, 1)$ (Use the change of variables methods in 8.1) We'll use this result to compute probabilities for X, and to show that $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$

8.5.2 Mean

If $f(x)$ is an odd function, then $\int_{-\infty}^{\infty} f(x)dx = 0$, provided the integral exists. Consider:

$$E(X - \mu) = \int_{-\infty}^{\infty} (x - \mu) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

Let $y = x - \mu$ then,

$$E(Y) = \int_{-\infty}^{\infty} y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y}{\sigma})^2} dy$$

where the integrand is an odd function so that $E(Y) = 0$, but since

$$E(Y) = E(X) - \mu$$

this implies that $E(X) = \mu$

8.5.3 Variance

To obtain variance we have

$$\begin{aligned}
\text{Var}(X) &= E[(X-\mu)^2] \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \\
&= 2\int_{\mu}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx && \text{(since the function is symmetric about } \mu\text{)}
\end{aligned}$$

We can obtain a gamma function by letting $y = \frac{(x-\mu)^2}{2\sigma^2}$ and noting that

$$\begin{aligned}
(x-\mu)^2 = 2\sigma^2 y,\ x-\mu = \sigma\sqrt{2y} \text{ since } x > \mu \\
dx = \frac{\sigma\sqrt{2}dy}{2\sqrt{y}} = \frac{\sigma}{\sqrt{2y}} dy
\end{aligned}$$

Then

$$\begin{aligned}
\text{Var}(X) &= 2\int_{0}^{\infty} (2\sigma^2 y) \frac{1}{\sigma\sqrt{2\pi}} e^{-y} \left(\frac{\sigma}{\sqrt{2y}} dy\right) \\
&= \frac{2\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} y^{\frac{1}{2}} e^{-y} dy \\
&= \frac{2\sigma^2}{\sqrt{(\pi)}} \Gamma(\frac{3}{2}) = \frac{2\sigma^2}{\sqrt{\pi}} \left(\frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right) \\
&= \frac{\sigma^2\sqrt{\pi}}{\sqrt{\pi}} && \text{(since } \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)\text{)} \\
&= \sigma^2
\end{aligned}$$

Theorem 8.5.1. *Let $X \sim N(\mu, \sigma^2)$ and define $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0,1)$ and*

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

Proof. The fact that $Z \sim N(0,1)$ has the probability density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad z \in \mathbb{R}$$

follows immediately by change of variables

Alternatively, we can just note that

$$\begin{aligned}
P(X \le x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2} dy \text{ let } z = \frac{y-\mu}{\sigma} \\
&= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\
&= P(Z \le \frac{x-\mu}{\sigma})
\end{aligned}$$

□

Example 8.5.1. *Find the following probabilities, where $Z \sim N(0,1)$*

(a) $P(Z \leq 2.11)$

(b) $P(Z < 3.40)$

(c) $P(Z > 1.06)$

(d) $P(Z \leq -1.06)$

(e) $P(-1.06 < Z \leq 2.11)$

Solution. (a) Look up 2.11 in the table by going down the left column to 2,1 then across to the heading 0.01, we find the number 0.98257. Then $P(Z \leq 2.11) = 0.98357$

(b) $P(Z < 3.40) = P(Z \leq 3.40) = 0.99966$

(c) $P(Z > 1.06) = 1 - P(Z \leq 1.06) = 1 - 0.85543 = 0.14457$

(d) Now we have to use symmetry:

$$P(Z \leq -1.06) = P(Z > 1.06) = 1 - P(Z \leq 1.06) = 1 - 0.85543 = 0.14457$$

(e)

$$\begin{aligned} P(-1.06 < Z < 2.11) &= P(Z < 2.11) - P(Z \leq 1.06) \\ &= P(Z \leq 2.11) - P(Z > 1.06) \\ &= P(Z \leq 2.11) - 1 + P(Z \leq 1.06) \\ &= 0.98257 - 1 + 0.85543 = 0.83800 \end{aligned}$$

■

We also have the tables to find desired values.

Example 8.5.2. *(a) Find a number c such that $P(Z \leq c) = 0.85$*

(b) Find a number d such that $P(Z > d) = 0.90$

(c) Find a number b such that $P(|z| \leq b) = 0.95$

Solution. (a) We can look in the body of the table to get an entry close to 0.85.This occurs for z between 1.03 and 1.04, $z = 1.04$ gives the closest value to 0.85. For greater accuracy, the table at the bottom of the last page is designed for finding numbers, given the probability. Looking beside hte entry 0.85, we find $z = 1.0364$

(b) Since $P(Z > d) = 0.90$ we have

$$P(Z \leq d) = 1 - P(Z > d) = 0.10$$

There is no entry for which $P(Z \leq z) = 0.10$ so we again have to use symmetry, since d will be negative. From the table we hae $P(Z \leq 1.2816) = 0.90$, then by symmetry, $P(Z > -1.2816) = 0.90$ and therefore $d = 1.2816$.

(c) We first note that $P(|z| < b) = P(-b < Z < b) = 0.95$ By symmetry, the probability outside the interval $(-b, b)$ must be 0.05, and this evenly split between the area between b and the area below $-b$. Therefore,

$$P(Z \leq -b) = P(Z > b) = 0.0025$$

and

$$P(Z \leq b) = 0.975$$

Looking within the body of the top table, we can see $P(Z \leq 1.96) = 0.975$ so $b = 1.96$

■

Finding $N(\mu, \sigma^2)$ probabilities:

To find $N(\mu, \sigma^2)$ probabilities in general, we use theorem given earlier that

$$\begin{aligned} P(a \leq X \leq b) &= P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}) \\ &= P(Z \leq \frac{b-\mu}{\sigma}) - P(Z \leq \frac{a-\mu}{\sigma}) \end{aligned}$$

where $Z \in N(0, 1)$

Example 8.5.3. *Let $X \sim N(3, 25)$*

(a) Find $P(X < 2)$

(b) Find a number c such that $P(X > c) = 0.95$

Solution. (a)

$$\begin{aligned} P(X < 2) &= P(\frac{X-\mu}{\sigma} < \frac{2-3}{5}) \\ &= P(Z < -0.20) \text{ where } Z \sim N(0,1) \\ &= 1 - P(Z < 0.20) = 1 - 0.57926 = 0.42074 \end{aligned}$$

(b)

$$\begin{aligned} P(X > c) &= P(\frac{X-\mu}{\sigma} > \frac{c-3}{5}) \\ &= P(Z > \frac{c-3}{5}) \text{ where } Z \sim N(0,1) \\ &= 0.95 \end{aligned}$$

Therefore, $\frac{c-3}{5} = -1.6449$, $c = -5.2245$

■

8.5.4 Gaussian Distribution

The normal distribution is also known as the **Gaussian Distribution**.The notation $X \sim G(\mu, \sigma)$ means that X has Guassian (Normal) distribution with mean μ and standard deviation σ, so, for example, if $X \sim N(1, 4)$, then we could write $X \sim G(1, 2)$

Example 8.5.4. *The distribution of heights of adult males in Canada is well approximate by a Guassian distribution with* $\mu = 69.0$ *inches and standard deviation* $\sigma = 2.4$ *inches. Find the* 10^{th} *and* 90^{th} *percentiles of the height distribution.*

Solution. We have $X \sim G(69.0, 2.4)$, or equivalently, $X \sim N(69.0, 5.76)$. To find the 90^{th} percentile c, we use

$$\begin{aligned} P(X \leq c) &= P(\frac{X - 69.0}{2.4} \leq \frac{c - 69.0}{2.4}) \\ &= P(Z \leq \frac{c - 69.0}{2.4}) \text{ where } Z \sim G(0,1) \\ &= 0.90 \end{aligned}$$

From the table we see $P(Z \leq 1.2816) = 0.90$ so we need

$$\frac{c - 69.0}{2.4} = 1.2816$$

which gives $c = 72.08$ inches as the 90^{th} percentile.

Similarly, to find c such that $P(X \leq c) = 0.10$ we find that $P(Z \leq -1.2816) = 0.10$, so we need

$$\frac{c - 69.0}{2.4} = -1.2816$$

or $c = 65.92$ inches as the 10^{th} percentile. ■

Chapter 9

Multivariate Distributions

9.1 Basic Terminology and Techniques

Many problems involve more than one single random variable. We need to extend the ideas introduced for single variables to deal with multivariate problems.

9.1.1 Joint Probability Function

Suppose there are two discrete random variables X and Y, and define the function

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

We call $f(x, y)$ the **joint probability function** of (X, Y). The properties of a joint probability function are similar to those for a single variable; for two random variables we have $f(x, y) \geq 0$ for all (x, y) and

$$\sum_{\text{all } (x,y)} f(x, y) = 1$$

In general,

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

if there are n random variables $X_1, \ldots, X_n$

Example: Consider the following, where we show $f(x, y)$ in a table:

		x			
$f(x,y)$		0	1	2	
y	1	0.1	0.2	0.3	
	2	0.2	0.1	0.1	
					1

Example: Suppose a fair coin is tossed 3 times. Define the random variable X = number of Heads and $Y = 1(0)$ if Heads (Tails) occurs on the first toss. Find the joint probability function for (X, Y)

Solution. First we should note the range for (X,Y), which is the set of possible values (x,y) which can occur. Clearly X can be $0,1,2,3$ and Y can be $0,1$, but we'll see that not all 8 combinations (x,y) are possible. We can find $f(x,y) = P(X = x, Y = y)$ by just writing down the sample space

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

that we have used before for this process. Then sample sounting gives $f(x,y)$ as shown in the following table:

		x				
$f(x,y)$		0	1	2	3	
y	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	
	1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	
						1

For example, $(X,Y) = (0,0)$ if and only if the outcome is TTT; $(X,Y) = (1,0)$ if and only if the outcome is either THT or TTH
Note that we just use a table for joint probability function since it's a little awkward to write it down in a formula. ■

9.1.2 Marginal Distributions

We may be given a joint probability function involving more variables than we're interested in using. Looking at the previous example, if we're only interested in X, and don't care what value Y takes, we can see that

$$\begin{aligned} P(X = 0) &= P(X = 0, Y = 1) + P(X = 0, Y = 2) \\ &= f(0,1) + f(0,2) \\ &= 0.3 \end{aligned}$$

Similarly,

$$\begin{aligned} P(X = 1) &= f(1,1) + f(1,2) = 0.3 \\ P(X = 2) &= f(2,1) + f(2,2) = 0.4 \end{aligned}$$

The distribution of X obtained in this way from the joint probability funciton is called the **marginal probability function** of X:

x	0	1	2	Total
$f_1(x) = P(X = x)$	0.3	0.3	0.4	1

In the same way, if we were only interested in Y, we obtain

$$P(Y = 1) = f(0,1) + f(1,1) + f(2,1) = 0.6$$

Since X can be 0,1,2 when $Y = 1$, the marginal probability of Y would be:

y	1	2	Total
$f_2(y) = P(Y = y)$	0.6	0.4	1

Note that we use the notation $f_1(x)$ and $f_2(y)$ to avoid confusion with $f(x, y) = P(X = x, Y = y)$. An alternative notation would be $f_X(x)$ and $f_Y(y)$

In general, to find $f_1(x)$ we add over all values of y where $X = x$, and to find $f_2(y)$ we add over all values of x with $Y = y$, then

$$f_1(x) = \sum_{\text{all } y} f(x, y)$$
$$f_2(y) = \sum_{\text{all } x} f(x, y)$$

The same reasoning can be extended beyond two variables, for example, with three variables (X_1, X_2, X_3)

$$f_1(x_1) = \sum_{\text{all } (x_2, x_3)} f(x_1, x_2, x_3)$$
$$f_{1,3}(x_1, x_3) = \sum_{\text{all } x_2} f(x_1, x_2, x_3) = P(X_1 = x_1, X_3 = x_3)$$

where $f_{1,3}(x_1, x_3)$ is the marginal joint probability function of X_1, X_3

Note that if the joint probability function is given in a table then the marginal probability functions are obtained by simply summing over the rows and columns as shown in the table below for the coin example above:

		x				
$f(x, y)$		0	1	2	3	$f_2(y)$
y	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{4}{8}$
	1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{4}{8}$
	$f_1(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

9.1.3 Independent Random Variables

Definition 9.1.1. *X and Y are **independent random variables** if $f(x, y) = f_1(x)f_2(y)$ for all values (x, y)*

Definition 9.1.2. *In general, $X_1, X_2, \ldots, X_n$ are **independent random variables** if and only if*

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

Be careful with this definition, we can only conclude that X and Y are independent variables after checking all (x, y) combinations. Even a single case where $f_1(x)f_2(y) \neq f(x, y)$ makes X and Y dependent random variables.

9.1.4 Conditional Probability Function

Definition 9.1.3. *The conditional probability function of X given $Y = y$ is*

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

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

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

Example: Suppose X and Y have joint probability function:

$f(x,y)$		$x=0$	$x=1$	$x=2$	$f_2(y)$
y	1	0.1	0.2	0.3	0.6
	2	0.2	0.1	0.1	0.4
	$f_1(x)$	0.3	0.3	0.4	1

Find $f_1(x|Y=1) = P(X=x|Y=1)$

Solution. Since

$$f_1(x|Y=1) = \frac{f(x,1)}{f_2(1)}$$

we obtain

x	0	1	2	Total
$f_1(x\|Y=1)$	$\frac{0.1}{0.6}=\frac{1}{6}$	$\frac{0.2}{0.6}=\frac{1}{3}$	$\frac{0.3}{0.6}=\frac{1}{2}$	1

As you would expect, the marginal and conditional probability functions are probability functions in that they are always ≥ 0 and their sum is 1. ■

9.1.5 Functions of Random Variables

We often encounter problems where we need to find the probability distribution of a function of two or more random variables. The most general method for finding the probability function for some function of random variables X and Y involves looking at every combination (x,y) to see what value the function takes.

Example: Suppose X and Y have joint probability function

$f(x,y)$		$x=0$	$x=1$	$x=2$	
y	1	0.1	0.2	0.3	
	2	0.2	0.1	0.1	
					1

and we want to find the probability function of $U = 2(Y-X)$. The possible values of U are seen by looking at hte value of $u = 2(y-x)$ for each (x,y) in the range of (X,Y)

u		$x=0$	$x=1$	$x=2$
y	1	2	0	-2
	2	4	2	0

Since

$$\begin{aligned}
P(U=-2) &= P(X=2, Y=1) = f(2,1) = 0.3 \\
P(U=0) &= P(X=1, Y=1) + P(X=2, Y=2) = f(1,1) + f(2,2) = 0.3 \\
P(U=2) &= f(0,1) + f(1,2) = 0.2 \\
P(U=4) &= f(0,2) = 0.2
\end{aligned}$$

the probability function of U is

u	-2	0	2	4	Total
$P(U=u)$	0.3	0.3	0.2	0.2	1

For some functions it's possible to approach this problem more systemetially. One of the most common functions of this type is the total. Let $T = X + Y$, this gives:

t		x: 0	1	2
y	1	1	2	3
	2	2	3	4

Then

$$P(T=3) = f(1,2) + f(2,1) = 0.4$$

Continuing in this way, we obtain

t	1	2	3	4	Total
$P(T=t)$	0.1	0.4	0.4	0.1	1

In fact, to find $P(T=t)$, we are simply adding the probabilities for all (x,y) combinations with $x+y=t$, this could be written as

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

However, if $x + y = t$, then $y = t - x$. To systematically pick out the right combinations of (x,y), all we really need to do is to sum over values of x and then substitute $t-x$ for y, then

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

So $P(T=3)$ would be

$$P(T=3) = \sum_{\text{all } x} f(x, 3-x) = f(0,3) + f(1,2) + f(2,1) = 0.4$$

(Note: $f(0,3) = 0$ since Y can't be 3)

We cam summarize the method of finding probability function for a function $U = g(X,Y)$ of two random variables X and Y as follows

Theorem 9.1.1. *Let $f(x,y) = P(X=x, Y=y)$ be the probability function for (X,Y), then the probability function for U is*

$$P(U=u) = \sum_{\textit{all } (x,y),\ g(x,y)=u} f(x,y)$$

This can also be extended to functions of three or more random variables $U = g(X_1, X_2, \ldots, X_n)$:

$$P(U=u) = \sum_{(x_1,\ldots,x_n):\ g(x_1,\ldots,x_n)=u} f(x_1,\ldots,x_n)$$

Theorem 9.1.2. *If $X \sim Poisson(\mu_1)$ and $Y \sim Poisson(\mu_2)$ independently then $T = X + Y \sim Poisson(\mu_1 + \mu_2)$*

Proof. Since $X \sim Poisson(\mu_1)$ and independently $Y \sim Poisson(\mu_2)$ their joint probability function is given by

$$f(x,y) = \frac{\mu_1^x e^{-\mu_1}}{x!} \cdot \frac{\mu_2^y e^{-\mu_2}}{y!} \text{ for } x = 0,1,\ldots \text{ and } y = 0,1,\ldots$$

The probability function of T is

$$\begin{aligned}
P(T=t) &= P(X+Y=t) \\
&= \sum_{\text{all } x} P(X=x, Y=t-x) \\
&= \sum_{x=0}^{t} \frac{\mu_1^x e^{-\mu_1}}{x!} \cdot \frac{\mu_2^{t-x} e^{-\mu_2}}{(t-x)!} \\
&= \mu_2^t e^{-(\mu_1+\mu_2)} \sum_{x=0}^{t} \frac{1}{x!(t-x)!} \left(\frac{\mu_1}{\mu_2}\right)^x \\
&= \frac{\mu_2^t e^{-(\mu_1+\mu_2)}}{t!} \sum_{x=0}^{t} \binom{t}{x} \left(\frac{\mu_1}{\mu_2}\right)^x \\
&= \frac{\mu_2^t e^{-(\mu_1+\mu_2)}}{t!} \left(1 + \frac{\mu_1}{\mu_2}\right)^t && \text{(by the Binomial Theorem)} \\
&= \frac{\mu_2^t e^{-(\mu_1+\mu_2)}}{t!} \cdot \frac{(\mu_1+\mu_2)^t}{\mu_2^t} \\
&= \frac{(\mu_1+\mu_2)^t}{t!} e^{-(\mu_1+\mu_2)} \text{ for } t = 0,1,2,\ldots
\end{aligned}$$

which we recognize as the probability function of a $Poisson(\mu_1 + \mu_2)$ and we've proven the desired result. □

Theorem 9.1.3. *If $X \sim Binomial(n,p)$ and $Y \sim Binomial(m,p)$ independently then $T = X + Y \sim Binomial(n+m,p)$*

9.2 Multinomial Distribution

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

Example: Three sprinters, A, B, and C, compete against each other in 10 independent 100m races. The probability of winning any single race are 0.5 for A, 0.4 for B, and 0.1 for C. Let X_1, X_2 and X_3 be the number of races A, B and C win respectively.

(a) Find the joint probability function, $f(x_1, x_2, x_3)$

Solution. We know that $x_1 + x_2 + x_3 = 10$ since there are 10 races in all. We only need to have two variables since $x_3 = 10 - x_1 - x_2$. However, it's convenient just to use x_3 to save writing and preserve symmetry.

The reasoning to this problem is similar to the Binomial distribution except there are now 3 types of outcomes. There are $\frac{10!}{x_1!x_2!x_3!}$ different outcomes. Each of these arrangements has a probability of $(0.5)^{x_1}$, $(0.4)^{x_2}$, and $(0.1)^{x_3}$ times in some order, that is $(0.5)^{x_1}(0.4)^{x_2}(0.1)^{x_3}$

Therefore,

$$f(x_1, x_2, x_3) = \frac{10!}{x_1!x_2!x_3!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{x_3}$$

The domain of f is the set $\{(x_1, x_2, x_3) : \ x_i = 0, 1, \ldots, 10, i = 1, 2, 3 \text{ and } x_1 + x_2 + x_3 = 10\}$ ■

(b) Find the marginal probability function, $f_1(x_1)$

Solution. It would also be acceptable to drop x_3 as a variable, and write down the probability function for X_1, X_2 only:

$$f(x_1, x_2) = \frac{10!}{x_1!x_2!(10 - x_1 - x_2)!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{10-x_1-x_2}$$

We now have $f_1(x_1) = \sum_{x_2} f(x_1, x_2)$ The limits of summation need care: x_2 could be as small as 0, but since $x_1 + x_2 \leq 10$, we also require $x_2 \leq 10 - x_1$, thus

$$\begin{aligned}
f_1(x_1) &= \sum_{x_2=0}^{10-x_1} \frac{10!}{x_1!x_2!(10 - x_1 - x_2)!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{10-x_1-x_2} \\
&= \frac{10!}{x_1!}(0.5)^{x_1}(0.1)^{10-x_1} \sum_{x_2=0}^{10-x_1} \frac{1}{x_2!(10 - x_1 - x_2)!} \left(\frac{0.4}{0.1}\right)^{x_2} \\
&= \frac{10!}{x_1!(10 - x_1!)}(0.5)^{x_1}(0.1)^{10-x_1} \sum_{x_2=0}^{10-x_1} \binom{10 - x_1}{x_2} \left(\frac{0.4}{0.1}\right)^{x_2} \\
&= \binom{10}{x_1}(0.5)^{x_1}(0.1)^{10-x_1} \left(1 + \frac{0.4}{0.1}\right)^{10-x_1} && \text{(by the Binomial Theorem)} \\
&= \binom{10}{x_1}(0.5)^{x_1}(0.1)^{10-x_1} \frac{(0.1 + 0.4)^{10-x_1}}{(0.1)^{10-x_1}} \\
&= \binom{10}{x_1}(0.5)^{x_1}(0.5)^{10-x_1} && \text{(for } x_1 = 0, 1, 2, \ldots, 10)
\end{aligned}$$

This derivation is included as an example of how to find marginal distribution by summing a joint probability function, there is a much simpler method fo rthis problem. Note that each race is either won by A ("success") or it is not won by A ("failure"). Since the races are independent and X_1 is now just the number of "success" outcomes, X_1 must have a Binomial distribution, with $n = 10$ and $p = 0.5$. Hence,

$$f_1(x_1) = \binom{10}{x_1}(0.5)^{x_1}(0.5)^{10-x_1} \text{ for } x_1 = 0, 1, \ldots, 10$$

■

(c) Find the conditional probability function, $f(x_2|x_1)$

Solution. Since $f(x_2|x_1) = P(X_2 = x_2|X_1 = x_1)$, so that

$$\begin{aligned}
f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} &= \frac{\frac{10!}{x_1!x_2!(10-x_1-x_2)!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{10-x_1-x_2}}{\frac{10!}{x_1!(10-x_1)!}(0.5)^{x_1}(0.5)^{10-x_1}} \\
&= \frac{(10 - x_1)!}{x_2!(10 - x_1 - x_2)!} \frac{(0.4)^{x_2}(0.1)^{10-x_1-x_2}}{(0.5)^{x_2}(0.5)^{10-x_1-x_2}} \\
&= \binom{10 - x_1}{x_2} \left(\frac{4}{5}\right)^{x_2} \left(\frac{1}{5}\right)^{10-x_1-x_2} && \text{(for } x_2 = 0, 1, \ldots, (10 - x_1))
\end{aligned}$$

The range of X_2 depends on the value of x_1, which makes sense: if B wins x_1 races then the most A can win is $10 - x_1$

Note: This result can be obtained more simply by general reasoning. Once we're given A wins x_1 races, the remaining $10-x_1$ races ("trials") are all won by either B or C. Since $P(B \text{ wins}) = 0.4$ and $P(C \text{ wins}) = 0.1$, then for the races won by either B or C, the probability that B wins ("Success") is

$$P(B \text{ wins}|B \text{ or } C \text{ wins}) = \frac{P(B \text{ wins})}{P(B \text{ or } C \text{ wins})} = \frac{0.4}{0.4+0.1} = 0.8$$

So the probability function of the number of wins ("Sucesses") in $10 - x_1$ races ("trials") is

$$f(x_2|x_1) = \binom{10-x_1}{x_2}(0.8)^{x_2}(0.2)^{10-x_1-x_2} \text{ for } x_2 = 0, 1, \ldots, 10 - x_1$$

■

(d) Are X_1 and X_2 independent? Why?

Solution. X_1 and X_2 are clearly not independent random variables since the more races A wins, the fewer races there are for B to win. More formally,

$$f_1(x_1)f_2(x_2) = \binom{10}{x_1}(0.5)^{x_1}(0.5)^{10-x_1}\binom{10}{x_2}(0.4)^{x_2}(0.6)^{10-x_2} \neq f(x_1, x_2)$$

(In general, if the range for X_1 depends on the value of X_2, then X_1 and X_2 cannot be independent random variables). ■

(e) Find the probability function of $T = X + Y$

Solution. If $T = X_1 + X_2$ then

$$\begin{aligned} f_T(t) = P(T = t) &= \sum_{x_1}^{t} f(x_1, t - x_1) \\ &= \sum_{x_1=0}^{t} \frac{10!}{x_1!(t-x_1)!\underbrace{(10-x_1-(t-x_1))!}_{(10-t)!}}(0.5)^{x_1}(0.4)^{t-x_1}(0.1)^{10-t} \end{aligned}$$

The upper limit on x_1 is t because, for example, if $t = 7$ then A could not have won more than 7 races. Then

$$\begin{aligned} f_T(t) = P(T = t) &= \frac{10!}{(10-t)!}(0.4)^t(0.1)^{10-t}\sum_{x_1=0}^{t}\frac{1}{x_1!(t-x_1)!}\left(\frac{0.5}{0.4}\right)^{x_1} \\ &= \binom{10}{t}(0.4)^t(0.1)^{10-t}\left(1+\frac{0.5}{0.4}\right)^t \\ &= \binom{10}{t}(0.4)^t(0.1)^{10-t}\frac{(0.4+0.5)^t}{(0.4)^t} \\ &= \binom{10}{t}(0.9)^t(0.1)^{10-t} \text{ for } t = 0, 1, \ldots, 10 \end{aligned}$$

■

Physical setup for the Multinomial distribution: Suppose an experiment is repeated independently n times with k distinct types of outcome each time. Let the probabilities of these k types be $p_1, p_2, \ldots, p_k$ each time. Let X_1 be the number of times the 1^{st} type occurs, X_2 the number of times the 2^{nd} occurs,..., X_k the number of times the k^{th} type occurs, then $(X_1, X_2, \ldots, X_k)$ has a Multinomial distribution.

Note:

- $p_1 + p_2 + \cdots + p_k = 1$

- $X_1 + X_2 + \cdots + X_k = n$

9.2.1 Joint Probability Function

There are $\frac{n!}{x_1!x_2!\cdots x_k!}$ different outcomes of the n trials. Therefore,

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

The domain on the x_i's are $x_i = 0, 1, \ldots, n$ and $\sum_{i=1}^{k} x_i = n$

We sometimes use the notation $(X_1, \ldots, X_k) \sim Multinomial(n;\ p_1, \ldots, p_k)$ to indicate that $(X_1, \ldots, X_k)$ have a Multinomial distribution.

Example: A potter is producing teapots one at a time. Assume that they are produced independently of each other and with probability p the pot produced will be "satisfactory"; the rest are sold at a lower price. The number, X, of rejects before producing a satisfactory teapot is recorded. When 12 satisfactory teapots are produced, what is the probability the 12 values of X will consist of six 0's, three 1's, two 2's and one value which is ≥ 3?

Solution. Each time a "satisfactory" pot is produced the value of X falls in one of the four categories, $X = 0, X = 1, X = 2, X \geq 3$. Under the assumptions, X has a Geometric distribution with

$$P(X = x) = f(x) = p(1-p)^x \text{ for } x = 0, 1, 2, \ldots$$

therefore, we have

$$\begin{aligned}
P(X = 0) &= f(0) = p \\
P(X = 1) &= f(1) = p(1-p) \\
P(X = 2) &= f(2) = p(1-p)^2 \\
P(X \geq 3) &= f(3) + f(4) + \ldots \\
&= p(1-p)^3 + p(1-p)^4 + \cdots \\
&= \frac{p(1-p)^3}{1-(1-p)} \qquad \text{(by the Geometric Series)} \\
&= (1-p)^3
\end{aligned}$$

Therefore,

$$\begin{aligned}
&P(\text{six 0's, three 1's, two 2's, and one value } \geq 3) \\
&= \frac{12!}{6!3!2!1!}[p]^6[p(1-p)]^3[p(1-p)^2]^2[(1-p)^3]^1 \\
&= \frac{12!}{6!3!2!1!}p^{6+3+2}(1-p)^{3+4+3} \\
&= \frac{12!}{6!3!2!1!}p^{11}(1-p)^{10}
\end{aligned}$$

■

9.3 Markov Chains

9.4 Expection for Multivariate Distributions: Covariance and Correlation

Definition 9.4.1.

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

and

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

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

	$f(x,y)$	x = 0	1	2	$f_2(y)$
y	1	0.1	0.2	0.3	0.6
	2	0.2	0.1	0.1	0.4
	$f_1(x)$	0.3	0.3	0.4	1

Find $E(XY)$ and $E(X)$

Solution.

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

To find $E(X)$ we have a choice of methods. First, taking $g(x,y) = x$ we get

$$\begin{aligned}
E(X) &= \sum_{\text{all }(x,y)} xf(x,y) \\
&= (0\times 0)(0.1) + (1\times 0.2) + (2\times 0.3) + (0\times 0.2) + (1\times 0.1) + (2\times 0.1) \\
&= 1.1
\end{aligned}$$

Alternatively, since $E(X)$ only involves X, we could find $f_1(x)$ and use

$$E(X) = \sum_{x=0}^{2} x f_1(x) = (0 \times 0.3) + (1 \times 0.3) + (2 \times 0.4) = 1.1$$

■

Property of Multivariate Expectation

It is easy proved that

$$E[ag_1(X,Y) + bg_2(X,Y)] = aE[g_1(X,Y)] + bE[g_2(X,Y)]$$

This can be extended beyond 2 functions g_1 and g_2, and beyond 2 variables X and Y

9.4.1 Relationships between Variables

Definition 9.4.2. *The* ***covariance*** *of X and Y, dentoed $\text{Cov}(X,Y)$ or σ_{XY}, is*

$$\text{Cov}(X,Y) = E[(X - \sigma_X)(Y - \sigma_Y)]$$

Note that

$$\begin{aligned}
\text{Cov}X,Y &= E[(X - \sigma_X)(Y - \sigma_Y)] \\
&= E(XY - \sigma_X Y - X\sigma_Y + \sigma_X\sigma_Y) \\
&= E(XY) - \sigma_X E(Y) - \sigma_Y E(X) + \sigma_X\sigma_Y \\
&= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

and $\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$ is the formula we usually use for calculation purposes.

Theorem 9.4.1. *If X and Y are independent then* $\text{Cov}(X,Y) = 0$

Proof. Recall $E(X - \sigma_X) = E(X) - \sigma_X = 0$. Let X and Y be independent, then $f(x,y) = f_1(x)f_2(y)$

$$\begin{aligned}
\text{Cov}(X,Y) = E[(X - \sigma_X)(Y - \sigma_Y)] &= \sum_{\text{all } y}[\sum_{\text{all } x}(x - \sigma_X)(y - \sigma_Y)f_1(x)f_2(y)] \\
&= \sum_{\text{all } y}[(y - \sigma_Y)f_2(y)\sum_{\text{all } x}(x - \sigma_X)f_1(x)] \\
&= \sum_{\text{all } y}[(y - \sigma_Y)f_2(y)E(X - \sigma_X)] \\
&= \sigma_{\text{all } y} 0 = 0
\end{aligned}$$

□

Theorem 9.4.2. *Suppose random variables X and Y are independent random variables. Then, if $g_1(X)$ and $g_2(Y)$ are any two functions,*

$$E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)]$$

Proof. Since X and Y are independent, $f(x,y) = f_1(x)f_2(y)$, thus

$$\begin{aligned} E[g_1(X)g_2(Y)] &= \sum_{\text{all } (x,y)} g_1(x)g_2(y)f(x,y) \\ &= \sum_{\text{all } x}\sum_{\text{all } y} g_1(x)f_1(x)g_2(y)f_2(y) \\ &= [\sum_{\text{all } x} g_1(x)f_1(x)][\sum_{\text{all } y} g_2(y)f_2(y)] \\ &= E[g_1(X)]E[g_2(Y)] \end{aligned}$$

□

To prove Theorem 9.4.1, we just note that if X and Y are independent then by Theorem 9.4.2

$$\begin{aligned} \text{Cov}(X,Y) &= E[(X-\sigma_X)(Y-\sigma_Y)] \\ &= E(X-\sigma_X)E(Y-\sigma_Y) \\ &= 0 \times 0 = 0 \end{aligned}$$

Caution: This result is not reversible. If $\text{Cov}(X,Y) = 0$ we cannot conclude that X and Y are independent random variables!!

Definition 9.4.3. *The **correlation coefficient** of X and Y is*

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

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

Properties of ρ:

1. Since σ_X and σ_Y, the standard deviation of X and Y, are both positive, 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. $-1 \le \rho \le 1$ and $\rho \to \pm 1$ the relation between X and Y becomes one-to-one and linear.

9.5 Mean and Variance of a Linear Combination of Random Variables

9.5.1 Results for Means

1. $E(aX + bY) = aE(X) + bE(Y) = a\sigma_X + b\sigma_X$, when a and b are constants.

2. Let a_i be constants (real numbers) and $E(X_i) = \mu_i, i = 1, 2, \ldots, n$ Then $E(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i \mu_i$. In particular, $E(\sum_{i=1}^n X_i = \sum_{i=1}^n E(X_i))$

3. Let $X_1, X_2, \ldots, X_n$ be random variables which have mean μ. The sample mean is $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then $E(\overline{X}) = \mu$

9.5.2 Result for Covariance

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

9.5.3 Results for Variance

1. **Variance of a Linear Combination**

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

2. **Variance of a sum of independent random variables**: Let X and Y be independent, since $\mathrm{Cov}(X, Y) = 0$, result 1 gives:

$$\mathrm{Var}(X + Y) = \sigma_X^2 + \sigma_Y^2$$

that is, for independent variables, the *variance of a sum is the sum of the variances*. Also note

$$\mathrm{Var}(X - Y) = \sigma_Z^2 + (-1)^2\sigma_Y^2 = \sigma_X^2 + \sigma_Y^2$$

3. **Variance of a general linear combination of random variables**: Let a_i be constants and $\mathrm{Var}(X_i) = \sigma_i^2$ then

$$\mathrm{Var}(\sum_{i=0}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j)$$

4. **Variance of a linear combination of independent random variables**: Special cases of result 3 are

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

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2\sigma_i^2$$

   b) If $X_1, \ldots, X_n$ are independent and all have the same variance σ^2, then

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

9.6 Linear Combinations of Independent Normal Random Variables

Theorem 9.6.1. *1. Let $X \sim N(\mu, \sigma^2)$, and $Y = aX + b$, where a and b are constant real numbers, then $Y \sim N(a\mu + b, a^2\sigma^2)$*

2. Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ independently, and let a and b be constants, then $aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$. In general, if $X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \ldots, n$ independently and $a_1, \ldots, a_n$ are constants, then $\sum_{i=1}^n a_i X_i \sim N(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2)$

3. Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables, then $\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$ and $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

Chapter 10

C.L.T., Normal Approximations and M.G.F's

10.1 Central Limit Theorem (C.L.T.) and Normal Approximations

Theorem 10.1.1. ***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 of the random variables

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

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

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

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

This is a theorem about limits. We'll use it when n is large, but finite, to approximate the distirbution of S_n or $\overline{X}$ by a Normal distribution. That is, we will use

$$S_n = \sum_{i=1}^{n} X_i \text{ has approximately a } N(n\mu, n\sigma^2) \text{ distribution for lage } n$$

and

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ has approximately a } N(\mu, \frac{\sigma^2}{n}) \text{ distribution for large } n$$

Notes:

1. As $n \to \infty$, both distirbutions $N(n\mu, n\sigma^2)$ and $N(\mu, \frac{\sigma^2}{n})$ fail to exist.
2. The Central Limit Theorem does not hold if the common mean μ and covariance σ^2 do not exist.
3. We use the Central Limit Theorem to approximate the distribution of the sum $S_n = \sum_{i=1}^{n} X_i$ or average $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The accuracy of the approximation depends on n (bigger is better) and also the actual distribution of X_i's. The appximation works better for smaller n when the shape of probability function/probability density function of X_i is symmetric (for exmaple, the $U(a, b)$ probability density function) or nearly symmetric (for exmaple, the $Poisson(5)$ probability function).

4. In section 9.6, the distributions of linear combination of indendent Normal random variables were given. In particular, if $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ random variables then

$$S_n = \sum_{i=1}^{n} X_1 \sim N(n\mu, n\sigma^2) \text{ and } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$$

Thus, if the X_i's themselves form a Normal distribution, then S_n and $\overline{X}$ have exactly Normal distributions for all values of n. If the X_i's do not have a Normal distribution themselves, then S_n and $\overline{X}$ have approximately Normal distribution when n is large.