Applications Of Probability

Niyathi Srinivasan, Yuxian "Iris" Liu, Bella Qu

May 21, 2023

Abstract

The applications of probability, which stems from the study of the likeliness of certain events, determine much of the workings of our world, whether we realize it or not. This paper explores several ideas and concepts in probability theory, starting from the fundamental axioms of probability and then discussing conditional probability, and independence and dependence of events. We will apply these fundamental theories to more complicated distributions of random events, discussing discrete random variables (the Binomial, Poisson, and Geometric distributions especially), and continuous random variables. Further, we will discuss the properties of expected value and variance of a data set, all the while solving challenging, real-world problems to support our conjectures.

1 Introduction

1.1 Definitions

Definition 1. *A* **sample space** *is defined to be the union of all outcomes, which is denoted by S.*

Definition 2. *An* **event** *is a subset under the sample space; it consists of all possible outcomes of an experiment. It is often denoted by E.*

Definition 3. *A* ***union*** *of two events is denoted as $E \cup F$, which represents the outcomes that are in the event E, F, or both E and F.*

Definition 4. *An* ***intersection*** *represents the outcomes that are in both events E and F and is denoted by EF or $E \cap F$.*

Definition 5. *A* ***complement****, denoted as E^c, represents all outcomes that are not contained in the event E.*

2 Axioms of Probability

2.1 Definitions

We will start by defining some basic axioms of probability:

1. For any event E, $0 \leq p(E) \leq 1$, where $p(E)$ is the probability that the event occurs. In other words, the probability of an event occurring is always between 0 and 1.

2. Let p_i be individual probabilities of events in a sample space. Then,

$$\sum_{i=1}^{\infty} p_i.$$

   In other words, the sum of all probabilities within a sample space is equal to 1.

3. Let E_i be one of i events that could occur. For all such mutually exclusive events E, we have

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i).$$

   In other words, the union of all events is equal to the sum of the individual probabilities of the events.

2.2 Some propositions

From the central axioms mentioned above, we can now introduce several propositions that will be useful to the development of more complex ideas later on.

Proposition 1. *Let E be an event and let $P(E)$ be the probability that the event E occurs. Then, the probability that the complement of E occurs is $P(E)^c$, where*

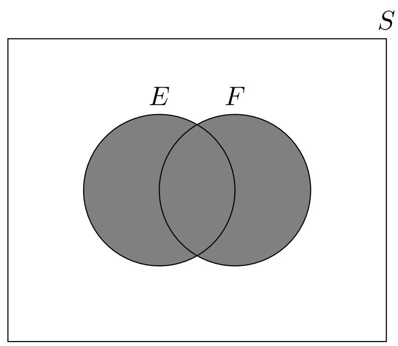
$$P(E)^c = 1 - P(E).$$

This is intuitive as the sum of the event and its complement will be equal to the sample space, which has probability 1.

Proposition 2. *Let E and F be two different events. Then the union of E and F can simply be expressed as*

$$P(E \cup F) = P(E) + P(F) - P(EF).$$

One can think about this proposition as if it was a problem on finding the area of two overlapping shapes: simply sum the individual areas and subtract the overlapping area.



We can extend this proposition to form a more general statement.

Proposition 3. *Let $E_1, E_2, ..., E_i$ be independent events. Then the probability of the union of all those events can be expressed in the **Inclusion-Exclusion Identity**, which states that*

$$\begin{aligned} P(E_1 \cup E_2 \cup ... \cup E_n) = &\sum_{i=1}^{n} P(E_i) - \sum_{i_1<i_2} P(E_{i_1} \cdot E_{i_2}) + ... \\ &+ (-1)^{r+1} \sum_{i_1<i_2<...<i_r} P(E_{i_1} \cdot E_{i_2} ... E_{i_r}) + ... \\ &+ (-1)^{n+1} P(E_1 E_2 ... E_n). \end{aligned}$$

2.3 The Matching Problem

Suppose there are N men at a party and they all throw their hats into a box in the center of the room. The hats are then mixed up and each man randomly takes one of the hats. What is the probability that no man will get his own hat?

We can approach this problem by finding the complementary of what the problem asked for, i.e. the probability that at least one man selects his own hat. Let E_i be the event that the ith man selects his own hat, where $i = 1, 2, 3, ..., N$, and let n be the number of men that selects his own hat. By the inclusion-exclusion identity, we can evaluate the probability that at least one of the men selects his own hat to be

$$\begin{aligned} P(E_1 \cup E_2 \cup ... \cup E_n) = &\sum_{i=1}^{N} P(E_i) - \sum_{i_1<i_2} P(E_{i_1} E_{i_2}) + ... \\ &+ (-1)^{n+1} \sum_{i_1<i_2<...<i_n} P(E_{i_1} \cdot E_{i_2} ... E_{i_n}) + ... \\ &+ (-1)^{N+1} P(E_1 E_2 ... E_N). \end{aligned}$$

If we view the outcomes of this experiment as an ordered list, where the *ith* element is the number of hat that the *ith* man selects, then there are a total of $N!$ possible outcomes.

We can also see that the event $E_{i_1}E_{i_2}...E_{i_n}$, which represents the event that each of the n men selects his own hat, can occur in $(N-n)!$ ways. This holds since for the remaining $N-n$ men, the first one can select any of the $N-n$ hats, the second one can select any of the $N-n-1$ hats, and so on.

Thus, if we assume all possible outcomes to be equally likely ($P(E_{i_1}) = P(E_{i_2}) = ... = P(E_{i_n})$, then

$$P(E_{i_1}E_{i_2}...E_{i_n}) = \frac{(N-n)!}{N!}.$$

Additionally, since there are $\binom{N}{n}$ terms in $\sum_{i_1<i_2<...<i_n} P(E_{i_1} \times E_{i_2}...E_{i_n})$, by expanding and simplifying, we get

$$\frac{N!}{(N-n)!n!}\frac{(N-n)!}{N!} = \frac{1}{n!}.$$

This is one over all possible permutations of n men selecting their own hat.

Hence, the probability $P(E_1 \cup E_2 \cup ... \cup E_n)$ now becomes

$$1 - \frac{1}{2!} + \frac{1}{3!} - ... + (-1)^{N+1}\frac{1}{N!}$$

and the desired probability (that no man selects his own hat) is simply

$$1 - 1 + \frac{1}{2!} - \frac{1}{3!} + ... - (-1)^{N+1}\frac{1}{N!}$$

.

This can be rewritten as

$$\sum_{i=0}^{N} \frac{(-1)^i}{i!}$$

.

We can then apply an identity, which states that $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. Upon letting $x = -1$, we can see that, for very large values of N, the desired probability tends to

$$e^{-1} \approx 0.3679,$$

which is about 37%.

Notice that this is very different from what one would expect, which is that probability of no man selecting his own hat tends to 1 when N is very large.

3 Conditional Probability and Independence

3.1 Conditional Probability

Conditional probability is used when the probability of an event happening depends on the occurrence of another event. Given two events E and F, the conditional probability that E occurs given that F has occurred is denoted by $P(E|F)$. Since E can only occur if F occurred, essentially we are trying to find the probability of both events happening within the new sample space F. The definition and general formula for $P(E|F)$ if $P(F) > 0$ is

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

3.1.1 The Multiplication Rule

Multiply both sides of the general formula by $P(F)$ and we obtain

$$P(EF) = P(F)P(E|F).$$

A generalization of the above equation gives

$$P(E_1E_2E_3 \cdots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2) \cdots P(E_n|E_1 \cdots E_{n-1}).$$

This is sometimes referred to as the *multiplication rule*, and it gives the probability of the intersection of n number of events.

3.1.2 Bayes's Formula

If E and F are events, E can be expressed as

$$E = EF \cup EF^c.$$

As EF and EF^c are mutually exclusive, by axiom 3 we have

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c). \end{aligned}$$

Equivalently, we have Proposition 4 as follows.

Proposition 4.

$$P(E) = P(E|F)P(F) + P(E|F^c)[1 - P(F)]$$

Essentially, this expression states that the probability of E is a weighted average of the conditional probability of E (given that F has occurred) and the conditional probability of E (given that F has not occurred). Each conditional probability has as much weight as the probability of the condition happening. This formula is extremely useful because sometimes it is much easier to calculate the probability of an event when we know whether or not another event it depends on has occurred.

The next proposition is referred to as the *law of total probability*, which states that for given events F_1, F_2,..., F_n, of which one and only one must occur, $P(E)$ is equal to a weighted average of $P(E|F_i)$. Each term is weighted by the probability of the vent on which it is conditioned.

Proposition 5.

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

Proof. Proposition 4 can be generalized as

$$\bigcup_{i=1}^{n} F_i = S.$$

When exactly one of the events F_1, F_2,..., F_n must occur,

$$E = \bigcup_{i=1}^{n} EF_i.$$

Knowing that events F_1, F_2,..., F_n are mutually exclusive, we obtain

$$\begin{aligned} P(E) &= \sum_{i=1}^{n} P(EF_i) \\ &= \sum_{i=1}^{n} P(E|F_i)P(F_i). \end{aligned}$$

□

Let F_1, F_2,..., F_n be a set of mutually exclusive events where exactly one of these events must occur. If E has occurred and we are trying to determine which one of F_j also occurred, then by Proposition 5, we have obtain the following proposition.

Proposition 6.

$$\begin{aligned} P(F_j|E) &= \frac{P(EF_j)}{P(E)} \\ &= \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)} \end{aligned}$$

Proposition 6 is known as *Bayes's Formula*, and this formula is extremely useful because sometimes it is much easier to calculate the probability of an event when we know whether or not another event it depends on has occurred.

3.1.3 Crime Investigation Problem

This problem is adapted from [5]. To see how conditional probability is used in real life, let us look at a problem regarding crime investigation. A crime has been committed by an individual who left some of their DNA at the crime scene. Scientists who studied the recovered DNA state that each innocent person individually has a probability of 10^{-5} of matching the recovered DNA. The attorney supposes that any of the 1 million residents o the town could have committed the crime. Ten thousand of those residents have been released from prison within the past 10 years, and a copy of their DNA is on file. Each of the 10,000 ex-criminals has a probability α of being guilty of the crime, and each of the remaining 990,000 residents has probability β. Each ex-criminal is c times more likely to be the perpetrator than those who were not past convicts. When the database of the ex-criminals is analyzed, A.J. is the only one whose DNA matches the recovered DNA. What is the probability that A.J. is guilty?

Solution Since probability sum up to 1, we have

$$1 = 10,000\alpha + 990,000\beta = (10,000c + 990,000)\beta.$$

Thus,

$$\beta = \frac{1}{10,000c + 990,000}, \quad \alpha = \frac{c}{10,000c + 990,000}.$$

Let G denote the event that A.J. is guilty, and let M denote the event that A.J. is the only one of the 10,000 to match the recovered DNA. Then,

$$\begin{aligned} P(G|M) &= \frac{P(GM)}{P(M)} \\ &= \frac{P(G)P(M|G)}{P(M|G)P(G) + P(M|G^c)P(G^c)}. \end{aligned}$$

If A.J. is guilty, he will be the only one to have a DNA match if no other DNA on file matches. Therefore,

$$P(M|G) = (1 - 10^{-5})^{9999}.$$

However, if A.J. is innocent, then in order for his DNA to be the only match, his DNA must match. Additionally, all other DNA on file must be innocent and not match. Given that A.J. is innocent, the probability that all other ex-criminals are also innocent is

$$\begin{aligned} P(\text{all others innocent—AJ innocent}) &= \frac{P(\text{all other ex-criminals innocent})}{P(\text{AJ innocent})} \\ &= \frac{1 - 10,000\alpha}{1-\alpha}. \end{aligned}$$

Given their innocence, the probability of all other ex-criminals matching the recovered DNA is $(1 - 10^{-5})^{9999}$. Therefore,

$$P(M|G^c) = 10^{-5}(\frac{1 - 10,000\alpha}{1-\alpha})(1 - 10^{-5})^{9999}.$$

Since $P(G) = \alpha$, we have

$$P(G|M) = \frac{\alpha}{\alpha + 10^{-5}(1 - 10,000\alpha)} = \frac{1}{0.9 + \frac{10^{-5}}{\alpha}}.$$

If an ex-criminal was 10 times more likely to commit the crime than non-convicts, then $\alpha = \frac{1}{109,000}$ and

$$P(G|M) = \frac{1}{1.99} \approx 0.5025.$$

3.2 Independence

E is independent of F if $P(E|F) = P(E)$. That is, the knowledge that F has occurred does not change the probability of E. Since $P(E|F) = \frac{P(EF)}{P(F)}$, two events are independent if $P(EF) = P(E)P(F)$.

Three events E, F, and G are said to be independent if

$$\begin{aligned} P(EFG) &= P(E)P(F)P(G) \\ P(EF) &= P(E)(P(F) \\ P(EG) &= P(E)P(G) \\ P(FG) &= P(F)P(G). \end{aligned}$$

That is, E, F, and G are independent if E is independent of any event formed by F and G. If there are more than three events, events E_1, E_2, ..., E_n are independent if for every subset $E_{1'}$, $E_{2'}$, ..., $E_{r'}$, $r \leq n$ of those events

$$P(E_{1'}, E_{2'}, \ldots, E_{r'}) = P(E_{1'})P(E_{2'})\ldots P(E_{r'}).$$

An infinite set of events are independent if every finite subset of those events is independent.

3.3 The Gambler's Ruin Problem using Conditional Probability

Next, let's try to solve a classic probability problem using conditional probability and independence. Suppose two gamblers, A and B, bet on the outcomes of successive coin flips. On each flip, if the result is a head, A collects 1 unit from B, and if the result is a tail, A pays 1 unit to B. They continue to do this until one gambler runs out of money and loses. If successive coin flips are independent and each flip results in a head with probability p, what is the probability that A wins if A start with i units and B starts with $N - i$ units?

Proposition 7.

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)^N} & \textit{if } p \neq \frac{1}{2} \\ \frac{i}{N} & \textit{if } p = \frac{1}{2} \end{cases}$$

Proof. This solution is adopted from [5]. Let E denote the event that A wins when they start with i units, and let $P_i = P(E)$. We can first obtain an expression for P(E) by making the outcome of the first flip a condition. Let H denote the event that the first flip lands on a head. Then,

$$\begin{aligned} P_i = P(E) &= P(E|H)P(H) + P(E|H^c)P(H^c) \\ &= p \times P(E|H) + (1-p) \times P(E|H^c). \end{aligned}$$

If the first flip lands on heads, then A has $i+1$ units and B has $N-(i+1)$ units. Since the coin flips are independent of each other, A's probability of winning is now exactly as if the game started with A having an initial amount of $i+1$ units and B having an initial amount of $N-(i+1)$ units. Therefore,

$$\begin{aligned} P(E|H) &= P_{i+1} \\ P(E|H^c) &= P_{i-1}. \end{aligned}$$

If we let$q = 1-p$, we obtain

$$P_i = p \cdot P_{i+1} + q \cdot P_{i-1} \text{ for } i = 1, 2, \ldots, N-1.$$

The boundary conditions $P_0 = 0$ and $P_N = 1$ allow us to solve the above equation. Since $p + q = 1$, then

$$p \cdot P_i + q \cdot P_i = p \cdot P_{i+1} + q \cdot P_{i-1},$$

or

$$P_{i+1} - P_i = \frac{q}{p}(P_i - P_{i-1}) \text{ for } i = 1, 2, \ldots, N-1.$$

Since $P_0 = 0$, from the above equation, we obtain

$$\begin{aligned} P_2 - P_1 &= \frac{q}{p}(P_1 - P_0) = \frac{q}{p}P_1 \\ P_3 - P_2 &= \frac{q}{p}(P_2 - P_1) = (\frac{q}{p})^2 P_1 \\ &\vdots \\ P_I - P_{i-1} &= \frac{q}{p}(P_{i-1} - P_{i-2}) = (\frac{q}{p})^{i-1} P_1 \\ &\vdots \\ P_N - P_{N-1} &= \frac{q}{p}(P_{N-1} - P_{N-2} = (\frac{q}{p})^{N-1} P_1. \end{aligned}$$

Adding the first $i-1$ equations from the above equations, we obtain

$$P_i - P_1 = P_1 \left[(\frac{q}{p} + (\frac{q}{p})^2 + \ldots + (\frac{q}{p})^{i-1} \right]$$

or

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)} P_1 & \text{if } \frac{q}{p} \neq 1 \\ iP_1 & \text{if } \frac{q}{p} = 1 \end{cases}$$

Knowing the fact that $P_N = 1$, we obtain

$$P_1 = \begin{cases} \frac{1-(q/p)}{1-(q/p)^N} & \text{if } \frac{q}{p} \neq \frac{1}{2} \\ \frac{1}{N} & \text{if } p = \frac{1}{2} \end{cases}$$

Therefore,

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

□

Let Q_i denote the probability that B wins when A starts with i and B starts with $N-i$. Then by symmetry, we can replace p by q and i by $N-i$. Therefore, we have

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

Furthermore, since $q = \frac{1}{2}$ is equal to $p = \frac{1}{2}$, when $q \neq \frac{1}{2}$, we have

$$\begin{aligned} P_i + Q_i &= \frac{1-(q/p)^i}{1-(q/p)^N} + \frac{1-(p/q)^{N-i}}{1-(p/q)^N} \\ &= \frac{p^N - p^N(q/p)^i}{p^N - q^N} + \frac{q^N - q^N(p/q)^{N-i}}{q^N - q^N} \\ &= \frac{p^N - p^{N-i}q^i - q^N + q^i p^{N-i}}{p^N - q^N} \\ &= 1 \end{aligned}$$

This equation holds when $p = q = \frac{1}{2}$, so

$$P_i + Q_i = 1.$$

Essentially, this equation shows that either A or B will win with probability 1, which means there does not exist a scenario where the game goes on forever with nobody winning.

4 Random Variables

A random variable is a variable whose value is unknown, represented by a function that assigns values to each of an experiment's outcomes. Random variables can be classified as discrete or continuous and are used to quantify the outcome of a random experiment.

For instance, let a random variable X be the number of heads flipped when flipping three coins. We are typically concerned only with the outcome of the random variable (the number of heads flipped), not the outcome of each flip (i.e., that there was 1 head flipped, irrespective of the actual outcome (H,T,T), (T,H,T), or (T,T,H)).

4.1 Measures of Spread and Center

4.1.1 Expected Value

We can now calculate the expected value of a random variable X, denoted as $E[X]$:

$$E[X] = \sum_{x:p(x)>0} x \cdot p(x).$$

Expected value, in simple terms, is the predicted value of a random variable, calculated as the sum of all possible values of the random variable, with each value being multiplied by the probability of its occurrence.

Essentially, $E[X]$ is a weighted average of all the possible values of X where the weight of each value is the probability that X takes on that value.

4.1.2 Variance

Although the expected value of X gives us the weighted average of the random variable X, the spread of these values is another important property of X. To measure the spread of a dataset, a property called **variance** can be used. Variance determines how well the mean represents an entire data set. One way to measure it is to calculate the average of the deviations from the mean, denoted as

$$\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}.$$

However, this method of measuring variance is often time-consuming and difficult to calculate, especially for larger random variables. We can derive an easier way of calculating the variance, using $E[X]$.

Proposition 8. $E[X^2] - E[X]^2$

Proof. Let $E[X] = \mu$

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

□

4.2 Discrete Random Variables

One subcategory of random variables are **discrete random variables**, which are random variables that can only take a countable amount of whole number values as outcomes of a random experiment.

For a discrete random variable X, the probability mass function $p(a)$ of X, is denoted by $p(a) = P(X = a)$, where $p(a)$ is only positive for a countable number of values for a.

The probability mass function, $p(a)$, is positive for a countable number of values of a. Assuming that X is one of x_1,x_2, . . . x_n, then

$$p(x_i) \geq 0 \text{ for } i = 1,2,... \tag{1}$$

$$p(x) = 0 \text{ for all other values of } x. \tag{2}$$

Since the total probabilities of all possible values for X must sum to 1, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1,$$

where $p(x_i)$ are the individual probabilities of the countable number of events.

In real life, a discrete random variable could be represented as the number of passengers on a train or the number of defective computers in a box of 100.

4.2.1 Binomial

A **Binomial random variable** is a random variable that can only have two outcomes, a success or a failure in an experiment consisting of n trials. $Bin(X)$ is usually equal to 1 when the outcome of the trial is a success and equal to 0 when the outcome is a failure. If X represents the number of successes that occur in those n trials, then X is a binomial random variable with parameters (n, p). The probability mass function of $Bin(X)$ is given by:

$$p(0) = P(X = 0) = 1 - p \tag{3}$$

$$p(1) = P(X = 1) = p \tag{4}$$

where n is the number of trials, and where p, $0 \leq p \leq 1$, is the probability that one trial is a success. A random variable X is said to be a Binomial Random Variable if its probability mass function is given by some $p \in [0, 1]$.

To get the probability of having a k number of desired successes in n trials for a Binomial random variable, we use

$$P(X = k) = \binom{n}{k} \cdot p^k (1 - p)^{n-k}. \tag{5}$$

The expected value of this type of variable is np and the variance is $np \times (1 - p)$.

Consider an experiment where you throw 12 dice, with each outcome being independent of one another. What is the probability of getting at least two sixes in the 12 rolls? To compute, we can combine the binomial distribution with the complement rule. Let X be the number of sixes rolled after throwing 12 dice, and let

$$\begin{aligned}
P(X \geq 2) &= 1 - P(X \leq 1) \\
&= 1 - (P(X = 1) + P(X = 0)) \\
&= 1 - \binom{12}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{12} - \binom{12}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^{11} \\
&= 1 - \left(\frac{5}{6}\right)^{12} - 12 \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^{11} \\
&\approx .6187.
\end{aligned}$$

4.2.2 Geometric

Now, consider an experiment with n independent trials, each having two outcomes with probabilities p and $1-p$. Let X be the random variable representing the number of trials required to obtain a success. To determine the probability of the first success being the kth trial, we use the Geometric probability distribution function given by:

$$P(X = i) = p(1 - p)^{i-1}, \tag{6}$$

where p is the probability of success for 1 trial, and n is the number of the final successful trial.

The expected value of a geometric random variable is $1/p$, which is relatively intuitive since you need $p \cdot 1/p$ to get 100% chance of success. The variance is similarly $\frac{1-p}{p^2}$

Suppose, as a homeowner, that you are ordering lightbulbs to replace a light fixture. In a box of 75 lightbulbs, on average, 3 are defective. What is the probability that the first defective light bulb is 6th?

First, we need to realize that the scenario given is a geometric distribution where X can represent the random variable of the number of trials to obtain a success. Since we are looking at 6 trials, $k = 6$ and the probability is as follows:

$$P(X = 6) = 0.04(1 - 0.04)^{6-1}$$
$$P(X = 6) = 0.04(0.96)^5 = 0.0326$$

.

There is approximately a 3.26% chance of the 6th bulb being defective.

4.2.3 Poisson

Suppose we want to approximate a binomial random variable with parameters (n, p). This approximation is represented as a Poisson random variable with parameter $\lambda = np$, when n is large and p is small enough so that np is of moderate size, given by

$$P(X = i) \approx e^{-\lambda} \cdot \frac{\lambda_i}{i!}.$$

In simple terms, a Poisson random variable is the probability of a given number of events occurring in a fixed interval of time or space, if these events occur with a known constant mean rate and independently of the time since the last event. The probability mass function of a Poisson random variable is given by

$$P(X = i) = e^{-\lambda} \cdot \frac{\lambda_i}{i!} \text{ for } i\text{=0,1,2,....} \tag{7}$$

Both the variance and expected value of a Poisson distribution are λ.

Let's say that the number of typos in any given mystery novel follows a Poison distribution of $\lambda = 0.5$. Suppose you purchase one such novel from a bookstore, what is the probability that there is at least one error in the book? To solve this, we can let X be the Poisson RV, which is the number of errors in the book. Then, the probability is given by

$$P(X \geq 1) = 1 - P(X = 1) = 1 - \frac{0.5^0 \cdot e^{-\frac{1}{2}}}{0!} = 1 - e^{-\frac{1}{2}} \approx 0.393.$$

4.3 Continuous Random Variables

A continuous random variable is a random variable that, unlike discrete random variables, has an uncountable set of possible values.

We define these random variables, denoted as X, with a non-negative function f defined for all $x \in (-\infty, \infty)$, as having the property that for any set of real numbers,

$$P\{X \in B\} = \int_B f(x)dx.$$

Here, f is called the **Probability Density Function (PDF)** of the continuous random variable X.

The above equation simply means that the probability that X will be in B can be obtained by integrating the PDF over the set B. Moreover, since X can take on any value, f must satisfy

$$P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)dx = 1.$$

By letting B= [a,b], we can find $P\{a \leq X \leq b\}$:

$$P\{a \leq X \leq b\} = \int_a^b f(x)dx.$$

If we let a=b, then we will see that $P\{X = a\} = 0$, implying that a continuous random variable can never take on a fixed value.

From this, we can say that, for a continuous random variable X,

$$P\{X < a\} = P\{X \leq a\} = F(a) = \int_{-\infty}^a f(x)dx,$$

where $F(a)$ is also called the cumulative distribution function (CDF).

4.3.1 Expected Value and Variance of Continuous Random Variables

We will start with the definition of the expected value of discrete random variables, which states that

$$E[x] = \sum_x xP\{X = x\}.$$

From this we can say that, since $f(x) \approx P\{x \leq X \leq x + dx\}$ for a small dx,

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

Similar to that of a discrete random variable, we can introduce the idea of the expected value of a function of a continuous random variable:

Proposition 9. *Let g(x) be any real-valued function. Then, when applied to some continuous random variable X, we have*

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

Now we are ready to find the variance of any continuous random variable. Starting from the definition of Variance in the discrete case, we have

$$Var[x] = E[X^2] - (E[X])^2.$$

Simply rewrite the expression above to obtain the formula for variance of a continuous random variable:

$$Var[x] = \int_{-\infty}^{\infty} x^2 f(x)dx - (\int_{-\infty}^{\infty} f(x)dx)^2.$$

4.3.2 Uniform Random Variables

A continuous random variable X is considered to be **uniformly distributed** over an interval (a, b) if the PDF of X is given by

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

We can use this definition to compute the expected value and variance of a uniform random variable, which yields

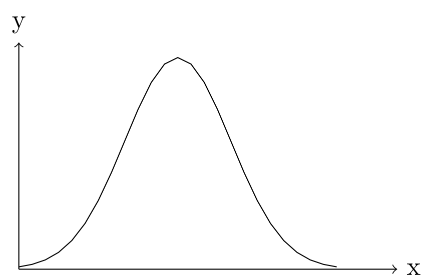
$$E[X] = \frac{b+a}{2},$$
$$Var[X] = \frac{(b-a)^2}{12}.$$

4.3.3 Normal Random Variables

A continuous random variable X is said to be normally distributed, with parameters μ and σ^2, if the density of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

where μ is the mean of the set and σ^2 is the standard deviation. This density function will yield a bell-shaped curve that is symmetrical about μ.



We can then compute the expected value and variance of a normal random variable, which yields

$$E[X] = \mu + \sigma E[Z] = \mu$$
$$Var[X] = \sigma^2 Var(Z) = \sigma^2$$

where $Z = (X - \mu)/\sigma$. This comes from the definition of a standard normal random variable.

4.3.4 Exponential Random Variables

A random variable X is said to be **exponential** if, for some $\lambda > 0$, its probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

From this definition we can compute the expected value and variance for an exponential random variable X, namely

$$E[X] = \frac{1}{\lambda}$$
$$Var(X) = \frac{1}{\lambda^2}.$$

5 Joint Distribution Functions

A **joint distribution function** is simply one that describes the probability of two or more random variables in one function. For any two random variables X and Y, the **joint probability distribution function**, which describes the related probabilities of X and Y, is defined as

$$F(a, b) = P(X \leq a, Y \leq b) \text{ for } \infty \leq a, b \leq \infty. \tag{8}$$

Their joint probability mass function is defined as

$$p(x, y) = P(X = x, Y = y). \tag{9}$$

When computing a joint distribution, the two variables are independent of each other when

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B). \tag{10}$$

In other words, X and Y are independent if knowing the value of one does not change the distribution of the other. If the variables are independent, the desired probabilities are multiplied, in accordance to the laws of conditional probability.

Consider a situation where radioactive particles of a decaying substance reach a Geiger counter (a device used for measuring radiation in the atmosphere) according to a Poisson process at a rate of $\lambda = 0.8$ particles per second. What is the probability that the Geiger counter detects (exactly) 1 particle in the next second and 3 or more in the next 4 seconds?

The joint probability of

$$P(1 \text{ particle in } (0,\, 1) \text{ AND } 3 \text{ or more particles in } (0,\, 4))$$

involves two events that are not independent. However, we can rewrite the joint probability as

$$P(1 \text{ particle in } (0,\ 1) \text{ AND } 2 \text{ or more particles in } (1,\ 4)).$$

Now, we can split this problem into two intervals of time, and calculate the joint probability for both by multiplication, since they are independent events:

$$P(1 \text{ particle in } (0,\ 1)) \cdot P(2 \text{ or more particles in } (1,\ 4)).$$

The probability of 1 particle arriving at the detector in the next second – the interval $(0, 1)$ can be represented as a Poisson random variable, with parameter $\mu = 0.8$. Additionally, the probability of 3 or more particles reaching the Geiger counter in the next 4 seconds – interval $(0, 4)$ – can be written as the probability of 2 or more reaching the counter in the interval $(1, 4)$. This can by similarly represented as a Poisson random variable, X, with parameter $\mu = 3 \cdot 0.8 = 2.4$. If we represent these two random variables by X and Y, respectively, then the probability is

$$X(1) \cdot (1 - Y(0) - Y(1)) = e^{-0.8}\frac{0.8^1}{1!} \cdot \left(1 - e^{-2.4}\frac{2.4^0}{0!} - e^{-2.4}\frac{2.4^1}{1!}\right) \approx .2486.$$

There is approximately a 24.86% chance that the Geiger counter detects 1 radioactive particle in the next second and 3 or more in the next 4 seconds.

5.1 Joint Probability: Discrete Case

Suppose X and Y are two discrete random variables and that X takes values from $x_1, x_2, ..., x_n$ and Y takes values $y_1, y_2, ..., y_m$. Then, the ordered pair (X, Y) take values in the product, $(x_1, y_1), (x_2, y_2), ...(x_n, y_m)$.

The joint probability mass function of X and Y is the function $p(x_i, y_j)$ that gives the probability of the joint outcome $X = x_i, Y = y_j$.

The total probability must equal 1 in accordance to the axioms of probability, represented by the double sum

$$0 \leq P_{XY}(x, y) \leq 1 \tag{11}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) = 1. \tag{12}$$

A joint discrete distribution is best represented in a tabular distribution matrix:

X/Y	y_1	y_2	....	y_j	....	y_m
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	....	$p(x_1, y_j$	....	$p(x_1, y_m$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	....	$p(x_2, y_j$	....	$p(x_2, y_m)$
....	....	....	....	....	....	....
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$	....	$p(x_i, y_j$	....	$p(x_i, y_m)$
....	....	....	....	....	....	....
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$	....	$p(x_n, y_j$	....	$p(x_n, y_m)$

Suppose you are given data that shows the outcomes of vehicle crashes when using different types of safety equipment. The random variable X corresponds to the level of injury and the random variable Y corresponds to the type of safety equipment used.

X/Y	None (0)	Belt Only (1)	Belt and Airbag (2)	Total
None (0)	0.065	0.075	0.06	0.20
Minor (1)	0.175	0.16	0.115	0.45
Major (2)	0.135	0.10	0.065	0.30
Death (3)	0.025	0.015	0.01	0.05
Total	0.40	0.35	0.25	1

If you sustained minor injuries, what is the probability that you had both a seat belt and an airbag?

The probability of sustaining minor injuries is 0.45, and the probability of sustaining minor injuries while using both belt and bag is 0.115. Therefore, the probability that you used both restraints, given that you sustained minor injuries, is $\frac{0.115}{0.45} = 0.256$.

5.2 Joint Probability: Continuous Case

Joint probability with continuous random variables uses continuous intervals of values, instead of discrete sets. As with all continuous random variables, we similarly use the joint probability density function instead of a probability mass function. Accordingly, we use the sums by integrals to compute.

If X takes values in $[a, b]$ and Y takes values in $[c, d]$ then the pair (X, Y) takes values in the product $[a, b] \cdot [c, d]$. The total probability must equal 1, expressed as a double integral:

$$\int_c^d \int_a^b f(x, y) dx dy = 1. \tag{13}$$

Consider a situation where a manager wants to find the relationship between the time spent on a building project and the client satisfaction with the project. She uses x as the number of weeks between the launch of the project and its completion, and Y as the percent satisfaction as reported by a client survey. She finds that the probability distribution follows the function

$$f_{XY}(x, y) = \frac{9}{10} xy^2 + \frac{1}{5} \text{ where } 0 \leq X \leq 2 \text{ and } 0 \leq Y \leq 1.$$

What is the probability that a project is completed in less than half a week and the client is more than 50% satisfied with the result?

Here, the manager would like to find the joint probability of $P(X < 1)$, and $P(Y > 0.5)$, given by

$$
\begin{aligned}
P(x<1, y>0.5) &= \int_0^1 \int_{0.5}^1 f_{XY}(x,y)\,dy\,dx \\
&= \int_0^1 \int_{0.5}^1 \left(\frac{9}{10}xy^2 + \frac{1}{5}\right) dy\,dx \\
&= \int_0^1 \left[\frac{3}{10}xy^3 + \frac{1}{5}y\right]_{0.5}^1 dx \\
&= \int_0^1 \left(\frac{21}{80}x + \frac{1}{10}\right) dx \\
&= \left[\frac{21}{160}x^2 + \frac{1}{10}x\right]_0^1 \\
&= 0.13125 + 0.1 \\
&= 0.23125.
\end{aligned}
$$

6 Markov Chains

Consider an event with different states it could be in. A Markov chain is a model describing a sequence of events when the probability of each event depends only on the probability of the previous event. In other words, what happens next depends only on the present circumstances. A sequence of random variables X_0, X_1, ..., can be in states $0, 1, ..., M$. X_n is the state at time n, and the system is in state i at time n if $X_n = i$.

The sequence of random variables X_0, X_1, ..., is said to form a Markov chain if each time the system is in state i, there is a fixed probability, P_{ij}, that the system will next be in state j. The values P_{ij}, $0 \leq i \leq M$, $0 \leq j \leq N$ are called **transition probabilities** of a Markov chain if they satisfy:

1. $P_{ij} \leq 0$
2. $\sum_{j=1}^{m} p(x_i, y_j)$
3. $i = 0, 1, ..., M$

The transition probabilities can be represented as

$$P\{X_{N+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_1 = i_1, X_0 = i_0\} = P_{ij}.$$

However, it is more convenient to arrange the transition probabilities P_ij in a matrix, creating a matrix called the **transition probability matrix**.

$$T = \begin{bmatrix} P_{00} & P_{01} \cdots & P_{0M} \\ P_{10} & P_{11} \cdots & P_{1M} \\ \vdots & & \\ P_{M0} & P_{M1} \cdots & P_{MM} \end{bmatrix}$$

6.1 Weather Problem

Let us take a look at how the transition matrix could be used to calculate probability. Suppose any day in town Lexington can be sunny, cloudy, rainy, or windy. The sequence of days is X_0, X_1, ... The states are sunny, cloudy, rainy, or windy. The date would be time n.

In Lexington, if today is sunny, tomorrow has 0.5 chance of being sunny, 0.15 chance of being cloudy, 0.1 chance of being rainy, and 0.25 chance of being windy. If today is cloudy, tomorrow has 0.2 chance of being sunny, 0.15 chance of being cloudy, 0.5 chance of being rainy, and 0.15 chance of being windy. If today is rainy, tomorrow has 0.25 chance of being sunny, 0.5 chance of being cloudy, 0 chance of being rainy, and 0.25 chance of being windy. If today is windy, tomorrow has 0.6 chance of being sunny, 0.2 chance of being cloudy, 0 chance of being rainy, and 0.2 chance of being windy.

With the above information, we can create the transition matrix for the weather in Lexington.

$$T = \begin{bmatrix} 0.5 & 0.15 & 0.1 & 0.25 \\ 0.2 & 0.15 & 0.5 & 0.15 \\ 0.25 & 0.5 & 0 & 0.25 \\ 0.6 & 0.2 & 0 & 0.2 \end{bmatrix}$$

The rows represent the weather of the current day, and the columns represent the weather of the next day. Each row adds up to 1 since there is a probability of 1 that the current day will transition into the next day.

The transition matrix can be used as a transformation matrix. The probability of weather in Lexington in n days is:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} T^n$$

So in our example, the probability of each weather in 2 days if the first day is sunny is:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0.5 & 0.15 & 0.1 & 0.25 \\ 0.2 & 0.15 & 0.5 & 0.15 \\ 0.25 & 0.5 & 0 & 0.25 \\ 0.6 & 0.2 & 0 & 0.2 \end{bmatrix} \cdot \begin{bmatrix} 0.5 & 0.15 & 0.1 & 0.25 \\ 0.2 & 0.15 & 0.5 & 0.15 \\ 0.25 & 0.5 & 0 & 0.25 \\ 0.6 & 0.2 & 0 & 0.2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.455 & 0.1915 & 0.125 & 0.2225 \\ 0.345 & 0.3325 & 0.095 & 0.2275 \\ 0.375 & 0.1625 & 0.275 & 0.1875 \\ 0.46 & 0.16 & 0.16 & 0.22 \end{bmatrix}$$

$$= \begin{bmatrix} 0.455 & 0.1975 & 0.125 & 0.2225 \end{bmatrix}$$

This end matrix tells us that there's 45.5% chance Lexington will be sunny again, 19.75% it will be cloudy, 12.5% it will be rainy, and 22.25% it will be windy in 2 days.

Let us calculate the probability of each weather in increased days:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} T^3 \approx \begin{bmatrix} 0.4318 & 0.2049 & 0.1443 & 0.2191 \\ 0.3993 & 0.1946 & 0.2008 & 0.2054 \\ 0.4013 & 0.2556 & 0.1189 & 0.2244 \\ 0.434 & 0.217 & 0.126 & 0.223 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} T^5 \approx \begin{bmatrix} 0.422 & 0.2119 & 0.1482 & 0.2179 \\ 0.4189 & 0.209 & 0.1565 & 0.2159 \\ 0.4185 & 0.2194 & 0.143 & 0.2191 \\ 0.4222 & 0.214 & 0.1452 & 0.2186 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} T^{15} \approx \begin{bmatrix} 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} T^{30} \approx \begin{bmatrix} 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} T^{100} \approx \begin{bmatrix} 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \\ 0.4208 & 0.2129 & 0.1485 & 0.2178 \end{bmatrix}$$

Notice how after the number of transformations increases, each row eventually becomes similar. This is because as the duration of the sequence increases, the weather of the first day becomes less significant.

Also, notice how the probabilities do not really change after a large number of transformations. Eventually, the system reaches a state where further transformations do not change the probabilities. This is called a **steady state**, and it gives the long-term probability of a system. The steady state can also be calculated with precise math.

Let S denote the steady state and T denote the transition matrix. Further transformations of the steady state do not change the steady state.

$$ST = S$$

Let us denote the probability of sunny weather in the steady state as a, of cloudy weather as b, of rainy weather as c, and of windy weather as $1-a-b-c$. We can rewrite the above equation as

$$\begin{bmatrix} a & b & c & (1-a-b-c) \end{bmatrix} \begin{bmatrix} 0.5 & 0.15 & 0.1 & 0.25 \\ 0.2 & 0.15 & 0.5 & 0.15 \\ 0.25 & 0.5 & 0 & 0.25 \\ 0.6 & 0.2 & 0 & 0.2 \end{bmatrix} = \begin{bmatrix} a & b & c & (1-a-b-c) \end{bmatrix}$$

Computing the multiplication of these matrices gives us the following system of equations.

$$\left\{ \begin{array}{c} 0.5a + 0.2b + 025c + 0.6(1 - a - b - c) = a \\ 0.15a + 0.15b + 0.5c + 0.2(1 - a - b - c) = b \\ 0.1a + 0.5b + 0c + 0(1 - a - b - c) = c \\ 0.25a + 0.15b + 0.25c + 0.2(1 - a - b - c) = 1 - a - b - c \end{array} \right\}$$

We can simplify this to

$$\left\{ \begin{array}{c} 1.1a + 0.4b + 0.35c = 0.6 \\ 0.05a + 1.05b - 0.3c = 0.2 \\ 0.1a + 0.5b - c = 0 \\ 1.05a + 0.95b + 1.05c = 0.8 \end{array} \right\}$$

We can then put this system of equations into matrix form.

$$\begin{bmatrix} 1.1 & 0.4 & 0.35 & 0.6 \\ 0.05 & 1.05 & -0.3 & 0.2 \\ 0.1 & 0.5 & -1 & 0 \\ 1.05 & 0.95 & 1.05 & 0.8 \end{bmatrix}$$

Using RREF, a method used for matrix calculations, we obtain

$$\begin{bmatrix} 1.1 & 0.4 & 0.35 & 0.6 \\ 0.05 & 1.05 & -0.3 & 0.2 \\ 0.1 & 0.5 & -1 & 0 \\ 1.05 & 0.95 & 1.05 & 0.8 \end{bmatrix} \approx \begin{bmatrix} 1 & 0 & 0 & 0.4208 \\ 0 & 1 & 0 & 0.2129 \\ 0 & 0 & 1 & 0.1485 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\therefore S = \begin{bmatrix} 0.4208 & 0.2129 & 0.1485 & 0.2178 \end{bmatrix}$$

This steady state tells us that on any given day, Lexington has 42.08% chance of being sunny, 21.29% of being cloudy, 14.85% of being rainy, and 21.78% of being windy.

References

[1] *Continuous Joint Probability* Introduction to Engineering Statistics. (n.d.). http://matcmath.org/textbooks/engineeringstats/continuous-joint-probability/

[2] *Discrete Joint Probability* Introduction to Engineering Statistics. (n.d.). http://matcmath.org/textbooks/engineeringstats/discrete-joint-probability/

[3] *Introduction to probability and statistics: Mathematics.* MIT OpenCourseWare. https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2014/

[4] Ross, Sheldon. A First Course in Probability. 10th ed., Pearson, 2020.

[5] Sigman, K. (2009). *Gambler's Ruin Problem.* Columbia University. http://www.columbia.edu/ ks20/stochastic-I/stochastic-I-GRP.pdf