STATS 230 Course Notes

Joshua Allum

April 2017

Contents

Chapter 1

Introduction

1.1 Defining Probability

The Classical Definition

The probability of an event is

$$\frac{\text{the number of ways the event may occur}}{\text{the total number of possible outcomes}}$$

provided all outcomes are equally likely.

Example 1.1.1

The probability of a fair dice landing on 3 is 1/6 because there is one way in which the dice may land on 3 and 6 total possible outcomes of faces the dice may land on. The sample space of the experiment, $\mathbb{S}$, is $\{1, 2, 3, 4, 5, 6\}$ and the event occurs in only one of these six outcomes.

The main limitation of this definition is that it demands that the outcomes of a sample space are equally likely. This is a problem since a definition of "likelyhood" (probability) is needed to include this postulate in a definition of probability itself.

The Relative Frequency Definition

The probability of an event is the limiting proportion of times that an event occurs in a large number of repetitions of an experiment.

Example 1.1.2

The probability of a fair dice landing on 3 is 1/6 because after a very large series of repetitions (ideally infinite) of rolling the dice, the fraction of times the face with 3 is rolled tends to 1/6.

The main limitation of this definition is that we can never repeat a process indefinitely so we can never truly know the probability of an event from this definition. Additionally, in some cases we cannot even obtain a long series of repetitions of processes to produce an estimate due to restrictions on cost, time, etc.

The Subjective Definition

The probability of an event occurring is a measure of how sure the person making the statement is that the event will occur.

Example 1.1.3

The probability that a football team will win their next match can be predicted by experts who regard all the data of past matches and current situations to provide a subjective probability.

This definition is irrational and leads to many people having different probabilities for the same events, with no clear "right" answer. Thus, by this definition, probability is not an objective science.

Probability Model

To avoid many of the limitation of the definitions of probability, we can instead treat probability as a mathematical system defined by a set of axioms. Thus, we can ignore the numerical values of probabilities until we consider a specific application. The model is defined as follows

- A sample space of all possible outcomes of a random experiment is defined.
- A set of events, to which we may assign probabilities, is defined.
- A mechanism for assigning probabilities to events is specified.

Chapter 2

Mathematical Probability Models

2.1 Sample Spaces

A sample space, $\mathbb{S}$, is a set of distinct outcomes for an experiment or process, with the property that in a single trial, one and only one of these outcomes occurs. The outcomes that make up a sample space are called sample points or simply points.

Example 2.1.1

The sample space for a roll of a six-sided die is

$$\{a_1, a_2, a_3, a_4, a_5, a_6\} \quad \text{where } a_i \text{ is the event the top face is } i$$

More simply we could define the sample space as

$$\{1, 2, 3, 4, 5, 6\}$$

Note that a sample space of a probability model for a process is not necessarily unique. Often times, however, we try to chose sample points that are the smallest possible or "indivisible".

Example 2.1.2

If we define E to be the event that the top face of a six-sided die is even when rolled and O to be the event the top-face is odd, then the sample space, $\mathbb{S}$, can be defined as

$$\{E, O\}$$

This is the same process as Example 2.1.1 (rolling a six-sided die), so since the sample spaces differ, clearly, sample spaces are not unique. Moreover, if we are interested in the event that a 3 is rolled, this sample space is not suitable since it groups the event in question with other events.

A sample space can be either **discrete** or **non-discrete**. If a sample space is discrete, it consists of a finite or countably infinite number "simple events". A countably infinite set is one that can be put into a one-to-one correspondence with the set of real numbers. For example, $\left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right\}$ is countably infinite whereas $\{\, x \mid x \in \mathbb{R} \,\}$ is not.

Simple Events

An event in a discrete sample space is a subset of the sample space, i.e., $A \subset \mathbb{S}$. If the event is indivisible, so as to only contain one point, we call it a simple event, otherwise it is a compound event.

Example 2.1.3

A simple event for a roll of a six-sided die is $A = \{a_1\}$ where a_i is the event the top face is i. A compound event is $E = \{a_2, a_4, a_6\}$.

2.2 Assigning Probabilities

Let $\mathbb{S} = \{a_1, a_2, a_3, \ldots\}$ be a discrete sample space. We assign probabilities, $P(a_i)$, for $i = 1, 2, 3, \ldots$ to each sample point a_i such that the following two conditions hold

- $0 \leq P(a_i) \leq 1$
- $\sum\limits_{\text{all } i} P(a_i) = 1$

The set of probabilities $\{\, P(a_i) \mid i = 1, 2, 3, \ldots \}$ is called a **probability distribution** on $\mathbb{S}$.

Note that P is a function with the sample space as its domain.

The second condition, that the sum of the probabilities of all sample points is 1, relates to the property that for a given experiment one simple event in the sample space must occur. Every experiment or process always has an outcome thus the probability of any outcome being achieved must be 1.

Compound Events

The probability of an event A is the sum of the probability of all the simple events that make up A.

$$P(A) = \sum_{a \in A} P(a)$$

Example 2.2.1

In the previous example we saw that $E = \{a_2, a_4, a_6\}$ is a compound event. Thus, the probability of the compound event E is

$$P(E) = P(a_2) + P(a_4) + P(a_6)$$

Note that the probability model that we defined does not specify what actual numbers to assign to the simple events of a process. It only defines the properties that guarantee mathematical consistency. Thus, if we assigned $P(a_2)$ to be 0.9, our model would still be mathematically consistent but would not be consistent with the frequencies we obtain in multiple repetitions of the experiment.

In actual practice we try to define probabilities that are approximately consistent with the frequencies of the events in multiple repetition of the process.

Complements

The complement of an event, A, is the set of all outcomes not included in A and is denoted by $\overline{A}$.

Example 2.2.2

If $E = \{ a_1, a_3, a_5 \}$ is a compound event on the sample space $\{ a_1, a_2, a_3, a_4, a_5, a_6 \}$, then the complement of E is

$$\overline{E} = \{ a_2, a_4, a_6 \}$$

Because of the nature of complementary events, two complementary events cannot both occur in one process. The events are **mutually exclusive**.

Chapter 3

Counting Techniques

3.1 Counting Arguments

If we have a sample space, $\mathbb{S}$, of some experiment that has a **uniform distribution** (all sample points are equally likely), then we can calculate the probability of a compound event A as the number of sample points in A divided by the total number of sample points.

$$P(A) = \frac{k}{n}$$

where k is the number of sample points in A and n is the total number of sample points in the sample space.

Addition Rule

Consider we can perform process 1 in p ways and process 2 in q ways. Suppose we want to do process 1 **or** process 2 **but not both**, then there are $p + q$ ways to do so.

Example 3.1.1

Suppose a keyboard only has 26 letters and 20 special characters (!%#$), there are 46 ways in which a typist may type a **single** character. (Process 1: typing a letter. Process 2: typing a special character).

Multiplication Rule

Again, consider we can perform process 1 in p ways and process 2 in q ways. Suppose we want to do process 1 **and** process 2, then there are $p \times q$ ways to do so. This is because **for each way** of doing process 1 we can do process 2 in q ways.

Example 3.1.2

Suppose the same typist with the same keyboard wants to type a single letter **and** a single special character. The typist can do so in 520 ways, since there are 26 ways to select the letter and **for each** possible letter selection there are 20 possible special character selections.

Try to associate **OR** and **AND** with **addition** and **multiplication** respectively in your mind.

Often times, **OR**'s and **AND**'s are not explicit or obvious so you must re-word your problem to identify implicit **OR**'s and **AND**'s.

Example 3.1.3

A young boy gets to pick 2 toys from a store for his birthday. How many ways can he pick 2 toys if the store contains 12 toys? He may pick the same toy multiple times and picks the toys at random.

We can re-word this problem as follows: A young boy selects one of 12 toys **and** again, selects one of 12 toys. Thus there are $12 \times 12 = 144$ ways in which he can select 2 toys. Furthermore, we have that since selections are random, each selection is equally likely. So the probability that the boy selects any pair of toys is 1/144.

In this case the boy was allowed to select the same toy more than once. This is often referred to as **with replacement**. The addition and multiplication rules are generally sufficient to find probability of processes with replacement but if processes occur without replacement solutions become more complex and other techniques are often used.

The phrase **at random** or **uniformly**, indicates that each point in the sample space is equally likely.

Example 3.1.4

Consider a farmer with 500 different seeds. How many ways can he select 3 seeds randomly to plant?

We can re-word this problem to become: A farmer selects one seed from 500 **and** then selects one seed of 499 **and** then one seed of 498. So there are $500 \times 499 \times 498$ ways to do so.

Now, how many ways can he select 5 and 50 seeds randomly?

He can select 5 seeds in $500 \times 499 \times 498 \times 497 \times 496$ ways and 50 seeds in $500 \times \cdots \times 451$ ways.

Generally, if there are n ways of doing a process and it is done k times **without replacement**, that is you can only do the process a specific way once, there are $n \times \cdots \times (n-k+1)$ ways to do it.

3.2 Counting Arrangements

When the sample space of a process is a set of arrangements of elements, like $\{ abc, acb, bac, bca, cab, cba \}$, the sample points are called the **permutations**. Assuming all n elements we are arranging are unique, how many sample points are there?

Consider trying to fill n boxes: $\square\square\cdots\square\square$. We have n ways to fill the first box (each element can go in the first box), **and** we have $(n-1)$ ways to fill the second box, **and** so on until we have 1 way to fill the n^{th} box. So there are $n \times (n-1) \times \cdots \times 1$ total permutations in the sample space.

Example 3.2.1

Consider the letters of the word "fiesta". A baby (who cannot spell) randomly rearranges the letters of the word. What is the probability that "fiesta" is the outcome?

There are six boxes to fill: $\square\square\square\square\square\square$. We have 6 ways to fill the first position, 5 ways to fill the second and so on until we have 1 way to fill the 6^{th} position. The number of points in the sample

space is $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$. So the probability of each outcome in the sample space is $1/720$.

Example 3.2.2

Consider the letters of the word "snake". If arranged randomly what is the probability that the word formed begins with a vowel?

There are five boxes to fill: $\square\square\square\square\square$. There are two ways to fill the first box:

$$\boxed{a}\square\square\square\square \text{ and } \boxed{e}\square\square\square\square$$

and for each of these ways there are four remaining boxes to fill. The number of ways to fill the 4 remaining boxes is $4 \times 3 \times 2 \times 1 = 24$ so the total number of outcomes in which the first letter is a vowel is $2 \times 24 = 48$. Therefore, the probability of the event occurring is $\frac{48}{\text{number of sample points}}$.
The five boxes can be filled by any letter to obtain a point in the sample space, so there are $5 \times 4 \times 3 \times 2 \times 1 = 120$ sample points. So the probability of the event occurring is $48/120 = 4/15$.

Example 3.2.3

Suppose we have 7 meals to distribute randomly to 7 people (one each). Three of the meals are gluten free and the other four are not. Of the 7 people, two of them cannot eat gluten. How many ways are there to distribute the meals without giving gluten to someone who cannot eat it?

We can liken this to the boxes example with each person being a box. Let the first two boxes be the people who cannot eat gluten. We have

$$\square\square\square\square\square\square\square$$

Since we cannot place a gluten meal in boxes 1 or 2, we have that we have 3 ways to fill box 1 then 2 ways to fill box 2. So there are 6 ways distribute meals to the gluten-free people. We have

$$\boxed{G}\boxed{G}\square\square\square\square\square$$

Now there are 5 boxes to be filled with any of 5 meals. So there are $5 \times 4 \times 3 \times 2 \times 1 = 120$ ways to distribute the meals to the other 5 people. This is an implicit **and** statement, thus there are $6 \times 120 = 720$ ways to distribute the meals.

3.3 Notations

Because some calculations occur very frequently in statistics we define a notation that helps us to deal with such problems.

Factorial

We define $n!$ for any natural number n to be

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 1$$

and in order to maintain mathematical consistency we define 0! to be 1. This is the number of arrangements of n possible unique elements, using each once.

n to k Factors

We define $n^{(k)}$ to be

$$n^{(k)} = n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}$$

This is the number of arrangements of length k using each element, of n possible unique elements, at most once.

Power of

As in ordinary mathematics $n^k = \underbrace{n \times n \times \cdots \times n}_{k}$. This represents the number of arrangements that can be made of length k using each element, of n possible unique elements, as often as we wish (with replacement).

For many problems it is simply impractical to try to count the number of cases by conventional means because of how big the numbers become. Notations such as $n!$ and n^k allow us to deal with these large numbers effectively.

Example 3.3.1

An evil advertising company randomly chooses 7-digit phone numbers to call to try to sell products. Find the probabilities of the following events:

- A: the number is your phone number
- B: the first three number are less than 5
- C: the first and last numbers match your phone number

Now assume that all 7-digits are unique (chosen without replacement):

- D: the number is 210-3869
- E: the first three number are less than 5
- F: the first and last numbers are 1 and 2 respectively

A: The initial sample space contains all the ways that one can select 7 numbers from the numbers 0 to 9 **with replacement**. There are 10 choices for each of the seven numbers, therefore the sample space contains 10^7 points. Thus, since all points are equally likely, $P(A) = 1/10^7$.

B: Now if the first three numbers are less than 5, there are 5 ways (0 to 4) to select each of the first three numbers and there are 10 ways to select each of the next four numbers. So there are $5^3 \times 10^4$ points in B. Therefore, $P(B) = \frac{5^3 \times 10^4}{10^7}$

C: There is only one way to select the first number such that it matches your number and the same is true for the last number. Thus, we must only consider the middle digits. There are 10 choices each for the middle five numbers, so there are 10^5 points in C. Therefore, $P(C) = 1/10^5$.

D: The new sample space contains all the ways that one can select 7 numbers from the numbers 0 to 9 **without replacement**. There are 10 choices for the first number, 9 for the second and so on until there are 4 choices for the last number. Thus, there are $10^{(7)}$ points in the sample space and since each is equally likely, $P(D) = 1/10^{(7)} = \frac{1}{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4}$.

E: If the first three numbers are less than five, there are 5 ways to select the first number, 4 for the second and 3 for the third, so there are $5^{(3)}$ ways to select the first 3 numbers. The next 4 digits may be selected from any of the 7 digits that were not used as one of the first 3. So there are $7^{(4)}$ ways to select the final four digits. Therefore, there are $5^{(3)} \times 7^{(4)}$ points in E. So, $P(E) = \frac{5^{(3)} \times 7^{(4)}}{10^{(7)}}$.

F: There is only one way to select the first and last digits as 1 and 2 respectively, so we must only consider the middle 5 digits. The 5 digits are selected from 8 numbers without replacement, so there are $8^{(5)}$ ways to do this. Therefore, $P(F) = \frac{8^{(5)}}{10^{(7)}}$.

3.4 Counting Subsets

In many problems, you will encounter a sample space, $\mathbb{S}$, of some experiment that consists of fixed-length subsets of some set.

Combinations

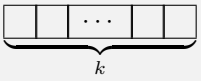
We define $\binom{n}{k}$ to be the number of subsets of size k that can be selected from a set of n elements. We have

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

It is read "n choose k".

Derivation of Choose

Suppose we have a set of n unique elements and we wish to select a subset of size k, such that $k \leq n$, and the elements of the subset are unique (selected without replacement). If we use the boxes metaphor we have k empty boxes.



There are n ways to select the first element of the subset, $(n-1)$ ways to select the second and so on until there are $(n-k+1)$ ways to select the k^{th} and last element.

So there are $n^{(k)}$ ways to fill the k boxes **but** note that some of the subsets will contain all the same elements as each other but in varying order. These subsets are not unique since we do not care for the arrangement of the items in a subset. Each unique subset can be arranged to form $k!$ permutations of its k elements. Thus, the number of unique subsets, $\binom{n}{k}$, multiplied by the number of arrangements of each subset, $k!$, is $n^{(k)}$. Therefore, we have

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

So it follows that

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

3.5 Properties of Combinations

Here are a few properties of $\binom{n}{k}$.

- $n^{(k)} = \frac{n!}{(n-k)!} = n(n-1)^{k-1}$ for $k \geq 1$
- $\binom{n}{k} = \frac{n!}{(n-k)!\,k!} = \frac{n^{(k)}}{k!}$
- $\binom{n}{k} = \binom{n}{n-k}$ for all $0 \leq k \leq n$
- $\binom{n}{0} = \binom{n}{n} = 1$
- $(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots + \binom{n}{n}x^n$ (Binomial Theorem)

3.6 Counting Arrangements of Set with Repeated Elements

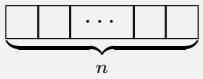
Thus far we have only discussed counting arrangements of unique items. Now, we consider a case in which we want to count the number of unique arrangements of size k of a set of n elements that are not necessarily unique.

Consider we have a set of n elements with k of those elements being unique. Let n_i be the number of appearances of the i^{th} element of the k unique elements. Thus, $n_1 + \cdots + n_k = n$. The number of different ways of selecting an arrangement of size n that uses all symbols is:

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

Derivation

Suppose we have a set of n elements with k being unique. Let the k unique items be labelled u_1 to u_k and let n_i be the number of appearance of u_i in the set of n elements. We want to form an arrangement of length n so using the boxes metaphor we have



We must use each of the n elements once so we must select n_1 boxes to fill with u_1's. This can be done in $\binom{n}{n_1}$ ways. Next, we must select n_2 of the remaining $n - n_1$ boxes to fill with u_2's, n_3 of the remaining $n - n_1 - n_2$ boxes to fill with u_3's, and so on until we must select n_k of the $n - n_1 - n_2 - \cdots - n_{k+1} = n_k$ remaining boxes to fill with u_k's. Therefore, there are

$$\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \cdots \times \binom{n_k}{n_k} = \frac{n!}{\cancel{(n-n_1)!}\,n_1!} \times \frac{\cancel{(n-n_1)!}}{(n-n_1-n_2)!\,n_2!} \times \cdots \times \frac{(n_{k-1}+n_k)!}{\cancel{n_k!}\,n_{k-1}!} \times \frac{\cancel{n_k!}}{0!\,n_k!}$$

ways, which simplifies to,

$$\frac{n!}{n_1!\,n_2!\,n_3!\cdots n_k!}$$

to do so.

Chapter 4

Probability Rules

4.1 General Rules

Here are a few basic rules of probabilities. They should be relatively straightforward.

Theorem 4.1.1

For a sample space, $\mathbb{S}$, the probability of a simple event in $\mathbb{S}$ occurring is 1. That is

$$P(\mathbb{S}) = 1$$

Proof 4.1.1:

$$P(\mathbb{S}) = \sum_{a \in \mathbb{S}} P(a) = \sum_{\text{all } a} P(a)$$

□

Theorem 4.1.2

Any event A in a sample space has a probability between 0 and 1 inclusive. That is

$$0 \leq P(A) \leq 1 \text{ for all } A \subseteq \mathbb{S}$$

Proof 4.1.2:

Note that A is a subset of $\mathbb{S}$, so

$$P(A) = \sum_{a \in A} P(a) \leq \sum_{a \in \mathbb{S}} P(a) = 1$$

Now, recall that $P(a) \geq 0$ for any sample point a by our probability model. Thus, since $P(A)$ is the sum of non-negative real numbers, $P(A) \geq 0$. So we have

$$1 \leq P(A) \leq 1$$

□

Theorem 4.1.3

If A and B are two events such that $A \subseteq B$, that is all the sample points in A are also in B, then

$$P(A) \leq P(B)$$

Proof 4.1.3:

$$P(A) = \sum_{a \in A} P(a) \leq \sum_{a \in B} P(a) = P(B)$$

□

4.2 Venn Diagrams

As we have seen already, it is helpful to think of events in a sample space as subsets of the sample space. Consider a sample space, $\mathbb{S} = \{ 1, 2, 3, 4, 5, 6 \}$. A number is picked at random, let E be the event that the number is even. We can think of E as the subsets of $\mathbb{S}$, $\{ 2, 4, 6 \}$ and the probability of E is the probability of any sample points in A occurs, that is 2, 4, or 6 is selected. We can represent the relationships of events in the sample space using Venn diagrams.

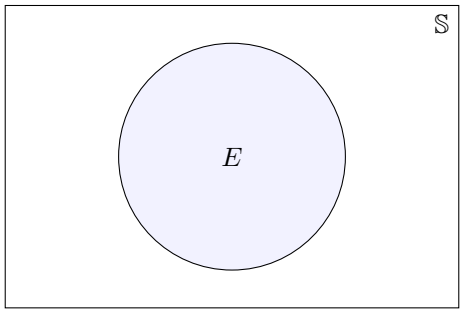


Figure 4.1: Single event E

Now, assuming the area of E is half the area of $\mathbb{S}$, we have that the probability of E is the probability that a randomly chosen point on the area of $\mathbb{S}$ will be within E.

Consider now we let $G = \{ 4, 5, 6 \}$ be the event that that the number selected is greater than or equal to 4. We have

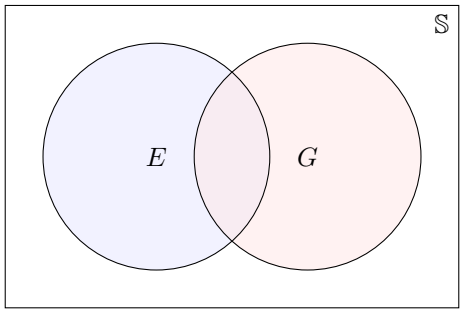


Figure 4.2: Events E and G

The total shaded region of the Venn diagram, $E \cup G$, contains all the sample points of E and G. It is the event that any outcome in either E or G, or both, occurs. Thus, $E \cup G$ is the event that E, G or both, occurs. Similarly, the union of three events is the event that at least one of the three events occur.

Consider now the intersection $E \cap G$. It is the set of all the points that are in both E and G, $\{4, 6\}$. Thus, it is the event that an outcome in both E and G occurs. So $E \cap G$ is the event that E and G both occur.

The sets $A \cap B$ and similarly $A \cap B \cap C$ are often denoted as AB and ABC respectively.

Finally, the unshaded space in Figure 4.1 is the set of all outcomes that are not in E. It is the complement of E and is denoted by $\overline{E}$. It is the event that E does not occur.

Note that the complement of $\mathbb{S}$ is the null set, that is $\overline{\mathbb{S}} = \emptyset$, and has a probability of 0.

4.3 De Morgan's Laws

Theorem 4.3.1

The following are De Morgan's Laws:

1. $\overline{A \cup B} = \overline{A} \cap \overline{B}$
2. $\overline{A \cap B} = \overline{A} \cup \overline{B}$

4.4 Rules for Unions of Events

Recall Figure 4.2, copied below.

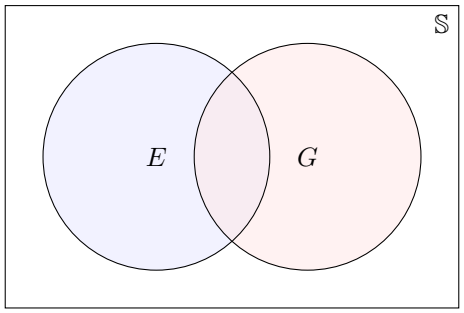


Figure 4.2: Events E and G (repeated from page 15)

We can see that the area of $E \cup G$ is not simply the sum of the areas of E and G. So we have that the probability of $E \cup G$ is not simply the sum of the probability of E and G. Rather, we must sum the probabilities and subtract the intersection (which gets included twice in the sum) to obtain $P(E \cup G)$.

Theorem 4.4.1

For any events, A and B, in a sample space, we have

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

Example 4.4.1

A number between 1 and 6 inclusive is chosen randomly. Let $E = \{2, 4, 6\}$ be the event the number is odd and let $G = \{4, 5, 6\}$ be the event that the number is greater than or equal to 4.

The probability of the number being even **or** greater than 4 is $P(E \cup G)$. Since both E and G contain 3 points of the six in the sample space, $P(E) = P(G) = 1/2$. Thus, we can see clearly that $P(E \cup G) \neq P(E) + P(G) = 1$ since $\{1\}$ is not in E or G and has a probability of 1/6. Now, note $E \cap G = \{4, 6\}$, so $P(E \cap G) = 1/3$. We have

$$P(E \cup G) = P(E) + P(G) - P(E \cap G) = 1/2 + 1/2 - 1/3 = 2/3$$

Now consider the case of the union of three events.

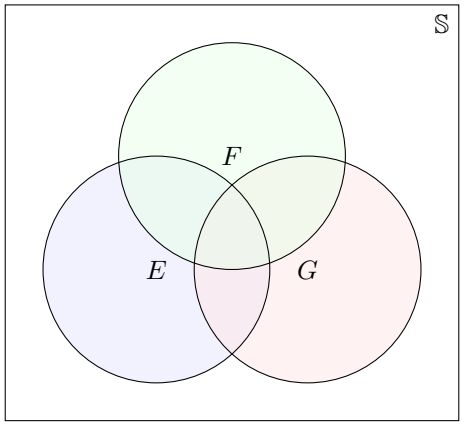


Figure 4.4: Three events

Let A_I be the area on the Venn diagram of the event I. The area of the union once again is not simply the sum of the areas $(A_E + A_G + A_F)$. Instead we can reason out that when we add the three areas we include $A_{E \cap G}$, $A_{G \cap F}$, and $A_{F \cap E}$ twice each and $A_{E \cap G \cap F}$ three times. The sum of these doubly counted areas $(A_{E \cap G} + A_{G \cap F} + A_{F \cap E})$ also includes $A_{E \cap G \cap F}$ three times. Thus, when we subtract the area of the doubly counted segments, $A_{E \cap G \cap F}$ is also subtracted three times leaving this area unaccounted for. Therefore we then add $A_{E \cap G \cap F}$ to find the complete area of $E \cup G \cup F$.

Theorem 4.4.2

For any events, A, B and C, in a sample space, we have

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

4.5 Mutually Exclusive Events

Events A and B are mutually exclusive if and only if $A \cap B = \emptyset$. More simply, the events A and B cannot both occur in one experiment because they share no points in common and only one sample point is achieved.

In general, events $A_1, A_2, A_3, \ldots, A_n$ are mutually exclusive if and only if $A_i \cap A_j = \emptyset$ for all $i \neq j$. This means that at most one of these events may occur in any one experiment.

Probability of the Unions of Mutually Exclusive Events

Consider the Venn diagram of two mutually exclusive events, E and G. Clearly the probability of the

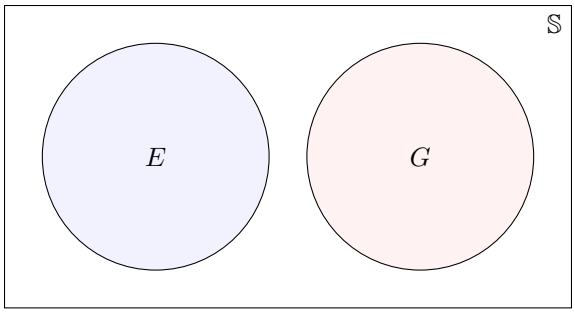


Figure 4.5: Two mutually exclusive events

intersection of two mutually exclusive events is 0, since it doesn't contains any sample points. So we have

$$P(E \cap G) = 0$$

Another intrinsic property of mutually exclusive events that we can see on a Venn diagram is that the area of $E \cup G$ is the sum of the areas of E and G. Therefore, unlike in previous examples, the probability of $E \cup G$ is the sum of the probabilities of E and G.

Theorem 4.5.1

For mutually exclusve events, A and B, in a sample space, we have

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

Theorem 4.5.2

More generally for n mutually exclusive events, $A_1, A_2, \ldots, A_n$, in a sample space, we have

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

Probabilities of Complements

Theorem 4.5.3

For any event A, we have

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

Proof 4.5.1:

Recall the complement of an event consists of all the sample points not in the event. Thus, for any event A, its complement $\overline{A}$ contains no points in common with A. So $A \cap \overline{A} = \emptyset$ and A and $\overline{A}$ are mutually exclusive, by definition. Now, consider $A \cup \overline{A}$, it spans the whole of the sample space so we have $P(A \cup \overline{A}) = 1$ and since A and $\overline{A}$ are mutually exclusive, we have

$$P(A) + P(\overline{A}) = 1$$

and it follows that $P(A) = 1 - P(\overline{A})$, as required. □

4.6 Independence of Events

Events A and B are said to be independent if and only if $P(A \cap B) = P(A)P(B)$. Otherwise they are dependent events.

In general, events $A_1, A_2, A_3 \ldots, A_n$ are independent if and only if

$$P(A_{i_1} \cap A_{i_2} \cap A_{i_3} \cap \cdots \cap A_{i_n}) = P(A_{i_1}) + P(A_{i_2}) + P(A_{i_3}) + \cdots + P(A_{i_n})$$

for all sets $\{ i_1, i_2, i_3, \ldots, i_k \}$ of distinct subscripts chosen from $\{ 1, 2, 3, \ldots, n \}$.

Example 4.6.1

Consider an experiment in which a fair die is tossed twice. We define the following events:

- A: The first number rolled is a six
- B: The second number rolled is a six
- C: The sum of the numbers rolled is less than or equal to seven
- D: The sum of the numbers rolled is equal to seven

Suppose the event A occurs. Does this have any impact on the probability of B, C or D occurring?

It is quite clear to see that the events A and B are independent events since rolling a six on the first toss has no impact on the number that will be rolled on the second toss. Now, events B and C from the onset appear to be dependent since if you roll a six on the first toss you must roll a one to make your total less than or equal to seven. To confirm this consider the sample space

$$\left\{ \begin{array}{l} (1,1),\ (1,2),\ (1,3),\ (1,4),\ (1,5),\ (1,6) \\ (2,1),\ (2,2),\ (2,3),\ (2,4),\ (2,5),\ (2,6) \\ (3,1),\ (3,2),\ (3,3),\ (3,4),\ (3,5),\ (3,6) \\ (4,1),\ (4,2),\ (4,3),\ (4,4),\ (4,5),\ (4,6) \\ (5,1),\ (5,2),\ (5,3),\ (5,4),\ (5,5),\ (5,6) \\ (6,1),\ (6,2),\ (6,3),\ (6,4),\ (6,5),\ (6,6) \end{array} \right\}$$

We can count that 21 of the sample points have sums less than or equal to seven. So the probability of C occurring is $P(C) = 21/36 = 7/12$. We also have that $P(A) = 1/6$. So $P(A)P(C) = 7/72$ but we can count that $A \cap C$ contains only one sample point and hence has a probability of $1/36$. Thus, $P(A)P(C) \neq P(A \cap C)$ so A and C are dependent events.

At first glance, we see that upon rolling a six as the first number you must roll a 1 for the sum to equal seven. So at first glance, events A and D seem to be independent however it would be naïve to assume this. We can count from the sample space that event D contains 6 points and so has a probability $P(D) = 6/36 = 1/6$ and $P(A) = 1/6$. So $P(A)P(D) = 1/36$. Now, we can count that the event $A \cap D$ contains only one point, (1,6) and so has a probability $P(A \cap D) = 1/36$. Therefore, $P(A \cap D) = P(A)P(D)$ and the events A and D are independent.

Chapter 5

Conditional Probability

Often we need to calculate the probability of some event A occurring while knowing that some other event B has already occurred. We call this the conditional probability of A **given** B and denote it by $P(A|B)$.

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

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

5.1 Theorems and Rules for Conditional Probability

Theorem 5.1.1

For any two events A and B defined on the same sample space, with $P(A) > 0$ and $P(B) > 0$, events A and B are independent if and only if $P(A|B) = P(A)$ or $P(B|A) = P(B)$.

Proof 5.1.1:

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

and by definition of independence, A and B are independent if and only if $P(A \cap B) = P(A)P(B)$ which is true if and only if $P(A|B) = P(A)$. Without loss of generality we can swap events A and B and arrive at the conclusion. □

Product Rules

Theorem 5.1.2

Let A, B, C and D be events on a sample space, with $P(A), P(B), P(C), P(D) > 0$. We have

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

and so on...

Proof 5.1.2:

The first statement come directly from the definition of conditional probability

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

For the second we have

$$\begin{aligned} P(A)P(B|A)P(C|A \cap B) &= P(A \cap B)P(C|A \cap B) && \text{by the first statement} \\ &= P(A \cap B)\frac{P(A \cap B \cap C)}{P(A \cap B)} && \text{by definition of conditional probability} \\ &= P(A \cap B \cap C) \end{aligned}$$

and so on... □

Law of Total Probability

Theorem 5.1.3

Let $A_1, A_2, A_3, \ldots, A_k$ be mutually exclusive events on a sample space and let B be an event on the same sample space. We have

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + \cdots + P(B \cap A_k) = \sum_{i=1}^{k} P(A_i)P(B|A_i)$$

Proof 5.1.3:

Note that the events $A_i \cap B$ for $1 \leq i \leq k$ are all mutually exclusive events since A_i's are mutually exclusive. Thus, the union of the $A_i \cap B$'s is B, that is

$$(A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \cup \cdots \cup (A_k \cap B) = B$$

So by Theorem 4.5.1, we have

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) + \cdots + P(A_k \cap B)$$

and by Theorem 5.1.2 (Product Rule), we have

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + \cdots + P(A_k)P(B|A_k) = \sum_{i=1}^{k} P(A_i)P(B|A_i)$$

as required. □

Bayes' Theorem

Theorem 5.1.4

Let A and B be events on a sample space, with $P(B) > 0$. We have

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

Proof 5.1.4:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} && \text{by Theorem 4.7.2 (Product Rule)} \\ &= \frac{P(B|A)P(A)}{P(A \cap B) + P(\overline{A} \cap B)} && \text{by Theorem 4.7.3 (Law of Total Probability)} \\ &= \frac{P(B|A)P(A)}{P(B|\overline{A})P(\overline{A}) + P(B|A)P(A)} && \text{by Theorem 4.7.2 (Product Rule)} \end{aligned}$$

□

Bayes' Theorem allows us to find the conditional probability of some event A given B, in terms of the probability of B given A. It allows us calculate conditional probabilities using the reversed order of conditioning.

5.2 Tree Diagrams

Tree diagrams are a technique that we can use to keep track of conditional probabilities. We start from a single node and draw new branches to separate nodes for each event. Each node represents the event occurring. On each branch we write the probability of event it leads to occurring. To find the probability of any node we multiply the probabilities of the branches leading to the node.

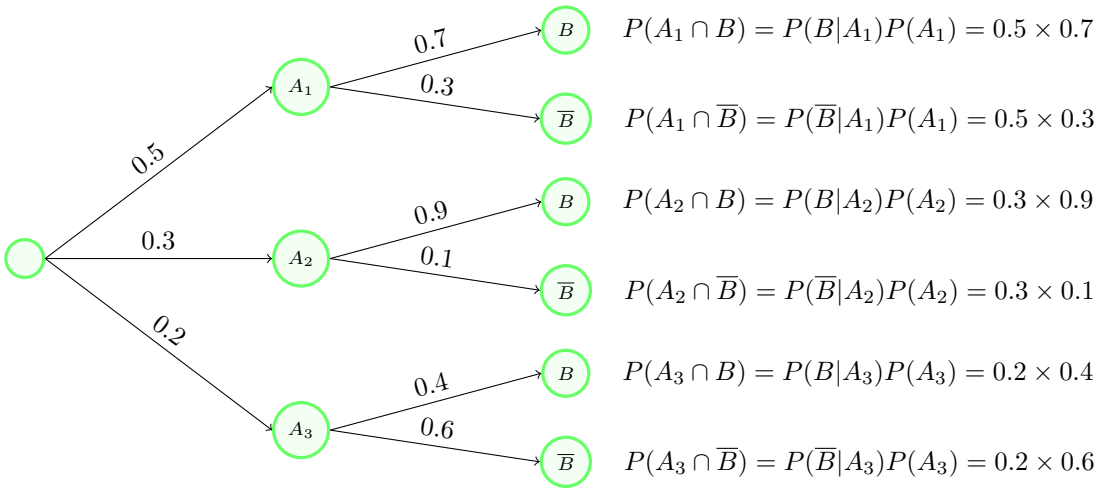


Figure 5.1: Tree diagram

The probability of all the the branches leading outward from each node must sum to 1 since at least one outcome must occur.

Chapter 6

Useful Sums and Series

This chapter includes a few useful sums and series that show up in the following chapters.

6.1 Geometric Series

$$\sum_{i=0}^{n-1} r^i = 1 + r + r^2 + \cdots + r^{n-1} = \frac{1 - r^n}{1 - r}$$

For $|r| < 1$, we have

$$\sum_{i=0}^{\infty} r^i = 1 + r + r^2 + \cdots = \frac{1}{1 - r}$$

Other identities can be obtained from this one by differentiation. For example we have

$$\frac{\mathrm{d}}{\mathrm{d}r} \sum_{i=0}^{\infty} r^i = \sum_{i=0}^{\infty} i r^{i-1} = \frac{\mathrm{d}}{\mathrm{d}r} \frac{1}{1 - r} = \frac{1}{(1 - r)^2}$$

6.2 Binomial Theorem

The binomial theorem describes the algebraic expansion of powers of a polynomial.

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

for any positive integer n and real number t.

A more general form of this theorem that holds even when n is not a positive integer is

$$(1 + t)^n = \sum_{x=0}^{\infty} \binom{n}{x} t^x, \quad \text{for } |t| < 1$$

It is an important skill to be able to recognize if an infinite, or otherwise, polynomial with binomial coefficients can be reduced to a simple polynomial raised to a power.

6.3 Multinomial Theorem

The multinomial theorem is a generalization of the binomial theorem. It describes the algebraic expansion of powers of a sum in terms of powers of the terms in the sum.

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{k_1+k_2+\cdots+k_m=n} \binom{n}{k_1, k_2, \ldots, k_m} \prod_{t=1}^{m} x_t^{k_t}$$

Another common form in which this theorem may be represented is

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{k_1+k_2+\cdots+k_m=n} \frac{1}{k_1!\, k_2! \cdots k_m!} (x_1^{k_1} x_2^{k_2} \cdots x_m^{k_m})$$

The summation is over all non-negative integers, $k_1, k_2, \ldots, k_m$ such that $k_1 + k_2 + \cdots + k_m = n$

6.4 Hypergeometric Identitiy

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

Proof 6.4.1:

We begin with the equality

$$(1+y)^{a+b} = (1+y)^a \times (1+y)^b$$

Now by Binomial Theorem we have

$$\sum_{k=0}^{a+b} \binom{a+b}{k} y^k = \sum_{i=0}^{a} \binom{a}{i} y^i \times \sum_{j=0}^{b} \binom{b}{j} y^j$$

Consider the coefficient of y^k on the right hand side. It is the sum of all the binomial terms such that $i + j = k$. Thus, the coefficient of y^k on the right hand side is

$$\sum_{i=0}^{\min\{a,k\}} \binom{a}{i} \binom{b}{k-i}$$

and since when $i > a$ or $i > k$ the term is 0 we can increase the sum to infinity. Thus, since the coefficient on the right hand side is equal to that on the left hand side we have

$$\binom{a+b}{k} = \sum_{i=0}^{\infty} \binom{a}{i} \binom{b}{k-i}$$

□

When x becomes significantly large, the terms of the summation become 0 since

$$\binom{n}{x} = \binom{n}{n-x} = 0, \quad \text{for } x > n$$

6.5 Exponential Series

This is an example of a Maclaurin series expansion.

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

The following limit definition of the exponential function is also useful

$$e^t = \lim_{n\to\infty} \left(1 + \frac{1}{t}\right)^n, \text{ for all } t \in \mathbb{R}$$

6.6 Integer Series

The following are useful equalities involving sums of integers.

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

Chapter 7

Discrete Random Variables and Probability Functions

7.1 Random Variables

A random variable (r.v. for short) is a numerical valued variable that represents the outcome of an experiment or process. Every random variable has a range associated with it, which is the set of all possible values the r.v. can take. We denote random variables by capital letters, e.g., A, X, Z.

Example 7.1.1

Suppose an experiment consists of tossing a coin three times. Let the random variable X be the number of heads that are rolled. And let the random variable Y be the number of tails rolled. Now, we have a nice short hand in that $X = 2$ is equivalent to the statement "two heads were rolled".

Moreover, we have useful equalities such as $X + Y = 3$ and $X = 3 - Y$.
The ranges of X and Y are both $\{\,0, 1, 2, 3\,\}$.

It is very important to understand the purpose of r.v.'s since the remainder of this course features them heavily.

The formal definition of a random variable is a function that assigns a real number to each point in a sample space.

Example 7.1.2

Consider the same experiment as above. The sample space is

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

Let us define X the number of heads that are rolled and the sample point $a = \text{HHT}$. The value of the function $X(a) = 2$ it is found by counting the number of heads in a. The range of X is $\{\,0, 1, 2, 3\,\}$. Each of the outcomes $X = x$ represents an event, simple or compound. In this case they are:

X	Event
0	$\{\,\text{TTT}\,\}$
1	$\{\,\text{TTH, THT, HTT}\,\}$
2	$\{\,\text{THH, HTH, HHT}\,\}$
3	$\{\,\text{HHH}\,\}$

Since some outcome in the range must always occur for a random variable for each event in the sample space, the events of a random variable are mutually exclusive subsets of the sample space such that their union is the total sample space. For r.v. X and outcome x, we have $X = x$ represents some event and we are interested in calculating its probability. We denote the probability of $X = x$ by $P(X = x)$.

Since the union of the events of values of a random variable is the total sample space, we have

$$\sum_{x \in \text{Range}(X)} P(X = x) = 1$$

Discrete Random Variables

Discrete random variables take integer values or, more generally, values in a countable set. Recall that a set is countable if its elements can be placed in a one-to-one correspondence with a subset of the positive integers.

Continuous Random Variables

Continuous random variables take values in some interval of real numbers like $(0, 1)$ or $(0, \infty)$ or (∞, ∞). You should be aware that there are infinitely numerical non-integer values that a r.v. with Range$(0, 1)$ could take. The values are separated by infinitesimally small intervals.

7.2 Probability Function

The probability function of a random variable X is a function that maps the value of X to the probability of that value. The probability function is represented by

$$f(x) = P(X = x) \text{ for } x \in \text{Range}(X)$$

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

Properties of Probability Functions

The following two properties hold for all probability functions.

- $f(x) \geq 0$ for all $x \in \text{Range}(X)$
- $\sum_{x \in \text{Range}(X)} f(x) = 1$

7.3 Cumulative Distribution Function

Another common function used to describe a probability model is the cumulative distribution function, usually denoted by $F(x)$. It is defined to be

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

Not that because the events "$X = x$" and "$X = y$" for $x \neq y$ are mutually exclusive we have

$$F(x) = P(X \leq x) = \sum_{z=0}^{x} P(X = z) \text{ for all } x \in \text{Range}(X)$$

It is the sum of the probabilities that the random variable takes values less than or equal to x.

Properties of Cumulative Distribution Functions

The following four properties hold for all cumulative distribution functions.

- $F(x)$ is a non-decreasing function
- $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$

Chapter 8

Discrete Distributions

As we briefly mentioned in the previous chapter, **probability distributions** are the set of pairs $(x, f(x))$ for all possible outcomes x of a random variable X. Many probability distributions appear commonly on r.v.'s of similar "real-life" processes. In this chapter we define a few of these common distributions on discrete random variables, when they occur and how to use them to calculate probabilities.

It is important to understand distributions early-on. Distributions, probability functions and cumulative distribution functions are defined on random variables **not** experiments/processes or sample spaces.

8.1 Uniform Distribution

Suppose X can take a finite set of consecutive values with each of the values being equally likely. That is $\text{Range}(X) = \{a, a+1, a+2, \ldots, b\}$ with each of $a, a+1, a+2, \ldots, b$ being equally likely. Then X has a discrete uniform distribution and we denote it
$X \sim$ *Discrete Uniform*

$$f(x) = P(X = x) = \begin{cases} \dfrac{1}{b-a+1} & \text{for all } x \in \text{Range}(X) \\ 0 & \text{otherwise} \end{cases}$$

Derivation of Probability Function

The probability of each value of the r.v. is easy to calculate since they are all equal and must add up to 1. Therefore, $k \times P(X = a) = 1$ where k is the number of possible values of X. The number of possible values of X is $b - (a-1) = b - a + 1$ since $\text{Range}(X)$ is between a and b inclusive.

Another way to define the probability of each value of a random variable with this sample space is

$$\frac{1}{\text{Number of possible values in Range}(X)}$$

8.2 Hypergeometric Distribution

Suppose we have a collections of N objects which can be classified into two different types, successes and failures. There are r successes and $N - r$ failures. We pick n objects at random without replacement, and

let the random variable X be the number of successes obtained. X has a hypergeometric distribution and we denote it
$X \sim Hypergeometric$

$$f(x) = P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \text{ for } x \leq \min(r, n)$$

Derivation of Probability Function

We will use the counting techniques we previously learnt to calculate the probability function. We note that there are $\binom{N}{n}$ ways to select n objects from the total of N so the sample space contains $\binom{N}{n}$ points. Now the number of ways of choosing x successes from the total of r is $\binom{r}{x}$ **and independently** the number of ways to choose the remainder of objects, $n-x$, from the total remaining objects, $N-r$, is $\binom{N-r}{n-x}$. Thus the probability of $X = x$ by the multiplication rule is the product of those expressions divided by the number of points in the sample space, $\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$.

It is important to understand that the terms "successes" and "failures" are simply placeholder that represent a type of outcome and its complement. They could be replaced by "wins" and "losses", "whites" and "colors", or any other titles that are distinct groups with a union that spans the whole sample space.

This is used when we know how many items (n) are chosen at random from a set with two different types and we know the amount of each type in the set.

Example 8.2.1

There is a basket with 11 fruit, 9 apples and 2 oranges. 4 fruit are picked at random from the basket. Let random variable X be the number of apples selected. Find $f(x) = P(X = x)$. Then find $f(3)$.
$X \sim Hypergeometric$. $N = 11, n = 4, r = 9$.

$$f(x) = P(X = x) = \frac{\binom{9}{x}\binom{2}{4-x}}{\binom{11}{4}}, \text{ for } x \leq 4$$

Hence

$$f(7) = P(X = 7) = \frac{\binom{9}{3}\binom{2}{1}}{\binom{11}{4}} \approx 0.509$$

Example 8.2.2

15 cards are drawn from a deck of 52 at random. Let X be the number of red cards drawn. Find $f(x) = P(X = x)$. Then find $f(7)$.

$X \sim$ *Hypergeometric*. $N = 52, n = 15, r = 26$.

$$f(x) = P(X = x) = \frac{\binom{26}{x}\binom{26}{15-x}}{\binom{52}{15}}, \text{ for } x \leq 15$$

Hence

$$f(7) = P(X = 7) = \frac{\binom{26}{7}\binom{26}{8}}{\binom{52}{15}} \approx 0.229$$

8.3 Binomial Distribution

Suppose we have an experiment with two distinct outcomes, success and failure, with the probability of a success being p and a failure being $(1 - p)$. The experiment is repeated n times independently (these are called trials). Let the random variable X be the number of successes obtained. X has a binomial distribution.
$X \sim$ *Binomial*(n, p)

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

Derivation of Probability Function

Since there are n positions in which to put the x successes there are $\binom{n}{x}$ unique arrangements of successes and failures that satisfy "$X = x$". Each of these arrangements has probability $p^x(1-p)^{n-x}$ since the probability of obtaining x successes is p^x and the probability of obtaining $n - x$ failures is $(1-p)^{n-x}$. So the probability that $X = x$, that is that any of the arrangements occur, is the sum of the probability of each unique arrangement, $\binom{n}{x}p^x(1-p)^{n-x}$.

The above formula describes the probability of x success and $(n-x)$ failures multiplied by the number of different ways of arranging those successes within the total number of trials of the experiment.

Each of the n individual experiments is called a "Bernoulli trial" and the entire process of n trials is called a Bernoulli process or a Binomial process.

Example 8.3.1

A loaded coin is flipped 10 times, with a probability of a heads occurring being 0.4. Let random variable X be the number of heads that occur. Find $f(x) = P(X = x)$, then find $f(3)$.
$X \sim$ *Binomial*$(10, 0.4)$.

$$f(x) = P(X = x) = \binom{10}{x}(0.4)^x(0.6)^{10-x}, \text{ for } x = 1, \ldots, 10$$

Hence

$$f(3) = P(X = 3) = \binom{10}{3}(0.4)^3(0.6)^7 \approx 0.215$$

Example 8.3.2

A football season in a university league has 22 games. The probability of each game being abandoned (because of bad weather or other hazards) is 0.02. Let X be the number of games abandoned throughout the whole season. Find $f(x) = P(X = x)$, then find $f(2)$ and $f(10)$.
$X \sim \textit{Binomial}(22, 0.02)$.

$$f(x) = P(X = x) = \binom{22}{x}(0.02)^x(0.98)^{22-x}, \text{ for } x = 1, \ldots, 22$$

Hence

$$f(2) = P(X = 2) = \binom{22}{2}(0.02)^2(0.98)^20 \approx 0.062$$

and

$$f(10) = P(X = 10) = \binom{22}{10}(0.02)^{10}(0.98)^{12} \approx 5.196\times10^{-12}$$

8.3.1 Comparison of Binomial and Hypergeometric Distributions

The Binomial and Hypergeometric distributions are similar in that they both model the distribution of the number of successes in n trials of an experiment. The difference is that the collection of objects in the hypergeometric distribution is selected from without replacement as apposed the Binomial distribution in which successes and failures do not affect the probability of future outcomes (with replacement).

The Hypergeometric distribution is used when there is a fixed number of objects (successes and failures) to choose from.

The Binomial distribution is used when there is no fixed number of objects to be selected from and instead we know the constant probability of a success for all the trials.

Example 8.3.3

Consider Lisa owns a car dealership and has only 750 red cars and 1250 blue cars in stock. A rich Swedish man enters and picks 50 cars randomly to purchase. Let X be the number of red cars the Swede purchases.

Since we know the number of successes (750 red cars) and failures (1250 blue cars) as well as the number of trials, we have that $X \sim \textit{Geometric}$ and

$$f(x) = P(X = x) = \frac{\binom{750}{x}\binom{1250}{50-x}}{\binom{2000}{50}}$$

Now, consider Lisa has run out of all her stock of cars. She goes to a Swedish car manufacturer's factory which is capable of producing any amount of cars. The factory has a 37.5% chance of producing a red car and otherwise produces a blue car. Lisa orders 50 cars. Let X be the number of red cars she receives.

Since there is no fixed number of cars to choose from but we do know the probability of each car being a success, we have that $X \sim \textit{Binomial}(50, 0.375)$ and

$$f(x) = P(X = x) = \binom{50}{x}(0.375)^x(0.625)^{50-x}$$

8.3.2 Binomial Estimate of the Hypergeometric Distribution

When the number of objects to choose from, N, is very large and the number of objects being chosen, n, is relatively small in a hypergeometric distribution, we have that the probability of a success changes only very slightly due to lack of replacement. Since the number of objects is so large choosing a small number of objects without replacement barely changes the probability of a success so p is relatively constant. Thus we can fairly accurately estimate the distribution with a binomial distribution with the original probability of a success for the first choice.

Example 8.3.4

Consider the previous example, suppose the rich Swedish man purchased 50 cars from Lisa. What is the probability that he purchases 20 red cars?

The number of cars Lisa has in stock is very large and the number of cars being bought is fairly small. Thus we can approximate the distribution with the probability of a success being $750/2000 = 0.375$. We have

$$f(20) = P(X = 20) = \binom{50}{20}(0.375)^{20}(0.625)^{30} \approx 0.1072$$

Now we can calculate the probability using the hypergeometric distribution to determine how good an estimate this is. We have

$$f(20) = P(X = 20) = \frac{\binom{750}{20}\binom{1250}{30}}{\binom{2000}{50}} \approx 0.1084$$

So the approximation is accurate to 2 decimal points.

8.4 Negative Binomial Distribution

This distribution is similar to the binomial distribution. We have an experiment with two distinct outcomes, success and failure, with the probability of a success being p and a failure being $(1 - p)$. The experiment is repeated until a specified amount of successes, k, have been obtained. Let the random variable X be the number of failures obtained before the k^{th} success. X has a negative binomial distribution.
$X \sim NB(k, p)$

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

Derivation of the Probability Function

The above formula describes the probability of x failures and k successes multiplied by the number of different ways of arranging those x failures within the total number of candidate trials $n + k - 1$. The final trial cannot be a failure as it is the k^{th} success.

The negative binomial distribution is used to model the number of trials of an experiment before the k^{th} success. Thus, if we know the number of trials, this distribution is not appropriate.

Example 8.4.1

A bad driver never stops at red lights and keeps driving and running red lights until he is arrested. The probability of him getting pulled over by a police man immediately after breaking a light is 0.53 and upon being pulled over 4 times he is arrested. Let X be number of red lights the driver runs

without being pulled over before he is arrested. Find $f(x) = P(X = x)$, then find $f(1)$ and $f(7)$.
$X \sim NB(4, 0.53)$

$$f(x) = P(X = x) = \binom{x+3}{x}(0.53)^4(0.47)^x, \text{ for } x = 0, 1, 2, \ldots$$

Hence

$$f(1) = P(X = 1) = \binom{4}{1}(0.53)^4(0.47) \approx 0.148$$

and

$$f(7) = P(X = 7) = \binom{10}{7}(0.53)^4(0.47)^7 \approx 0.048$$

Example 8.4.2

The probability of a football player scoring at least one goal in each game is 0.72. When the player scores in 26 games, she is awarded a bonus check. Let X be the number of games in which the player does not score before she is awarded the bonus. Find $f(x) = P(X = x)$, then find $f(7)$, and $f(0)$.
$X \sim NB(26, 0.72)$

$$f(x) = P(X = x) = \binom{x+25}{x}(0.72)^{26}(0.28)^x, \text{ for } x = 0, 1, 2, \ldots$$

Hence

$$f(7) = P(X = 7) = \binom{32}{7}(0.72)^{26}(0.28)^7 \approx 0.089$$

and

$$f(0) = P(X = 0) = \binom{25}{0}(0.72)^{26}(0.28)^0 = 0.72^{26} \approx 1.953 \times 10^{-4}$$

Note that $f(0)$ is simply the probability that the player scores in all of her first 26 games.

8.5 Geometric Distribution

This distribution is identical to the negative binomial distribution with $k = 1$. We have an experiment with two distinct outcomes, success and failure, with the probability of a success being p and a failure being $(1 - p)$. Let the random variable X be the number of failures obtained before the first success. X has a geometric distribution.
$X \sim Geometric(p)$

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

Example 8.5.1

A betting game involves flipping a coin repeatedly. The coin is fixed to that the probability of heads is 0.7 and tails is 0.3. On every flip, if you get heads you may flip again, but otherwise (if you get tails) the game is over. For each heads you flip you get $100. Let X be the number of heads you get. Find $f(2)$ and $F(3)$.

X is the number of trials before the first failure (flipping a tails) occurs. Thus, $X \sim Geometric$ so

$$f(x) = P(X = x) = (0.7)^x 0.3, \text{ for } x = 0, 1, 2 \ldots$$

Hence

$$f(2) = P(X = 2) = (0.7)^2 0.3 = 0.147$$

and

$$F(3) = P(X \leq 3) = (0.7)^3 0.3 + (0.7)^2 0.3 + (0.7)^1 0.3 + (0.7)^0 0.3 = 0.7599$$

8.6 The Poisson Distribution

This distribution is somewhat unlike the other distributions we have encountered. Suppose an event occurs an average of μ times per a specified interval (time, space, etc.) according to the following conditions:

- **Independence** - the number of occurrences in non-overlapping intervals are independent of one another.
- **Individuality** - events do not occur in clusters, that is, for sufficiently short intervals of length Δt, the probability of two or more events occurring is extremely close to 0 (negligible).
- **Homogeneity** - events occur at a uniform/homogeneous rate λ such that the probability of one occurrence in the interval $(t, t + \Delta t)$ is $\lambda \Delta t$ for small Δt.

Let X be the number of times the event occurs in the interval. X has a Poisson distribution.
$X \sim Poisson(\mu)$

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

Derivation of the Probability Function

Based on the conditions above we can derive the probability function. Let $f_t(x)$ be the probability of x occurrences in an interval of length t. Now we consider $f_{t+\Delta t}(x)$. We will use the the relationship between $f_t(x)$ and $f_{t+\Delta t}(x)$ and induction to show the result.

Firstly, we must find, $f_t(0)$, the probability of 0 occurrences in an interval of length t. Consider the probability, $f_{t+\Delta t}(0)$ that their are 0 occurrences in an interval of length $t + \Delta t$. That is, the probability of no events in the interval of length t and no events in the interval of length Δt. We have

$$f_{t+\Delta t}(0) = f_t(0)(1 - \lambda \Delta t)$$
$$\frac{f_{t+\Delta t}(0) - f_t(0)}{\Delta t} = -\lambda f_t(0)$$

As $\Delta t \to 0$, we have the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t(0) = -\lambda f_t(0)$$

which resolves to

$$f_t(0) = Ce^{-\lambda t}$$

Note for an interval of length 0, the probability of 0 zero events must be 1. So the constant C must be 1. Hence, we have

$$f_t(0) = e^{-\lambda t}$$

Note that there are only two ways to get a total of $x \neq 0$ occurrences in an interval of length $t + \Delta t$ for a sufficiently small Δt since by **individuality** the probability of two or more events in the interval

$(t, t+\Delta t)$ is negligible. Either there are x occurrences by time t or there are $(x-1)$ occurrences by time t and 1 in the interval $(t, t+\Delta t)$. This and the property of **independence** lead to

$$\begin{aligned} f_{t+\Delta t}(x) &= f_t(x)(1-\lambda\Delta t) + f_t(x-1)(\lambda\Delta t) \\ f_{t+\Delta t}(x) &= f_t(x) - f_t(x)(\lambda\Delta t) + f_t(x-1)(\lambda\Delta t) \\ \frac{f_{t+\Delta t}(x) - f_t(x)}{\Delta t} &+ \lambda f_t(x) = \lambda f_t(x-1) \end{aligned}$$

Now as $\Delta t \to 0$ we have the differential equation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} f_t(x) + \lambda f_t(x) &= \lambda f_t(x-1) \\ \frac{\mathrm{d}}{\mathrm{d}t}[e^{\lambda t} f_t(x)] &= e^{\lambda t}\lambda f_t(x-1) \end{aligned} \tag{8.1}$$

Now consider when $n = 1$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}[e^{\lambda t} f_t(1)] = e^{\lambda t}\lambda f_t(0)$$

Substituting the result, $f_t(0)$, we got earlier we have

$$\frac{\mathrm{d}}{\mathrm{d}t}[e^{\lambda t} f_t(1)] = e^{\lambda t}\lambda e^{-\lambda t} = \lambda$$

Integrating both sides we have

$$e^{\lambda t} f_t(1) = \lambda t + C$$

Note for an interval of length 0, the probability of 0 zero events must be 1. So the constant C must be 1. Hence, we have

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

We now use induction to generalize this result for an arbitrary x. Our inductive hypothesis is as follows

$$f_t(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

We have already shown that our hypothesis holds for $x = 0$ and $x = 1$. Now we assume our hypothesis is true and recall the differential equation (8.1) with $x + 1$. We have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}[e^{\lambda t} f_t(x+1)] &= e^{\lambda t}\lambda f_t(x) \\ \frac{\mathrm{d}}{\mathrm{d}t}[e^{\lambda t} f_t(x+1)] &= e^{\lambda t}\lambda \frac{(\lambda t)^x e^{-\lambda t}}{x!} = \frac{\lambda^{x+1} t^x}{x!} \\ e^{\lambda t} f_t(x+1) &= \int \frac{\lambda^{x+1}}{x!} t^x \,\mathrm{d}t = \frac{\lambda^{x+1}}{x!}\frac{t^{x+1}}{(x+1)} + C \end{aligned}$$

Again using the boundary condition, with an interval of length 0, we have that C must be 1. Thus, we have

$$f_t(x+1) = \frac{(\lambda t)^{x+1} e^{-\lambda t}}{(x+1)!}$$

Thus, if the inductive hypothesis holds for x then it also holds for $x+1$. So by principle of mathematical induction we have that the hypothesis is true for all natural numbers x.

Note that this derivation is fairly complex. If at first you do not understand don't worry. Try reading it again later.

8.6.1 Poisson Estimate to the Binomial Distribution

Derivation from Binomial Distribution

The Poisson distribution is a limiting case of the Binomial distribution as $n \to \infty$ and $p \to 0$. If we take $\mu = np$ as a constant as n tends to infinity we have that p tends to 0. Thus consider the following:

$$\begin{aligned}
f(x) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n^{(x)}}{x!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^n}{x!} \times \frac{n^{(x)}}{n^x} \left(1 - \frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^n}{x!} \times \frac{n \times (n-1) \times (n-2) \times \cdots \times (n-x+1)}{n^x} \times \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-x} \\
&\quad \text{(Note the middle term's numerator is the product of } n \text{ terms)} \\
&= \frac{\mu^n}{x!} \times \frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-x+1}{n} \times \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-x} \\
&= \frac{\mu^n}{x!} \times 1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{x-1}{n}\right) \times \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-x}
\end{aligned}$$

Now as n approaches infinity we have

$$\begin{aligned}
\lim_{n\to\infty} f(x) &= \frac{\mu^x}{x!} \times (1)(1)(1) \times \cdots \times (1) \times e^{-\mu}(1)^{-x}, \ \left(\text{since } e^k = \lim_{n\to\infty} \left(1 + \tfrac{k}{n}\right)^n\right) \\
&= \frac{\mu^x e^{-\mu}}{x!}, \ \text{ for } x = 0, 1, 2, \ldots
\end{aligned}$$

Thus, when n is very large and p is very small we can use the Poisson distribution to approximate a Binomial distribution.

8.6.2 Parameters μ and λ

The parameters μ and λ are, as you might expect, strongly linked. The first, μ, is a average number of occurrences in a specified interval, whereas λ is the uniform rate of occurrence per unit of the interval. Thus if t is the amount of units of time, space, etc., then $\lambda t = \mu$.

Example 8.6.1

Suppose a fire station gets 15 phone call every 5 minutes. The rate of occurrence per minute is $\lambda = 3$. Then, if we interested in the number of phone calls in 10 minutes, we have that the average number of phone calls in an interval of ten minutes is $\mu = 10\lambda = 30$.

8.6.3 Distinguishing the Poisson Distribution from other Distributions

In order to distinguish when to use and not to use a Poisson distribution, we can ask ourselves a few simple questions to rule out the Poisson process.

- *Is it reasonable to ask how often an event does not occur?*
- *Is it possible to specify an upper limit on the value the random variable in question can take?*

If the answer to either of these questions is yes, then the random variable in question does not follow a Poisson distribution and the experiment is not a Poisson process.

8.7 Combining Models

It is possible for the solution to a problem to require using more than one distribution to model the probability of a complex event. Below are a few examples where it is necessary to use more than one distribution.

Example 8.7.1

Suppose a type of spider catches flies in their webs at a rate of 2 per hour. If there are 10 such spiders, what is the probability that more than 6 spiders catch less than 4 flies in 2 hours?

First we find the probability that a single spider catches less than 4 flies in 2 hours. Let X be the number of flies the spider catches in 2 hours. The average number of flies caught in 2 hours is $\mu = 4$. Thus, we have $X \sim \textit{Poisson}(4)$.

$$P(X < 4) = \frac{4^0 e^{-4}}{0!} + \frac{4^1 e^{-4}}{1!} + \frac{4^2 e^{-4}}{2!} + \frac{4^3 e^{-4}}{3!} \approx 0.43$$

Now, we can find the probability that more than 6 spiders catch less than 4 flies in 2 hours with the knowledge that each spider has a 0.43 chance of doing so. Let Y be the number of spiders that catch less than 4 flies in 2 hours. We have $Y \sim \textit{Binomial}(10, 0.43)$.

$$\begin{aligned} P(Y > 6) &= \binom{10}{7} 0.43^7 (0.57)^3 + \binom{10}{8} 0.43^8 (0.57)^2 + \binom{10}{9} 0.43^9 (0.57)^1 + \binom{10}{10} 0.43^{10} (0.57)^0 \\ &\approx 0.081 \end{aligned}$$

It is important to remember that more than one distribution can be necessary. A common mistake is to correctly use one distribution and not realize the need for another.

Chapter 9

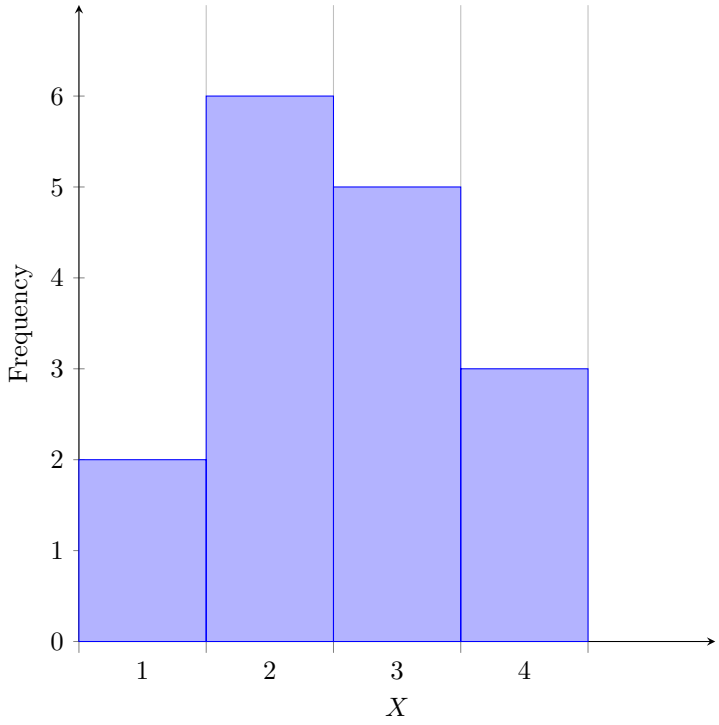
Mean and Variance

9.1 Summarizing Data on Random Variables

Often times, listing out all of the outcomes of a sample is not a very helpful way of communicating the information obtained from the sample. A common, more helpful way to present the data of a sample is a **frequency distribution**. A frequency distribution gives the number of times each value of a random variable X occurred.

X	Frequency Count	Frequency
1	$\|\|$	2
2	$\|\|\|\|\|$	6
3	$\|\|\|\|$	5
4	$\|\|\|$	3

We could also draw a **frequency histogram** of these frequencies.



Frequency distributions are good summaries of data because they show the variability in the observed outcomes clearly. Another way to summarize results are single-number summaries such as the following:

The **mean** of a sample of outcomes is the average value of the outcomes. It is the sum of the outcomes divide by the total number of outcomes. The mean of n outcomes, $x_1, \ldots, x_n$, for a random variable X is

$$\sum_{i=0}^{n} \frac{x_i}{n} = \frac{x_1, \ldots, x_n}{n}$$

The **median** of a sample is an outcome such that half the outcomes are before it and half the outcomes are after it when the outcomes are arranged in numerical order.

The **mode** of a sample is the outcome that occurs most frequently. There can be multiple equal modes in a sample.

Example 9.1.1

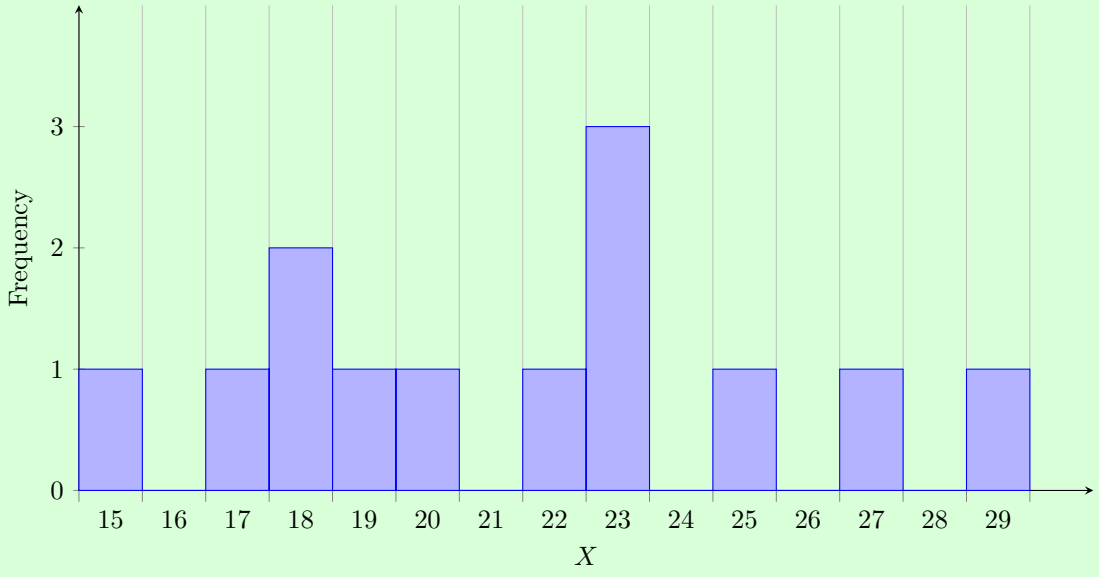
A fisherman records the weight of each fish he catches for a week. These are his results. Each value represents the weight, in pounds, of a fish he caught.

$$\{\, 20, 23, 19, 27, 17, 22, 18, 15, 23, 25, 18, 23, 29 \,\}$$

A frequency distribution of the sample above is

X	Frequency Count	Frequency
15	|	1
17	|	1
18	||	2
19	|	1
20	|	1
22	|	1
23	|||	3
25	|	1
27	|	1
29	|	1

And the following is a frequency histogram



The mean weight of sample of fishes is

$$\frac{20+23+19+27+22+18+15+23+18+23+29}{13} = \frac{237}{13} \approx 18.231$$

The median weight of the sample of fishes is found by rearranging the sample into numerical order and selecting the middle outcome. The median is 22.

$$15, 18, 18, 19, 20, \underset{\uparrow}{22}, 23, 23, 23, 27, 29$$

The mode is the weight that occurs most frequently. It corresponds to tallest bar on the histogram. 23lbs occurs the most (3 times) so it is the median.

9.2 Expected Value of a Random Variable

The expected value of a random variable X with a range of A and a probability function $f(x)$, is given by

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

Note that in order to calculate the expected value of a random variable X, we often need to know the distribution, and hence the probability function, of X.

Derivation of the Expected Value

Suppose we have a frequency distribution of a random variable X, as shown below:

X	Frequency Count	Frequency
5	𝍸 𝍸	10
10	𝍸 ||	7
25	𝍸 𝍸 |||	13
100	||||	4
200	||	1

As we learnt in the previous section, we can calculate the mean as

$$\begin{aligned} &\frac{(5\times 10)+(10\times 2)+(25\times 13)+(100\times 4)+(200\times 1)}{30} \\ &= (5)\left(\frac{10}{30}\right)+(10)\left(\frac{7}{30}\right)+(25)\left(\frac{13}{30}\right)+(100)\left(\frac{4}{30}\right)+(200)\left(\frac{1}{30}\right) \\ (A \text{ is the range of } X) \quad &= \sum_{x\in A} x \times \text{the fraction of times } x \text{ occurs} \end{aligned}$$

Now suppose we know the probability function of X is as follows

x	5	10	25	100	200
$f(x)$	$\frac{1}{3}$	$\frac{7}{30}$	$\frac{13}{30}$	$\frac{2}{15}$	$\frac{1}{30}$

Using the relative frequency definition of probability, we know that if we observed a very large number of outcomes, the fraction of times $X = x$ occurs (relative frequency of x) is $f(x)$.

Thus, *in theory*, we would expect the mean of a sample of infinitely many outcomes to be

$$(5)\left(\frac{1}{3}\right)+(10)\left(\frac{7}{30}\right)+(25)\left(\frac{13}{30}\right)+(100)\left(\frac{2}{15}\right)+(200)\left(\frac{1}{30}\right)\approx 33.167$$

This theoretical mean is denoted by μ or $E(X)$, and is known as the expected value of X.

Example 9.2.1

A slots machine in a casino costs \$5 to play. It has probabilities of 0.5 to pay out \$2, 0.2 to pay out \$5, a 0.1 to pay out \$10 and otherwise does not pay out anything. Let the random variable X be the amount of money (in dollars) the machine pays out in one play, and Y be the amount of money won or lost in one play. Find $E(X)$ and $E(Y)$.

$$E(X)=(0)(0.2)+(2)(0.5)+(5)(0.2)+(10)(0.1)=3$$

$$E(Y)=(-5)(0.2)+(-3)(0.5)+(0)(0.2)+(5)(0.1)=-2$$

Note that $E(Y)=E(X-5)=E(X)-5$

Example 9.2.2

A nightclub lets groups of up to 6 people enter at reduced fees. A randomly selected group in the nightclub's line has the following probabilities for its size and cost of entry:

Size of Group (X)	Cost of Entry (Y)	Probability
1	\$10	0.1
2	\$18	0.15
3	\$26	0.1
4	\$34	0.3
5	\$42	0.15
6	\$50	0.2

1. Let X be the size of a randomly selected group. Find $E(X)$.

$$E(X)=(0.1)(1)+(0.15)(2)+(0.1)(3)+(0.3)(4)+(0.15)(5)+(0.2)(6)=3.85$$

2. If the cost of entry of a group (Y) is $8\times$ the size of the group $+2$. Find the expected value of the cost of entry, in dollars, of a randomly selected group.

$$E(8X+2)=E(Y)=(0.1)(10)+(0.15)(18)+(0.1)(26)+(0.3)(34)+(0.15)(42)+(0.2)(50)=32.8$$

3. Show that the expected value of the cost of entry of a randomly selected group is $8\times$ the expected value of the size of the group $+2$.

$$8E(X)+2=8\times 3.85+2=30.8+2=32.8=E(8X+2)$$

Theorem 9.2.1

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

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

Proof 9.2.1:

Let the random variable $Y = g(X)$ have a range of B and a probability function $f_Y(y) = P(Y = y)$.

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

Now, let C_y be $\{\, x \mid g(x) = y \,\}$, that is the set of all values of x such that $g(X)$ is y. So

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

That is, the probability that $Y = y$ is the sum of the probabilities that $X = x$ such that $g(x) = y$.
Now, we have

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

Note that the inner summation is for all x such that $g(x) = y$ and the outer is for all y. Thus the equation is the sum for all x. So

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

where A is the range of X, as required. □

9.3 Linear Properties of Expected Value

Theorem 9.3.1

For constants a, b and c,

$$E[a g_1(X) + b g_2(X) + c] = a E[g_1(X)] + b E[g_2(X)] + c$$

Proof 9.3.1:

$$\begin{aligned}
E[aE[g_1(X)] + bE[g_2(X)] + c] &= \sum_{\text{all } x} [ag_1(x) + bg_2(x) + c]f(x) \\
&= \sum_{\text{all } x} [ag_1(x)f(x) + bg_2(x)f(x) + cf(x)] \\
&= \sum_{\text{all } x} ag_1(x)f(x) + \sum_{\text{all } x} bg_2(x)f(x) + \sum_{\text{all } x} cf(x) \\
&= a\sum_{\text{all } x} g_1(x)f(x) + b\sum_{\text{all } x} g_2(x)f(x) + c\sum_{\text{all } x} f(x) \\
\left(\text{recall } \sum_{\text{all } x} f(x) = 1\right) \qquad &= aE[g_1(X)] + b[g_2(X)] + c
\end{aligned}$$

□

9.4 Variance of a Random Variable

The variance of a random variable X is given by

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

where σ is the standard deviation of X. It is the average squared deviation of a random variable from its mean. It measures how far out from the mean the values of a random variable are spread.

The definition and formula above is useful in understanding the variance's importance but it can be difficult to use to actually calculate the variance. Here are a few other useful formulas for calculating the variance of a random variable:

$$Var(X) = E(X^2) - [E(X)]^2 = E(X^2) - \mu^2 \tag{9.1}$$

$$Var(X) = E[X(X-1)] + E(X) - [E(X)]^2 = E[X(X-1)] + \mu - \mu^2 \tag{9.2}$$

Derivation of Alternative Formulas

$$\begin{aligned}
Var(X) = \sigma^2 &= E\left[(X - \mu)^2\right] \\
&= E\left[X^2 - 2X\mu + \mu^2\right] \\
&= E(X^2) - 2\mu E(X) + \mu^2, \text{ (by linear property since } \mu \text{ is a constant)} \\
&= E(X^2) - 2\mu^2 + \mu^2, \text{ (since } E(X) = \mu) \\
&= E(X^2) - \mu^2
\end{aligned}$$

Now note that $X^2 = X(X-1) + X$, so we have

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

Chapter 10

Continuous Random Variables

10.1 Computer Generated Random Variables

Virtually all computer software have built in "psuedo-random number generators" that simulate observations of a random variable U, from a uniform distribution, $U(0,1)$. From this uniform distribution, we can apply functions on U in order to form non-uniform distributions with a given cumulative distribution functions $F(X)$.

Theorem 10.1.1

If F is an arbitrary cumulative distribution function and U is uniform on $[\,0,1\,]$ then the random variable X, defined by $X = F^{-}(U)$, where $F^{-}(y) = \min\{\, x \mid F(x) \geq y \,\}$, has a cumulative distribution function of $F(x)$.

Proof 10.1.1:

Note that, for all $U < F(x)$, we have that $X \leq x$ by applying the inverse function F^{-} to both sides. Now, by applying F to both sides of $X \leq x$, we have $U \leq F(x)$, for all x. So we can say

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

Taking probabilities of the across the equation we have,

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

Note that $P[U < F(x)] = P[U \leq F(x)] = F(x)$ since U is uniform and continuous. Thus,

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

so $P[X \leq x] = F(x)$, for all x, as required. □

Example 10.1.1

Suppose we have a random variable U that is uniform on $U[\,0,1\,]$ and we want to generate a random variable X with exponential distribution. We have that the cumulative distribution function of X is $F_X(x) = 1 - e^{-\lambda x}$ for some λ. Since $F_X(x)$ is a continuous, strictly increasing function for $x > 0$,

let $y = F_X(x)$. Now,

$$\begin{aligned} y &= 1 - e^{-\lambda x} \\ 1 - y &= e^{-\lambda x} \\ \ln(1-y) &= -\lambda x \\ x &= \frac{-\ln(1-y)}{\lambda} \end{aligned}$$

So $F_X^{-1}(y) = \dfrac{-\ln(1-y)}{\lambda}$.
Thus, by Theorem 3.1.1, $X = F_X^{-1}(U)$ has cumulative distribution function $F_X(x)$.

$$X = -\frac{1}{\lambda}\ln(1-U)$$

Now, to find $f_X(x)$, the probability density function of X, we differentiate the cumulative distribution function. So,

$$f_x(x) = \frac{1-y}{\lambda}$$

10.2 Normal Distribution

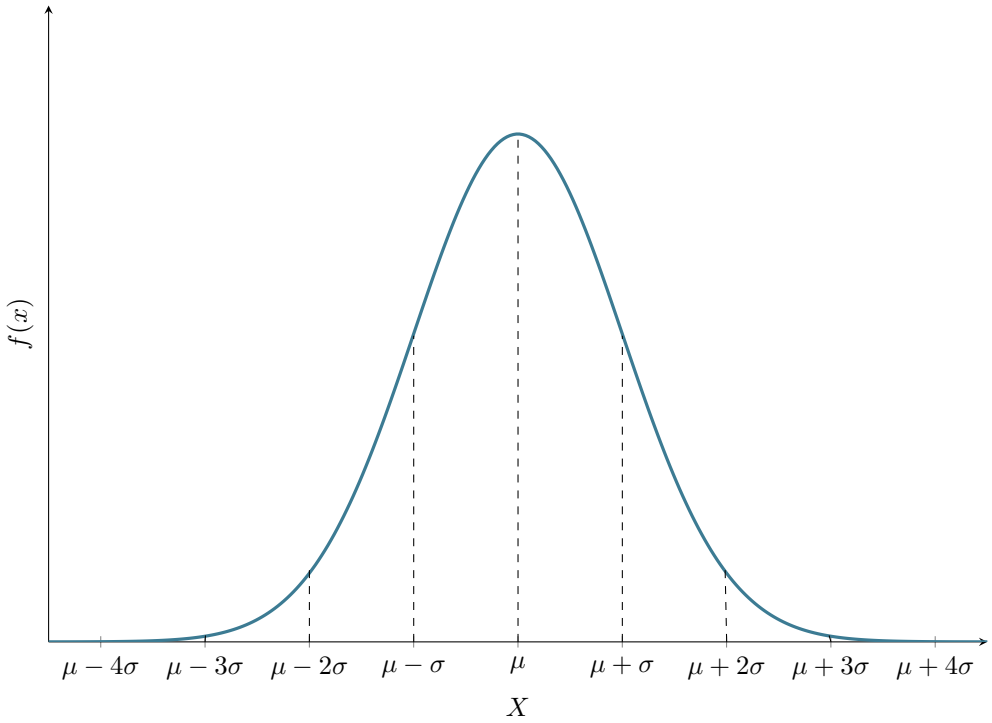
A random variable X has a normal distribution if it has probability density function of the form

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

where $\mu \in \mathbb{R}$ and $\theta > 0$ are parameters. It turns out that $E(X) = \mu$ and $Var(X) = \sigma^2$.
$X \sim N(\mu, \sigma^2)$ where X has expected value μ and variance σ^2.

The Normal distribution is the most widely used distribution in probability and statistics. Physical processes leading to the Normal distribution exist but are a little complicated to describe.

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



The **cumulative distribution function** of a Normal Distribution is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \, \mathrm{d}y$$

This integral cannot be given a simple mathematical expression so numerical methods are used to compute its value for given x, μ, σ. Before computers could solve such problems, tables of probabilities $F(x)$ where created by numerical integration. Only the table of the standard normal distribution, $N(0,1)$, is required to solve for $F(x)$ for all μ, σ, since with a change of variable, the c.d.f. any normal distribution can be related to that of the standard normal distribution.

Theorem 10.2.1

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

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

Proof 10.2.1:

Let $X \sim N(\mu, \sigma^2)$.

$$\begin{aligned} P(X \leq x) &= \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \, \mathrm{d}y \\ \left(\text{Let } z = \frac{y-\mu}{\sigma}\right) \quad &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\ &= P\left(Z \leq \tfrac{x-\mu}{\sigma}\right) \end{aligned}$$

□