Probability theory, the mathematical science of uncertainty, plays an ever growing role in how we understand the world around us - whether it is the climate of the planet, the spread of an infectious disease, or the results of the latest new poll.

The word "stochastic" comes from the Greek stokhazesthai, which means to aim at, or guess at. A stochastic process, also called a random process, is simply one in which outcomes are uncertain. By contrast, in a deterministic system there is no randomness. In a deterministic system, the same output is always produced from a given output.

Functions and differential equations are typically used to describe deterministic processes. Random variables and probability distributions are the building blocks for stochastic system.

Stochastic processes play a very important role in financial engineering. They are used to model random movements of the values of financial instruments, such as stocks, bonds, etc.

1 References

1. An introduction to mathematical finance with applications: understanding and buiding financial institution, Arlie O. Peter and Xiaoying Dong.
2. Simulation and inference for stochastic differential equations- With R examples- Stefano M. Iacus
3. Introduction to stochastic processes with R, Robert P. Dobrow
4. Probability with applications and R, Robert P. Dobrow
5. Potential references

   `http://b-ok.org/g/stochastics+in+finance`

2 Mathematical language for things that are uncertain in the future

Before rolling a die you do not know the result. This is an example of a random experiment. In particular, **a random experiment is an action that can produce uncertain outcomes under the same condition**. The result of the random experiment is known only when it ends. In our lives, there are many such random experiments and we would like to know which outcomes could occur.

People are always curious about future. That is why many people comes to fortune tellers to ask about their future: marriage, wealth, etc. What happens in our future is unknown at present and we can consider these as random events. Investors want to have a reliable prediction about future stock prices so that they can invest in right stocks and earn good profit. Insurance want to know when a person might die so that they can create a reasonable insurance policy to that person. Although we do not know exactly what will happen in future, we could collect past outcomes for these events, and then use mathematics to predict which outcomes may occur, with a certain likelihood. By doing that, we will have a good preparation for our future.

To do so, we first translate our real problems into mathematical language. A result of a random experiment is called an outcome. The set of all possible outcomes (all what can happen) is termed the sample space. Mathematically, we will use the Greek capital letter Ω (omega) to represent the sample space. The Greek lowercase

omega ω will be used to denote these outcomes, the elements of Ω.

An event is a set of outcomes. The event of getting all heads in three coin tosses can be written as: $A = \{HHH\}$. The event of getting at least two tails is $B = \{HTT, THT, TTH, TTT\}$.

Example 2.1. The weather forecast for tomorrow says rain. The amount of rainfall can be considered a random experiment. If at most 24 inches of rain will fall, then the sample space is the interval $\Omega = [0, 24]$. The event that "the amount of rainfall is between 2 and 4 inches of rain" is $A = [2, 4]$.

Example 2.2. Roll a pair of dice. Find the sample space and identify the event that the sum of the two dice is equal to 7.

Answer 2.1. The random experiment is rolling two dice. Keeping track of the roll of each die gives the sample space $\Omega = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),...,(6,5),(6,6)\}$. The event is $A = \{\text{Sum is } 7\} = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$.

Example 2.3. Yolanda and Zach are running for president of the student association. One thousand students will be voting. Identify:

1. the sample space
2. the event that Yolanda beats Zach by at least 100 votes.

Answer 2.2.
1. The outcome of the vote can be denoted as $(x, 1000 - x)$, where x is the number of votes for Yolanda, and $1000 - x$ is the number of votes for Zach. Then the sample space of all voting outcomes is $\Omega = \{(0,1000),(1,999),(2,998),...,(999,1),(1000,0)\}$.
2. Let A be the event that Yolanda beats Zach by at least 100 votes. The event A consists of all outcomes in which $x-(1000-x) \geq 100$, or $550 \leq x \leq 1000$. That is, $A = \{(550,450),(551,449),...,(999,1),(1000,0)\}$.

Example 2.4. Joe will continue to flip a coin until heads appears. Identify the sample space and the event that it will take Joe at least three coin flips to get a head.

Answer 2.3. The sample space is the set of all sequences of coin flips with one head preceded by some number of tails. That is, $\Omega = \{H, TH, TTH, TTTH, TTTTH, TTTTTH, ...\}$. The desired event is $A = \{TTH, TTTH, TTTTH, ...\}$. Note that in this case both the sample space and the event A are infinite.

3 The likelihood of occurrence of an event

With above terms and notations, we could already record random experiment in mathematical language. But what we would eventually like to know is the likelihood of occurrence of an event, which is termed as the probability of that event.

What does it mean to say that the probability that A occurs is equal to x?

From a practical, empirical point of view, a probability matches up with our intuition of the likelihood or "chance" that an event occurs. An event that has probability 0 "never" happens. An event that has probability 1 is "certain" to happen. In repeated coin flips, a coin comes up heads about half the time, and the probability of heads is equal to one-half.

Let A be an event associated with some random experiment. One way to understand the probability of A is to perform the following **thought exercise**: imagine conducting the experiment over and over, infinitely

often, keeping track of how often A occurs. Each experiment is called a trial. If the event A occurs when the experiment is performed, that is a success. The proportion of successes is the probability of A, written $P(A)$. This is the relative frequency interpretation of probability, which says that **the probability of an event is equal to its relative frequency in a large number of trials.**

When the weather forecaster tells us that tomorrow there is a 20% chance of rain, we understand that to mean that if we could repeat today's conditions - the air pressure, temperature, wind speed, etc. - over and over again, then 20% of the resulting "tomorrows" will result in rain. **Closer to what weather forecasters actually do in coming up with that** 20% **number, together with using satellite and radar information along with sophisticated computational models, is to go back in the historical record and find other days that match up closely with today's conditions and see what proportion of those days resulted in rain on the following day.**

There are definite limitations to constructing a rigorous mathematical theory out of this intuitive and empirical view of probability. One cannot actually repeat an experiment infinitely many times. To define probability carefully, we need to take a formal, axiomatic, mathematical approach. Nevertheless, the relative frequency viewpoint will still be useful in order to gain intuitive understanding.

We now construct a rigorous mathematical theory, precisely a probability measure, for computing the likelihood of occurrence of an event. Ideally, one would like to build a probability measure $\mathcal{P}$, which assigns a probability for every subset of Ω. It is, however, often the case that one can build a probability measure $\mathcal{P}$ only for a collection $\mathcal{F}$ of subsets of Ω (not all the subsets of Ω). So what kind of subset of set $\mathcal{F}$ is suitable?

To make sense, the probability of all outcomes should sum up to 1, that is $P(\Omega) = 1$. Thus $\Omega \in \mathcal{F}$. In addition, if we know the probability of an event A, we should also know the probability of the opposite event, i.e., the complement of A, denoted by A^c. The two probabilities is related by $P(A) = 1 - P(A^c)$. Thus if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$. Moreover, a probability measure $\mathcal{P}$ should be additive, that is the probability of two disjoint events (two disjoint subsets of Ω) is equal to the sum of the probability of each event. Mathematically, this can be expressed as $\mathcal{P}(A \cup B) = P(A) + P(B)$, if $A \cap B = \emptyset$. We can easily extend this to a finite union of disjoint events: $P(A_1 \cup A_2 \cup \cdots A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$. Even more, for a countable pairwise disjoint collection of events $A_i, i = 1, 2, ...$, we may require: $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$. Therefore, if we know the probability of each events (which of course belong to $\mathcal{F}$), we may be able to calculate the probability of their union event (which thus also belong to $\mathcal{F}$). As a result, $\mathcal{F}$ must be closed under finite union (or should be closed under countable union).

A set $\mathcal{F}$ satisfying all above requirements is called may a $\sigma - algebra$. **This is the set of all events that we would like to know the likelihood of occurrence.**

4 $\sigma - algebra$ of a sample space

Let Ω be a sample space of a random experiment. A collection of subsets of Ω is called an $\sigma - algebra$ over Ω if it satisfies the following conditions:

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_i \in \mathcal{F},\ \forall i = 1, 2 \cdots \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$

Let $\mathcal{F}_1$ and $\mathcal{F}_2$ are two $\sigma-algebra$ on a non-empty sample space Ω. If $\mathcal{F}_1 \subseteq \mathcal{F}_2$ then $\mathcal{F}_1$ is called a sub-$\sigma-algebra$ of $\mathcal{F}_1$.

Let G be an event of the sample space Ω. The smallest $\sigma - algebra$ containing G, which is the intersection of all $\sigma - algebras$ including G, is called the $\sigma - algebra$ generated by G and denoted by $\sigma(G)$. In particular, $\sigma(G) = \{\emptyset, \Omega, G, G^c\}$, where G^c is the complement of G.

Question? What are the smallest and largest $\sigma - algebras$ of over sample space Ω.

Example 4.1. (Borel $\sigma - algebra$) Consider a random experiment of picking a random real number $\mathbb{R}$. The sample space, containing all possible outcomes, would be $\mathbb{R}$. We now build a σ-algebra, which is called Borel σ-algebra, on this sample space. **We start with closed intervals $[a, b] \subset \mathbb{R}$ and add all other sets necessary to have a σ-algebra. Unions of sequences of closed intervals are Borel sets. In particular, every open interval is a Borel set because an open interval can be written as the union of a sequence of closed intervals. Furthermore, every open set (whether or not an interval) is a Borel set because every open set is the union of a sequence of open intervals. Every closed set is a Borel set because it is the complement of an open set. We denote the collection of Borel subsets of $\mathbb{R}$ by $\mathcal{B}(\mathbb{R})$ and call it the Borel σ-algebra of $\mathbb{R}$.**

Important note 4.1. Borel $\sigma - algebra$ is one of the most important one in the theory of stochastic processes.

We now build a probability measure that assigns each event in the $\sigma - algebra$ to its likelihood of occurrence.

5 Probability spaces

Definition 5.1. (Probability measure) Let Ω be a sample space and $\mathcal{F}$ a $\sigma - algebra$ over Ω. A function $\mathcal{P}$ from $\mathcal{F}$ to $[0, 1]$ is said to be a probability measure on $\mathcal{F}$ if it satisfies the following conditions:

- $\mathcal{P}(\Omega) = 1$;
- $\mathcal{P}(A) = 1 - \mathcal{P}(A^c)$;
- For each countable collection $\{A_i \in \mathcal{F}, i \in I\}$ of pairwise disjoint sets, $\mathcal{P}(\cup_{i\in I} A_i) = \sum_{i\in I} P(A_i)$.

You may not be familiar with some of the notations in this definition. The symbol $\in$ means "is an element of." So $\omega \in \Omega$ means ω is an element of Ω. We are also using a generalized σ-notation to specify the summation. The notation $\sum_{i\in I} P(A_i)$ means that the sum is over all i that are elements of the set I. In the case of a finite set $I = \{1, 2, ..., n\}$, the sum becomes $P(A_1) + P(A_2) + ... + P(A_n)$.

Important note 5.1. There are many ways to build a probability measure for a random experiment. The chosen probability measure usually has some intuition. The most usual way is to define the probability of each outcome as its relative frequency of a large number of trials. The probability of an event is then the sum of the probabilities of all outcomes, which are elements of the event.

Definition 5.2. (Probability space) A probability space is denoted by the triple $(\Omega, \mathcal{F}, \mathcal{P})$, where:

- Ω denotes the sample space, that is a nonempty set of all possible outcomes of a random experiment. Each subset of Ω, a set of some outcomes, is called an event.
- $\mathcal{F}$ denotes a $\sigma - algebra$ on Ω, which is a collection of subsets of Ω (satisfying some requirements).

- $\mathcal{P}$ denotes a probability measure on $\mathcal{F}$ that assigns probabilities to events. Here $\mathcal{P}$ is a function from $\mathcal{F}$ to $[0, 1]$.

It can be said that **σ-algebra $\mathcal{F}$ forms a collection of events for which a probability can be assigned**. More specifically, $\mathcal{F}$ is measurable with probability measure $\mathcal{P}$. Any subset A of $\mathcal{F}$ is then said $\mathcal{F}$-measurable, i.e., we can determine the probability of A.

Example 5.1. (Tossing a biased coin). Consider a random experiment of tossing a biased coin. Assume that the chance of landing heads up is 30% and that of landing tails up is 70%. Then the corresponding probability space is represented by $(\Omega, \mathcal{F}, \mathcal{P})$, where

- Sample space Ω contains all possible outcomes of tossing the coin, which are landing heads up, denoted by H, and landing tails up, denoted by T. Thus, $\Omega = \{H, T\}$.
- $\sigma - algebra$ $\mathcal{F}$, the only possible one in this case, is defined as $\mathcal{F} = \{\emptyset, \Omega, H, T\}$.
- Probability measure $\mathcal{P}$ is defined as $\mathcal{P} : \mathcal{F} \to [0, 1]$ such that $\mathcal{P}(H) = 0.3$ and $\mathcal{P}(T) = 0.7$. Obviously, the probability of the other two events are $\mathcal{P}(\emptyset) = 0, \mathcal{P}(\Omega) = 1$.

Example 5.2. A college has six majors: biology, geology, physics, dance, art, and music. The proportion of students taking these majors are $20, 20, 5, 10, 10$, and 35, respectively (100 students in total). Choose a random student. Build a reasonable probability space to compute the probability that the chosen student has a science major?

Answer 5.1. The random experiment is choosing a student. The sample space is Ω = {Bio, Geo, Phy, Dan, Art, Mus}. The probability function is given in the below table:

Major	Bio	Geo	Phy	Dan	Art	Mus
P	0.2	0.2	0.05	0.1	0.1	0.35

The event in question is A = {Science major} = {Bio, Geo, Phy}. Finally, $P(A) = P(\{\text{Bio, Geo, Phy}\}) = P(Bio) + P(Geo) + P(Phy) = 0.20 + 0.20 + 0.05 = 0.45$.

Example 5.3. Tossing three coins simultaneously, and observe whether heads or tails of each coin lands up. Build a reasonable probability space to compute the probability of getting at least two tails?

Answer 5.2. Although the probability model here is not explicitly stated, the simplest and most intuitive model for fair coin tosses is that every outcome is equally likely. Since the sample space

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

has eight outcomes, the model assigns to each outcome the probability 1/8. The event of getting at least two tails can be written as $A = \{HTT, THT, TTH, TTT\}$. This gives $P(A) = 4/8 = 1/2$.

Since counting plays a fundamental role in probability when outcomes are equally likely, we introduce some basic counting principles.

5.1 MULTIPLICATION PRINCIPLE

Theorem 1. If there are m ways for one thing to happen, and n ways for a second thing to happen, there are $m * n$ ways for both things to happen.

More generally - and more formally - consider an n-element sequence $(a_1, a_2, ..., a_n)$. If there are k_1 possible values for the first element, k_2 possible values for the second element, . . ., and k_n possible values for the nth element, there are $k_1 \times k_2 \times ... \times k_n$ possible sequences.

Example 5.4. License plates in Minnesota are issued with three letters from A to Z followed by three digits from 0 to 9. If each license plate is equally likely, what is the probability that a random license plate starts with G-Z-N?

Answer 5.3. The solution will be equal to the number of license plates that start with G-Z-N divided by the total number of license plates. By the multiplication principle, there are $26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17,576,000$ possible license plates. For the number of plates that start with G-Z-N, think of a 6-element plate of the form G-Z-N- - -. For the three blanks, there are $10 \times 10 \times 10$ possibilities. Thus

Definition 5.3. (SAMPLING WITH AND WITHOUT REPLACEMENT.) When sampling with replacement, a unit that is selected from a population is returned to the population before another unit is selected. When sampling without replacement, the unit is not returned to the population after being selected.

Important note 5.2. When solving a probability problem involving sampling (such as selecting cards or picking balls from urns), make sure you know the sampling method before computing the related probability.

5.2 PROBLEM-SOLVING STRATEGIES: COMPLEMENTS

Consider a sequence of events $A_1, A_2, ...$ In this section, we consider strategies to find the probability that at least one of the events occurs, which is the probability of the union $\cup_i A_i$.

Sometimes the complement of an event can be easier to work with than the event itself. The complement of the event that at least one of the A_is occurs is the event that none of the A_is occur, which is the intersection $\cap A_i$. Complements turn unions into intersections, and vice versa (DeMorgan's laws). In particular, $(\cup_{i=1}^{\infty} A_i)^c = \cap_{i=1}^{\infty} A_i^c$ and $(\cap_{i=1}^{\infty} A_i)^c = \cup_{i=1}^{\infty} A_i^c$.

Example 5.5. Four dice are rolled. Find the probabilty of getting at least one 6.

Answer 5.4. The sample space is the set of all outcomes of four dice rolls $\Omega = \{(1,1,1,1), (1,1,1,2), ..., (6,6,6,6)\}$. By the multiplication principle, there are $6^4 = 1296$ elements. If the dice are fair, each of these outcomes is equally likely.

Let A be the event of getting at least one 6. Then the complement A^c is the event of getting no sixes in four rolls. An outcome has no sixes if the dice rolls a $1, 2, 3, 4$, or 5 on every roll. By the multiplication principle, there are $5^4 = 625$ possibilities. Thus $P(A^c) = 5^4/6^4 = 625/1296$ and $P(A) = 1 - P(A^c) = 1 - 625/1296 = 0.5177$.

Example 5.6. Mark is taking four final exams next week. His studying was erratic and all scores A, B, C, D, and F are equally likely for each exam. What is the probability that Mark will get at least one A?

Answer 5.5. Take complements. The complementary event of getting at least one A is getting no As. Since outcomes are equally likely, by the multiplication principle there are 4^4 exam outcomes with no As (four grade choices for each of four exams). And there are 5^4 possible outcomes in all. The desired probability is $1 - 4^4/5^4 = 0.5904$.

5.2.1 Birthday Problem

The birthday problem is a classic probability delight first introduced by the mathematician Richard Von Mises in 1939. Von Mises asked, "How many people must be in a room before the probability that some share a birthday, ignoring the year and ignoring leap days, becomes at least 50%?" For a group of k people, let B be the event that at least two people have the same birthday. We find $P(B)$. Remember the problem-solving strategy of taking complements for "at least" probabilities. The complement B^c is the event that none of the k people

have the same birthday.

Consider asking people one by one their birthday and checking whether their birthday is different from the birthdays of those previously asked. The first person's birthday is fixed. The second person's birthday either matches the first birthday, which occurs with probability 1/365, or does not, with probability 364/365. The probability that the third person has a birthday different from the previous two, given that the previous two birthdays are different, is 363/365 (since two birthdays have been picked and there are 363 available ones left). Continuing in this way, we see that the ith branch of the tree gives the probability that the (i + 1)st person's birthday is different from the previous i birthdays, given that the previous i birthdays are all different, which occurs with probability $(365 - i)/365$. This gives $P(B^c) = \Pi_{i=1}^{k-1} \frac{365 - i}{365}$. And thus the birthday probability that at least two people have the same birthday is $P(B) = 1 - P(B^c) = 1 - \Pi_{i=1}^{k-1} \frac{365 - i}{365}$.

At $k = 22, P(B) = 0.476$, and at $k = 23, P(B) = 0.507$. So the answer to Von Mises' question is, remarkably, 23 people. The number is much smaller than most people think. With just $k = 15$ people there is a 25% chance of a birthday match. And with $k = 50$ people the likelihood of a match is virtually certain with $P(B) = 0.970$.

To explain the seemingly paradoxical result, intuitively observe that in a group of 23 people there are actually 253 ways for people to be paired. And we just need one of those pairs to have a common birthday for the desired event to occur.

6 Random variables

Often the outcomes of a random experiment take on numerical values. For instance, we might be interested in how many heads occur in three coin tosses. More specifically, let us toss three coins simultaneously and observe the number of heads occurs. Possible outcomes of such random experiment are: zero heads, one heads, two heads, three heads. Mathematically, let ω be the outcome of the experiment. Then ω receives one of the following values: $0, 1, 2$, or 3, depending on the outcome of the coin tosses.

For those random experiments with outcomes are not numbers, we can always assign a number to each of the outcome. That assignment is called a random variable.

Definition 6.1. A random variable on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ is a function X from Ω to $\mathbb{R}$. That is,

$$\begin{aligned} X: \quad & \Omega \to R \\ & \omega \to X(\omega) \end{aligned}$$

and for each $a \in \mathbb{R}$, the set $\{X \leq a\} = \{\omega \in \Omega | X(\omega) \leq a\} \in \mathcal{F}$ or $\{X \leq a\}$ is $\mathcal{F}$-measurable. In other expression, we can determine the probability of the event $X \leq a$ and thus we can determine the cumulative distribution of X.

The set $\mathbf{S} = \{X(\omega) | \omega \in \Omega\}$ is called the state of the random variable X (i.e., the range of the function X). If $\mathbf{S}$ is a countable set (containing finite or countably infinite elements) then X is called a discrete random variable. If $\mathbf{S}$ is an uncountable set (containing infinitely uncountable elements) then X is called a continuous random variable.

Important note 6.1. A random variable transforms each outcome of a random experiment into a real number. Also, the condition $\{X \leq a\} \in \mathcal{F}$ ensures that the probability $P(X \leq a)$ (and the cumulative distribution of X) is well-defined on the probability space. Intuitively, the information contains in $\mathcal{F}$ is enough to determine X. A significant attribute of random variables is that they allow us to work on cumulative distributions without

recognizing original sample spaces. It may be understood now that everything of the random experiment can be represented by numbers and we now use probability theory to figure out the likelihood of random events.

Example 6.1. (One-period Binomial model). We now consider an oversimplified stock price behavior in which the stock price either goes up by a factor $u > 1$ if a good economy condition, denoted by G, occurs with probability p or goes down by a factor $d < 1$ if a bad economy condition, denoted by B, occurs with probability $1 - p$ over a period of time $[t_0, t_1]$.
Then the corresponding probability space is represented by $(\Omega), \mathcal{F}, \mathcal{P})$, where

- Sample space Ω contains all possible outcomes of economy conditions, which are good condition, denoted by G, and bad condition, denoted by B. Thus, $\Omega = \{G, B\}$.
- $\sigma - algebra$ $\mathcal{F}$, the only possible one in this case, is defined as $\mathcal{F} = \{\emptyset, \Omega, G, B\}$.
- Probability measure $\mathcal{P}$ is defined as $\mathcal{P} : \mathcal{F} \to [0, 1]$ such that $\mathcal{P}(G) = p$ and $\mathcal{P}(B) = 1 - p$. Obviously, the probability of the other two events are $\mathcal{P}(\emptyset) = 0, \mathcal{P}(\Omega) = 1$.

Denote S_0 the stock price at time t_0 and S_1 the stock price at time t_1. We define a real-valued function S_1 as

$$\begin{aligned} S_1 : \quad & \Omega \to R \\ & G \to S_0 u \\ & B \to S_0 d \end{aligned}$$

To prove S is a random variable defined on the above probability space, we need to prove $\{X \leq a\} \in \mathcal{F},\ \forall a \in \mathbb{R}$. In fact,

$$\{X \leq a\} = \{\omega \in \Omega | X(\omega) \leq a\} = \begin{cases} \emptyset & \text{if } a < S_0 d, \\ B & \text{if } S_0 d \leq a < S_0 u, \\ \Omega & \text{if } a \geq S_0 u. \end{cases}$$

As can be seen clearly that $\{X \leq a\} \in \mathcal{F},\ \forall a \in \mathbb{R}$. Thus S is a random variable defined on the above probability space.

The smallest σ-algebra contains all subset $\{X \leq a\}, a \in \mathbb{R}$ is called the σ-algebra generated by X, denoted by σ. It is clear that X is $\sigma(X)$-measurable. In the language of information, it can be said that $\sigma(X)$ is the least information required to determine X (in the sense that we can determine the distribution of X).

Intuition. A random variable X is a function $X : \Omega \to \mathbb{R}$ that connects each random outcome (an element of the sample space Ω) to a unique real number. Two different outcomes may be represented by a number, but an outcome cannot be represented by two real numbers.

Random variables are enormously useful and allow us to use algebraic expressions, equalities, and inequalities when manipulating events. In many examples, we work with random variables without using the name, for example, the number of threes in rolls of a die, the number of votes received, the number of palindromes, the number of heads in repeated coin tosses.

Example 6.2. In tossing three coins, let X be the number of heads. Then the event of getting two heads can be written as $\{X = 2\}$. The probability of getting two heads is thus $P(X = 2) = P(\{HHT, HTH, THH\}) = 3/8$.

Example 6.3. If we throw two dice, what is the probability that the sum of the dice is greater than four?

Answer 6.1. We can, of course, find the probability by direct counting. But we will use random variables. Let Y be the sum of two dice rolls. Then Y is a random variable whose possible values are $2, 3, ..., 12$. The event

that the sum is greater than 4 can be written as $\{Y > 4\}$. Observe that the complementary event is $\{Y \leq 3\}$. By taking complements, $P(Y > 4) = 1 - P(Y \leq 3) = P(\{Y = 2\} \cup \{Y = 3\}) = P(Y = 2) + P(Y = 3) = P(\{(1, 1)\}) + P(\{(1, 2), (2, 1)\}) = 1/36 + 2/36 = 1/12$.

Example 6.4. Yolanda and Zach are running for president of the student association. One thousand students will be voting. Suppose the number of votes that Yolanda receives is equally likely to be any number from 0 to 1000. What is the probability that Yolanda beats Zach by at least 100 votes?

Answer 6.2. We approach the problem using random variables. Let Y be the number of votes for Yolanda. Let Z be the number of votes for Zach. Then the total number of votes is $Y + Z = 1000$. Thus, $Z = 1000 - Y$. The event that Yolanda beats Zach by at least 100 votes is $\{Y - Z \geq 100\} = \{Y - (1000 - Y) \geq 100\} = \{2Y \geq 1100\} = \{Y \geq 550\}$. The desired probability is $P(Y - Z \geq 100) = P(Y \geq 550) = 451/1001$, since there are 1001 possible votes for Yolanda and 451 of them are greater than 550.

Important note 6.2. What is the difference between a random variable and a "deterministic function"? The main difference is that random variables operate on random outcomes (i.e. random inputs), while deterministic functions operate on deterministic domain (i.e., deministic inputs). For deterministic functions, you can choose specifically the input you wanted to evaluate, however, the random inputs cannot be chosen as they randomly occur.

6.1 Distribution measure of a random variable

Definition 6.2. (Distribution measure of a random variable.) Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. The distribution measure of X is the probability measure μ_X on the Borel σ-algebra of $\mathbb{R}$ such that

$$\begin{aligned} \mu_X : \quad & \mathcal{B}(\mathbb{R}) \to R \\ & B \to \mu_X(B) = P(X \in B) = P(X^{-1}(B)), \end{aligned}$$

where where $X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in B\}$.

The distribution measure of a continuous random variable X can be described via its density function $f(x)$, if exists, which satisfies:

$$\mu_X[a, b] = P\{a \leq X \leq b\} = \int_a^b f(x)dx, -\infty < a \leq b < \infty. \tag{1}$$

We also denote $\mu_X(x) = \mu_X[-\infty, x] = \int_{-\infty}^{x} f(t)dt$. Thus $\mu_X[a, b] = \mu_X(b) - \mu_X(a)$.

The distribution measure of a discrete random variable X can be described via its probability mass function if there exits a countable sequence of numbers $x_1, x_2, ...$ such that with probability 1, the random variable X takes one of the values in the sequence. We then denote $p_i = P(X = X_i)$. Each p_i is nonnegative and $\sum_{i=1}^{\infty} p_i = 1$. The distribution measure of X can be represented by $\mu_X(B) = \sum_{i:x_i \in B} p_i$.

6.2 Unconditional expectation

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. If Ω is a countable set, i.e., its elements can be listed in a sequence $\omega_1, \omega_2, ...$, we can define the unconditional expectation of X as $E(X) = \sum_{k=1}^{\infty} X(\omega_k)P(\omega_k)$. Difficulty arises, however, if Ω is uncountably infinite. Uncountable sum cannot be defined. The Lebesgue

integral is used to define the unconditional expectation in general.

The expectation of X is defined $E(X) = \int_\Omega X(\omega)dP(\omega)$. We now explain the integral in details. Suppose X can receive values in $S \subset \mathbb{R}$. We use a partition Π to divide S into subintervals $[y_i, y_{i+1}]$ and then define

$$\int_\Omega X(\omega)dP(\omega) = \lim_{\Pi\to 0}\sum_{k=1}^{\infty} y_k P(\omega \in \Omega | y_k \le X(\omega) \le y_{k+1}) = \lim_{\Pi\to 0}\sum_{k=1}^{\infty} y_k P(y_k \le X \le y_{k+1}) \quad (2)$$

$$= \lim_{\Pi\to 0}\sum_{k=1}^{\infty} y_k(\mu_X(y_{k+1}) - \mu_X(y_k)) = \lim_{\Pi\to 0}\sum_{k=1}^{\infty} y_k d\mu_X(y_k) = \int_{-\infty}^{\infty} x d\mu_X(x) \quad (3)$$

If X has a density function $f(x)$ then $d\mu_X(x) = f(x)dx$, thus $E(X) = \int_{-\infty}^{\infty} xf(x)dx$. The above formula shows the relationship between an integral over the sample space Ω and one over $\mathbb{R}$. This also give us a more convenient way to calculate the unconditional expectation (see (**??**) for more details on the definition of the probability measure μ_X).
As can be seen clearly the unconditional expectation of a random variable is a constant, which is the best prediction about possible outcomes that the random variable may receive (here no available information about the random variable).

6.2.1 Expectation of a function of a random variable

Suppose X is a random variable and f is some function. Then $Y = f(X)$ is a random variable that is a function of X. The values of this new random variable are found as follows. If $X = x$, then $Y = f(x)$.

Functions of random variables, like X^2, e^X, and $1/X$, show up all the time in probability. Often we apply some function to the outcomes of a random experiment, as we will see in many examples later. In statistics, it is common to transform data using an elementary function such as the log or exponential function.

Suppose X is uniformly distributed on $\{-2, -1, 0, 1, 2\}$. That is, $P(X = k) = 1/5$, for $k = -2, -1, 0, 1, 2$. Take $f(x) = x^2$, and let $Y = f(X) = X^2$. The possible outcomes of Y are $(-2)^2, (-1)^2, 0^2, 1^2, 2^2$, that is, $0, 1$, and 4. Since $Y = 0$ if and only if $X = 0$, $Y = 1$ if and only if $X = \pm 1$, and $Y = 4$ if and only if $X = \pm 2$, the probability mass function of Y is

$$P(Y = 0) = P(X^2 = 0) = P(X = 0) = 1/5.$$
$$P(Y = 1) = P(X^2 = 1) = P(X = \pm 1) = P(X = 1) + P(X = -1) = 2/5.$$
$$P(Y = 4) = P(X^2 = 4) = P(X = \pm 2) = P(X = 2) + P(X = -2) = 2/5.$$

Example 6.5. Timmy spends \$2 in supplies to set up his lemonade stand. He charges 25 cents a cup. Suppose the number of cups he sells in a day has a Poisson distribution with $\lambda = 10$. Describe his profit as a function of a random variable and find the probability that the lemonade stand makes a positive profit.

Answer 6.3. Let X be the number of cups Timmy sells in a day. Then $X \sim Pois(10)$. If he sells x cups then his profit is $25x - 200$ cents. The random variable $Y = 25X - 200$ defines his profit as a function of X, the number of cups sold. Since X takes values $0, 1, 2, ...$, Y takes values $-200, -175, -150, ...$ The probability that Timmy makes a positive profit is

$$P(Y > 0) = P(25X - 200 > 0) = P(X > 8) = 1 - P(X \le 8) = 1 - \sum_{k=0}^{8} e^{-10}\frac{10^k}{k!} = 0.667.$$

Theorem 2. Let X be a random variable that takes values in a set $\mathbf{S}$ and g be a function. Then,

$$E[g(X)] = \begin{cases} \sum_{x \in \mathbf{S}} f(x)P(X = x), \text{ if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} f(x)g(x)dx \text{ if } X \text{ is continuous with density } f. \end{cases}$$

Example 6.6. A number X is picked uniformly at random from 1 to 100. What is the expected value of X^2?

Answer 6.4. The random variable X^2 is equal to $f(X)$, where $f(x) = x^2$. This gives:

$$E[X^2] = \sum_{x=1}^{100} x^2 P(X = x) = \frac{1}{100} \sum_{x=1}^{100} x^2 = \frac{100 * 101 * 201}{6 * 100} = 3383.5,$$

using the fact that the sum of the first n squares is $n(n + 1)(2n + 1)/6$.

It is not true that $E[X2] = (E[X])^2$. More generally, it is not true that $E[f(X)] = f(E[X])$. The operations of expectation and function evaluation cannot be interchanged. It is very easy, and common, to make this kind of a mistake. To be forewarned is to be forearmed!

Example 6.7. Create a "random sphere" whose radius R is determined by the roll of a die. Let V be the volume of the sphere. Find $E[V]$.

Answer 6.5. The radius R is a uniform random variable on $\{1, 2, 3, 4, 5, 6\}$. If the radius $R = r$ then the volume of the sphere is $V = \frac{4}{3}\pi r^3$. The random variable $V = \frac{4}{3}\pi R^3$ defines the volume of the sphere.

$$E[V] = E[\frac{4}{3}\pi R^3] = \sum_{r=1}^{6} \frac{4}{3}\pi r^3 P(R = r) = \sum_{r=1}^{6} \frac{4}{6 * 3}\pi r^3 = \frac{2\pi}{9}(1^3 + 2^3 + 3^3 + 4^3 + 5^3 + 6^3) = 98\pi.$$

Example 6.8. Suppose X has a Poisson distribution with parameter λ. Find $E\Big[1/(X + 1)\Big]$.

Answer 6.6. We have $X \sim Pois(\lambda)$.

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

6.2.2 EXPECTATION OF FUNCTION OF TWO RANDOM VARIABLES

Random variables can arise as functions of two or more random variables. Suppose $f(x, y)$ is a real-valued function of two variables. Then $Z = f(X, Y)$ is a random variable, which is a function of two random variables. We can define expectations of such random variables as follows:

Definition 6.3. $E[f(X, Y)] = \sum_{x \in \mathbf{S}} \sum_{y \in \mathbf{T}} f(x, y)P(X = x, Y = y)$.

If X and Y are independent random variables, then knowledge of whether or not X occurs gives no information about whether or not Y occurs. It follows that if f and g are functions, then $f(X)$ gives no information about whether or not $g(Y)$ occurs and hence $f(X)$ and $g(Y)$ are independent random variables.

Theorem 3. Let X and Y be independent random variables. Then for any functions f and g, $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$. Letting f and g be the identity function gives $E[XY] = E[X]E[Y]$.

6.3 Unconditional variance

Expectation is a measure of the average behavior of a random variable. Variance and standard deviation are measures of variability. They describe how near or far typical outcomes are to the mean.

Definition 6.4. (VARIANCE.) Let X be a random variable with mean $E[X] = \mu < \infty$. The variance of X is $V[X] = E[(X-\mu)^2] = \sum_x (x-\mu)^2 P(X=x) = E[X^2] - (E[X])^2$.

The standard deviation of X is the square root of the variance.

In the variance formula, $(x-\mu)$ is the difference or "deviation" of an outcome from the mean. Thus, the variance is a weighted average of the squared deviations from the mean.

Important note 6.3. Variance and standard deviation are always nonnegative. The greater the variability of outcomes, the larger the deviations from the mean, and the greater the two measures. If X is a constant, and hence has no variability, then $X = E[X] = \mu$ and we see from the variance formula that $V[X] = 0$. The converse is also true. That is, if $V[X] = 0$, then X is almost certainly a constant.

The variance is a "cleaner" mathematical formula than the standard deviation because it does not include the square root. For simplicity and for connections with other areas of mathematics, mathematicians and probabilist often prefer variance when working with random variables.

However, the standard deviation can be easier to interpret particularly when working with data. In statistics, random variables are often used to model data, which have some associated units attached to their measurements. For instance, we might take a random person's height and assign it to a random variable H. The units are inches. The expected height $E[H]$ is also expressed in inches. However, because of the square in the variance formula, the units of the variance $V[H]$ are square inches. The square root in the standard deviation brings the units back to the units of the data.

In statistics, there are analogous definitions of mean, variance, and standard deviation for a collection of data. For a list of measurements $x_1, ..., x_n$, the sample mean is the average $\bar{x} = (x_1 + ... + x_n)/n$. The sample variance is defined as $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$. The sample standard deviation is the square root of the sample variance.

The R commands mean(vec), var(vec), and sd(vec) compute these quantities for a vector vec.

Many probability distributions are fairly concentrated near their expectation in the sense that the probability that an outcome is within a few standard deviations from the mean is high. This is particularly true for symmetric, "bell-shaped" distributions such as the Poisson distribution when λ is large and the binomial distribution when p is close to $1/2$. The same is true for many datasets. For instance, the average height of women in this country is roughly 64 inches with standard deviation 3 inches. Roughly 95% of all adult women's heights are within two standard deviations of the mean, which is within $64 \pm 2(3)$, between 58 and 70 inches.

When we discuss the normal distribution, this will be made more precise. In the meantime, a rough "rule of thumb" is that for many symmetric and near-symmetric probability distributions, the mean and variance (or standard deviation) are good summary measures for describing the behavior of "typical" outcomes of the underlying random experiment. Typically most outcomes from such distributions fall within two standard deviations of the mean.

Example 6.9. Let W, X, Y, Z be random variables with the following probability mass functions:

$$P(W=4)=1$$
$$P(X=k)=\begin{cases}1/25, & k=1,7\\ 3/25, & k=2,6\\ 5/25, & k=3,5\\ 7/25, & k=4.\end{cases}$$
$$P(Y=k)=1/7, \quad k=1,...,7$$
$$P(Z=1)=P(Z=7)=1/2.$$

Find the variance of the above random variables.

Answer 6.7. It is easy to see that all the variables have the same mean, that is 4.

The variance of W is $V[W]=(4-4)^2P(W=4)=0$.

The variance of X is

$$\begin{aligned}V[X] &= (1-4)^2P(X=1)+(2-4)^2P(X=2)+(3-4)^2P(X=3)+(4-4)^2P(X=4)\\ &\quad +(5-4)^2P(X=5)+(6-4)^2P(X=6)+(7-4)^2P(X=7)\\ &= 3^2*1/25+2^2*3/25+1^2*5/25+1^2*5/25+2^2*3/25+3^2*1/25=2.08.\end{aligned}$$

The variance of Y is

$$\begin{aligned}V[Y] &= (1-4)^2P(Y=1)+(2-4)^2P(Y=2)+(3-4)^2P(Y=3)+(4-4)^2P(Y=4)\\ &\quad +(5-4)^2P(Y=5)+(6-4)^2P(Y=6)+(7-4)^2P(Y=7)\\ &= 3^2*1/7+2^2*1/7+1^2*1/7+1^2*1/7+2^2*1/7+3^2*1/7=4.\end{aligned}$$

The variance of Z is $V[Z]=(1-4)^2P(Z=1)+(7-4)^2P(Z=7)=9$.

Example 6.10. Suppose X is uniformly distributed on $\{1,...,n\}$. Find the variance of X.

Answer 6.8. As X is uniformly distributed on $\{1,...,n\}$, we have $P(X=k)=1/n,\ \forall\ k\in\{1,...,n\}$. The expectation of X is $E[X]=\dfrac{1+...+n}{n}=\dfrac{n+1}{2}$. To compute the variance of X, we first calculate $E[X^2]$. We have $P(X^2=k^2)=P(X=k)=1/n\ \forall\ k\in\{1,...,n\}$. The expectation of X^2 is thus

$$E[X^2]=\frac{1^2+...+n^2}{n}=\frac{n(n+1)(2n+1)}{6n}=\frac{(n+1)(2n+1)}{6}.$$

The variance of X can now be calculated as: $V[X]=E[X^2]-(E[X])^2=\dfrac{(n+1)(2n+1)}{6}-(\dfrac{n+1}{2})^2=\dfrac{n^2-1}{12}$.

Example 6.11. For an event A, let I_A be the corresponding indicator random variable. Find the variance of I_A.

Answer 6.9. We have $I_A=\begin{cases}1, & \text{if } A \text{ occurs}\\ 0, & \text{if } A \text{ does not occur}\end{cases}$ and $I_A^2=\begin{cases}1, & \text{if } A \text{ occurs}\\ 0, & \text{if } A \text{ does not occur}\end{cases}$.

The expectation of I_A and I_A^2 are: $E[I_A^2]=E[I_A]=1*P(A)+0*P(A^c)=P(A)$. Thus the variance of I_A is $V[A]=E[I_A^2]-(E[I_A])^2=P(A)-P(A)^2=P(A)(1-P(A))=P(A)P(A^c)$.

Some important properties of Variance

1. For any constants a and b, and random variable X, we have $V[aX + b] = a^2V[X]$.
2. For any random variables X, Y we have $V[X + Y] = V[X] + V[Y] + 2(E[XY] - E[X]E[Y])$. In particular, if X and Y are independent, then $V[X + Y] = V[X] + V[Y]$.

Example 6.12. Suppose $X = I_1 + ... + I_n$ is the sum of n independent indicator variables with success probability p. Then X has a binomial distribution with parameters n and p, i.e., $X \sim Binom(n, p)$. Find the variance of X.

Answer 6.10. Since the I_k's are independent, the variance of the sum of indicators is equal to the sum of the variances and thus

$$V[X] = V[\sum_{k=1}^{n} I_k] = \sum_{k=1}^{n} V[I_k] = \sum_{k=1}^{n} p(1-p) = np(1-p).$$

Example 6.13. Danny said he flipped 100 pennies and got 70 heads. Is this believable?

Answer 6.11. The number of heads has a binomial distribution with parameters $n = 100$ and $p = 1/2$. The mean number of heads is $np = 50$. The standard deviation is $\sqrt{np(1-p)} = \sqrt{25} = 5$. We expect most outcomes from tossing 100 coins to fall with two standard deviations of the mean, that is between 40 and 60 heads. As 70 heads represents an outcome that is four standard deviations from the mean, we are a little suspicious of Danny's claim.

This example is very interesting as it show us how to evaluate random information rationally. It should be emphasized that most outcomes of a random experiment fall within two standard deviations from the mean.

6.4 UNCONDITIONAL COVARIANCE AND CORRELATION

Having looked at measures of variability for individual and independent random variables, we now consider measures of variability between dependent random variables. The covariance is a measure of the association between two random variables.

Definition 6.5. For random variables X and Y, the covariance between X and Y is

$$Cov(X, Y) = E[XY] - E[X]E[Y] = E[(X - E[X])(Y - E[Y])].$$

For independent random variables, $E[XY] = E[X]E[Y]$ and thus $Cov(X, Y) = 0$. The covariance will be positive when large values of X are associated with large values of Y and small values of X are associated with small values of Y. In particular, for outcomes x and $y$products of the form $(x - E[X])(y - E[Y])$ in the covariance formula will tend to either both be positive or both be negative, both cases resulting in positive values.

On the other hand, if X and Y are inversely related, most terms $(x - E[X])(y - E[Y])$ will be negative, since when X takes values above the mean, Y will tend to fall below the mean, and vice versa. In this case, the covariance between X and Y will be negative.

Important note 6.4. Covariance is a measure of linear association between two variables. In a sense, the "less linear" the relationship, the closer the covariance is to 0.

The sign of the covariance indicates whether two random variables are positively or negative associated. But the magnitude of the covariance can be difficult to interpret. The correlation is an alternative measure.

Definition 6.6. (CORRELATION). The correlation between X and Y is $Corr(X, Y) = \dfrac{Cov(X, Y)}{SD[X]SD[Y]}$.

Important note 6.5. (Properties of correlation)

1. $-1 \leq Corr(X, Y) \leq 1$.
2. If $Y = aX + b$ is a linear function of X for constants a and b, then $Corr(X, Y) = \pm 1$, depending on the sign of a.
3. Correlation is a common summary measure in statistics. Dividing the covariance by the standard deviations creates a "standardized" covariance, which is a unitless measure that takes values between -1 and 1. The correlation is exactly equal to ± 1 if Y is a linear function of X.
4. Random variables that have correlation, and covariance, equal to 0 are called uncorrelated.

Definition 6.7. We say random variables X and Y are uncorrelated if $E[XY] = E[X]E[Y]$, that is, if $Cov(X, Y) = 0$.

Important note 6.6. If random variables X and Y are independent, then they are uncorrelated. However, the converse is not necessarily true. For instance, let X be uniformly distributed on $\{-1, 0, 1\}$. Let $Y = X^2$. The two random variables are not independent as Y is a function of X. However, $Cov(X, Y) = Cov(X, X^2) = E[X^3] - E[X]E[X^2] = 0 - 0 = 0$. The random variables are uncorrelated.

Example 6.14. The number of defective parts in a manufacturing process is modeled as a binomial random variable with parameters n and p. Let X be the number of defective parts, and let Y be the number of non-defective parts. Find the covariance between X and Y.

Answer 6.12. Observe that $Y = n - X$ is a linear function of X. Thus $Corr(X, Y) = -1$. Since "failures" for X are "successes" for Y , Y has a binomial distribution with parameters n and $1 - p$.

To find the covariance, rearrange the correlation formula

$$Cov(X, Y) = Corr(X, Y)SD[X]SD[Y] = (-1)\sqrt{np(1-p)}\sqrt{n(1-p)p} = -np(1-p).$$

Example 6.15. Let R and B be random variables having the joint probability mass distribution illustrated in the following table: Find the covariance between R and B.

	$B = 0$	$B = 1$	$B = 2$	Sum
$R = 0$	3/36	6/36	1/36	10/36
$R = 1$	12/36	8/36	0	20/36
$R = 2$	6/36	0	0	6/36
Sum	21/36	14/36	1/36	

Table 1: Joint distribution table of R and B

Answer 6.13. The problem is to find $Cov(R, B) = E[RB] - E[R]E[B]$. From the table, we have:

$P(R = 0) = 10/36, P(R = 1) = 20/36, P(R = 2) = 6/36 \rightarrow E[R] = 0 * 10/36 + 1 * 20/36 + 2 * 6/36 = 32/36.$

$P(B = 0) = 21/36, P(R = 1) = 14/36, P(R = 2) = 1/36 \rightarrow E[B] = 0 * 21/36 + 1 * 14/36 + 2 * 1/36 = 16/36.$

$E[RB] = 0 * 0 * 3/36 + 0 * 1 * 6/36 + 0 * 2 * 10/36 + 1 * 0 * 12/36 + 1 * 1 * 8/36 + 1 * 2 * 0$

$+2 * 0 * 6/36 + 2 * 1 * 0 + 2 * 2 * 0 = 8/36.$

$Cov(R, B) = E[RB] - E[R]E[B] = 8/36 - 32/36 * 16/36 = -14/81.$

The negative result should not be surprising. There is an inverse association between R and B since the more balls there are of one color in the sample the less balls there will be of another color.

Theorem 4. (GENERAL FORMULA FOR VARIANCE OF A SUM.) For random variables X and Y with finite variance, we have:

$$V[X+Y] = V[X] + V[Y] + 2Cov(X,Y),$$

and

$$V[X-Y] = V[X] + V[Y] - 2Cov(X,Y).$$

If X and Y are uncorrelated, $V[X \pm Y] = V[X] + V[Y]$.

Theorem 5. For random variables X, Y, and Z, and constants a, b, c, we have:

$$Cov(aX + bY + c, Z) = aCov(X,Z) + bCov(Y,Z)$$

and

$$Cov(X, aY + bZ + c) = aCov(X,Y) + bCov(X,Z).$$

Given a random variable X with mean μ and variance σ^2, the standardized variable X^* is defined as $X* = \frac{X-\mu}{\sigma}$. Observe that $E[X^*] = E[\frac{X-\mu}{\sigma}] = \frac{E[X-\mu]}{\sigma} = \frac{E[X]-\mu}{\sigma} = 0$ and $V[X^*] = V[\frac{X-\mu}{\sigma}] = \frac{V[X-\mu]}{\sigma^2} = \frac{V[X]}{\sigma^2} = 1$. "Standardizing" a random variable gives a new random variable with mean 0 and variance one.

Theorem 6. For random variables X and Y, we have $-1 \leq Corr(X,Y) \leq 1$. If $Corr(X,Y) = \pm 1$, then there exists constants a and b such that $Y = aX + b$.

Example 6.16. After a severe storm, the number of claims received by an insurance company for hail H and tornado T damages are each modeled with a Poisson distribution with respective parameters 400 and 100. The correlation between H and T is 0.75. Let Z be the total number of claims from hail and tornado. Find the variance and standard deviation of Z.

Answer 6.14. We have $Z = H + T$, where $H \sim Pois(400), T \sim Pois(100)$. The correlation between H and T is $Corr(H,T) = 0.75$. The problem is to find $V[H] = V[H] + V[T] + 2Cov(H,T) = V[H] + V[T] + 2Corr(H,T)SD(H)SD(T) = 400 + 100 + 2 * 0.75 * \sqrt{400}\sqrt{100} = 800$. The standard deviation of Z is thus $20\sqrt{2}$.

7 Some common discrete random variables

The expectation is a numerical measure that summarizes the typical, or average, behavior of a random variable.

Definition 7.1. If X is a discrete random variable that takes values in a set $\mathbf{S}$, the expectation $E[X]$ is defined as $E[X] = \sum_{x \in \mathbf{S}} xP(X = x)$.

The sum in the definition is over all values of X. If the sum is a divergent infinite series, we say that the expectation of X does not exist. Expectation is a weighted average of the values of X, where the weights are the corresponding probabilities of those values. The expectation places more weight on values that have greater probability.

In the case when X is uniformly distributed on a finite set $\{x_1, ..., x_n\}$, that is, all outcomes are equally likely, we have $E[X] = \frac{1}{n}(x_1 + ... + x_n)$. With equally likely outcomes the expectation is just the regular average of the values. Other names for expectation are mean and expected value. In the context of games and random experiments involving money, expected value is often used.

We interpret $E[X]$ as a long-run average. If we repeat the random experiment many times, then $E[X]$ is the average value of the random variable. More formally, let $X_1, X_2, ...$ be an i.i.d. sequence of outcomes of an

experiment, where X_k is the outcome of the kth experiment. Then the interpretation of expectation is that $E[X] \approx \dfrac{X_1 + ... + X_n}{n}$ when n is large. This gives a prescription for simulating the expectation of a random variable X: choose n large, simulate n copies of X, and take the average as an approximation for $E[X]$.

7.1 Indicator (Bernoulli) Random Variables

Given an event A, define a random variable I_A such that $I_A = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{if } A \text{ does not occur} \end{cases}$. Therefore, I_A equals 1, with probability $P(A)$, and 0, with probability $P(A^c)$. Such a random variable is called an indicator variable. An indicator is a Bernoulli random variable with $p = P(A)$.

The expectation of an indicator variable is important enough to highlight.

Important note 7.1. (EXPECTATION OF INDICATOR VARIABLE.) $E[I_A] = (1)P(A)+(0)P(A^c) = P(A)$.

This is fairly simple, but nevertheless extremely useful and interesting, because it means that probabilities of events can be thought of as expectations of indicator random variables.

Often random variables involving counts can be analyzed by expressing the count as a sum of indicator variables.

(Expectation of binomial distribution.) Let $I_1, ..., I_n$ be a sequence of i.i.d. Bernoulli (indicator) random variables with success probability p. Let $X = I_1 + ... + I_n$. Then X has a binomial distribution with parameters n and p. By linearity of expectation,

$$E[X] = E[I_1 + ... + I_n] = E[I_1] + ... + E[I_n] = np.$$

This result should be intuitive. For instance, if you roll 600 dice, you would expect 100 ones. The number of ones has a binomial distribution with $n = 600$ and $p = 1/6$. And $np = 100$.

Example 7.1. At graduation ceremony, a class of n seniors, upon hearing that they have graduated, throw their caps up into the air in celebration. Their caps fall back to the ground uniformly at random and each student picks up a cap. What is the expected number of students who get their original cap back (a "match")?

Answer 7.1. Let X be the number of matches. Define A_k be the event that the kth student get his/her cap back. The corresponding indicator random variable is

$$I_{A_k} = \begin{cases} 1, & \text{if } A_k \text{ occurs} \\ 0, & \text{if } A_k \text{ does not occur.} \end{cases}$$

for $k = 1, ..., n$. Then $X = I_{A_1} + ... + I_{A_n}$. As the caps fall back to the ground uniformly at random, the probability of the kth student get his/her cap back is $1/n$ (there are n caps to choose from and only one belongs to the kth student.). Mathematically, $P(A_k) = \dfrac{1}{n}$.
The expected number of matches is

$$E[X] = E[\sum_{k=1}^{n} I_{A_k}] = \sum_{k=1}^{n} E[I_{A_k}] = \sum_{k=1}^{n} P(A_k) = \sum_{k=1}^{n} \frac{1}{n} = 1.$$

Remarkably, the expected number of matches is one, independent of the number of people n. If everyone in China throws their hat up in the air, on average about one person will get their hat back.

The indicator random variables $I_1, ..., I_n$ in the matching problem are not independent. In particular, if $I1 = ... = I_{n-1} = 1$, that is, if the first $n-1$ people get their hats back, then necessarily $I_n = 1$, the last person must also get their hat back.

7.2 Uniform random variable

If a random variable X takes values in a finite set all of whose elements are equally likely, we say that X is uniformly distributed on that set.

Definition 7.2. (Discrete uniform random variable.) Let $\mathbf{S} = \{s_1, ..., s_k\}$ be a finite set. A random variable X is uniformly distributed on $\mathbf{S}$ if $P(X = s_i) = 1/k$, for $i = 1, ..., k$. We write $X \sim Unif(\mathbf{S})$.

Example 7.2. Rachel picks an integer "at random" between 1 and 50.

1. Find the probability that she picks 13.
2. Find the probability that her number is between 10 and 20.
3. Find the probability that her number is prime.

Answer 7.2. Let X be Rachel's number. It is reasonable to assume that X is uniformly distributed on $\{1, ..., 50\}$ as all the numbers between 1 and 50 are equally likely selected.

1. The probability that Rachel picks 13 is $P(X = 13) = 1/50$.
2. There are 11 numbers between 10 and 20 (including 10 and 20). The desired probability is $P(10 \leq X \leq 20) = 11/50 = 0.22$.
3. One counts 15 prime numbers between 1 and 50. Thus, $P(X\text{is prime}) = 15/50 = 0.3$.

Important note 7.2. We write $\{X = 2\}$ for the event that the random variable takes the value 2. More generally, we write $\{X = x\}$ for the event that the random variable X takes the value x, where x is a specific number. The difference between the uppercase X (a random variable) and the lowercase x (a number) can be confusing but is extremely important to clarify.

Example 7.3. Roll a pair of dice and then sum the two dice. What is the probability of getting a 4?

Answer 7.3. Assuming the dice fair, each die number is equally likely. There are six possibilities for the first roll, six possibilities for the second roll, so $6 * 6 = 36$ possible rolls. We thus assign the probability of $1/36$ to each dice pair. Let X be the sum of the two dice. Then $P(X = 4) = P(\{(1,3),(3,1),(2,2)\}) = P((1,3)) + P((3,1)) + P((2,2)) = 3/36 = 1/12$.

Example 7.4. Find the expectation of uniform random variable X on $\{1, 2, ..., n\}$.

Answer 7.4. The expectation of X is $E[X] = \sum_{k=1}^{n} kP(X = k) = \sum_{k=1}^{n} k\frac{1}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$.

7.3 Poisson random variables

The binomial setting requires a fixed number n of independent trials. However, in many applications, we model counts of independent outcomes where there is no prior constraints on the number of trials. Examples include

1. The number of alpha particles emitted by a radioactive substance in 1 minute.

2. The number of wrong numbers you receive on your cell phone over a month's time.
3. The number of babies born on a maternity ward in one day.
4. The number of chocolate chips in a cookie.
5. The number of accidents on a mile-long stretch of highway.

Definition 7.3. A random variable X has a Poisson distribution with parameter $\lambda > 0$ if $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$, for $k = 0, 1, ...$ We write $X \sim Pois(\lambda)$.

The probability function is nonnegative and sums to 1, as

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = e^0 = 1. \tag{4}$$

Theorem 7. Let $X_1, ..., X_k$ be a sequence of independent Poisson random variables with respective parameters $\lambda_1, ..., \lambda_k$. Then $X_1 + ... + X_k \sim Pois(\lambda_1 + ... + \lambda_k)$.

Example 7.5. Find the expectation of a Poisson random variable $X \sim P(\lambda)$.

Answer 7.5. The expectation of X is

$$E[X] = \sum_{k=1}^{\infty} kP(X = k) = \sum_{k=1}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

8 Exercise

1. The following dice game costs \$10 to play. If you roll $1, 2$, or 3, you lose your money. If you roll 4 or 5, you get your money back. If you roll a 6, you win \$24.
   (a) Find the distribution of your winnings W.
   (b) Find the expected value of the game.
2. Suppose X takes values $-1, 0$, and 3, with respective probabilities $0.1, 0.3$, and 0.6. Find $V[X]$.
3. Suppose $E[X] = 2$ and $V[X] = 3$. Find
   (a) $E[(3 + 2X)^2]$.
   (b) $V[4 - 5X]$.
4. Let $X \sim Unif\{-2, -1, 0, 1, 2\}$.
   (a) Find $E[X]$.
   (b) Find $E[e^X]$.
   (c) Find $E[1/(X + 3)]$.
5. You are dealt five cards from a standard deck. Let X be the number of aces in your hand. Find $E[X]$.
6. There is a 70% chance that a tree is infected with either root rot or bark disease. The chance that it does not have bark disease is 0.4. Whether or not a tree has root rot is independent of whether it has bark disease. Find the probability that a tree has root rot.
7. Suppose X has a Poisson distribution and $P(X = 2) = 2P(X = 1)$. Find $P(X = 3)$.
8. The number of eggs a chicken hatches is a Poisson random variable. The probability that the chicken hatches no eggs is 0.10. What is the probability that she hatches at least two eggs?

9. Cars pass a busy intersection at a rate of approximately 16 cars per minute. What is the probability that at least 1000 cars will cross the intersection in the next hour?

10. * (Elevator problem) An elevator containing p passengers is at the ground floor of a building with n floors. On its way to the top of the building, the elevator will stop if a passenger needs to get off. Passengers get off at a particular floor with probability $1/n$. Find the expected number of stops the elevator makes. (Hint: Use indicators, letting $I_k = 1$ if the kth floor is a stop. Be careful: more than one passenger can get off at a floor.)

11. * In a class of 25 students, what is the expected number of months in which at least two students are born? Assume birth months are equally likely. Hint: Use indicators.

9 Joint distribution of two random variables

In the case of two random variables X and Y, a joint distribution specifies the values and probabilities for all pairs of outcomes. For constants $a < b$ and $c < d$,

$$P(a \le X \le b, c \le Y \le d) = \begin{cases} \sum_{x=a}^{b} \sum_{y=c}^{d} \mathbf{P}(X = x, Y = y), \text{ if } X, Y \text{ are discrete with joint probability mass } \mathbf{P}, \\ \int_a^b \int_c^d f(x,y)dydx \text{ if } X, Y \text{ are continuous with joint probability density } f \end{cases}$$

From the joint distribution of X and Y, we can easily obtain the (marginal) distribution of X and Y. In particular, if X and Y are discrete then the probability mass function P_X, P_Y are determined as:
$P_X(X = x) = \sum_{y=-\infty}^{\infty} \mathbf{P}(X = x, Y = y)$ and $P_Y(Y = y) = \sum_{x=-\infty}^{\infty} \mathbf{P}(X = x, Y = y)$.
If X and Y are continuous then the probability density f_X, f_Y are determined by:
$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$.

Example 9.1. The joint probability mass function of X and Y is $P(X = x, Y = y) = cxy$, for $x, y = 0, 1, 2$.
i) Find c
ii) Find $P(X \le 1, Y \ge 1)$.

Answer 9.1. i) The joint probability mass function of X and Y must satisfy:

$$\begin{aligned} 1 &= \sum_{x\in\{0,1,2\}} \sum_{y\in\{0,1,2\}} P(X = x, Y = y) = P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 0, Y = 2) + P(X = 1, Y = 0) \\ &\quad + P(X = 1, Y = 1) + P(X = 1, Y = 2) + P(X = 2, Y = 0) + P(X = 2, Y = 1) + P(X = 2, Y = 2) \\ &= 0 + 0 + 0 + 0 + 1c + 2c + 0 + 2c + 4c = 9c. \end{aligned}$$

Thus $c = 1/9$.

ii) $P(X \le 1, Y \ge 1) = \sum_{x\in\{0,1\}} \sum_{y\in\{1,2\}} P(X = x, Y = y) = P(X = 0, Y = 1) + P(X = 0, Y = 2) + P(X = 1, Y = 1) + P(X = 1, Y = 2) = 1/9 + 2/9 = 1/3$.

Example 9.2. A bag contains four red, three white, and two blue balls. A sample of two balls is picked without replacement. Let R and B be the number of red and blue balls, respectively, in the sample.
(i) Find the joint probability mass function of R and B.
(ii) Use the joint pmf to find the probability that the sample contains at most one red and one blue ball.

Answer 9.2. Consider the event $\{R = r, B = b\}$. The number of red and blue balls in the sample must be between 0 and 2. For $0 \le r + b \le 2$, if r red balls and b blue balls are picked, then $2 - r - b$ white balls must also be picked. Selecting r red, b blue, and $2 - r - b$ white balls can be done in $C_4^r C_2^b C_3^{2-r-b}$ ways. There are $C_9^2 = 36$ ways to select two balls from the bag. Thus, the joint probability mass function of (R, B) is $P(R = r, B = b) = \dfrac{C_4^r C_2^b C_3^{2-r-b}}{36}$, for $0 \le r + b \le 2, 0 \le r, b$. In particular, we have:

$$P(R=0,B=0) = \frac{C_4^0 C_2^0 C_3^2}{36} = \frac{3}{36}, \quad P(R=0,B=1) = \frac{C_4^0 C_2^1 C_3^1}{36} = \frac{6}{36}, \quad P(R=0,B=2) = \frac{C_4^0 C_2^2 C_3^0}{36} = \frac{1}{36},$$
$$P(R=1,B=0) = \frac{C_4^1 C_2^0 C_3^1}{36} = \frac{12}{36}, \quad P(R=1,B=1) = \frac{C_4^1 C_2^1 C_3^0}{36} = \frac{8}{36}, \quad P(R=1,B=2) = 0,$$
$$P(R=2,B=0) = \frac{C_4^2 C_2^0 C_3^0}{36} = \frac{6}{36}, P(R=2,B=1) = 0, \quad P(R=2,B=2) = 0.$$

ii) The desired probability is

$$P(R \le 1, B \le 1) = P(R=0,B=0) + P(R=0,B=1) + P(R=1,B=0) + P(R=1,B=1) = \frac{3+6+12+8}{36} = \frac{29}{36}.$$

Example 9.3. A bag contains four red, three white, and two blue balls. A sample of two balls is picked without replacement. Let R and B be the number of red and blue balls, respectively, in the sample.
(i) Find the joint probability mass function of R and B.
ii) Find the marginal distributions of the number of red and blue balls, respectively.
iii) Find the expected number of red balls and the expected number of blue balls in the sample.

Answer 9.3. i) The joint probability function is $P(R = r, B = b) = \dfrac{C_4^r C_2^b C_3^{2-r-b}}{36}$, for $0 \le r + b \le 2, 0 \le r, b$. In particular, we have:

$$P(R=0,B=0) = \frac{C_4^0 C_2^0 C_3^2}{36} = \frac{3}{36}, \quad P(R=0,B=1) = \frac{C_4^0 C_2^1 C_3^1}{36} = \frac{6}{36}, \quad P(R=0,B=2) = \frac{C_4^0 C_2^2 C_3^0}{36} = \frac{1}{36},$$
$$P(R=1,B=0) = \frac{C_4^1 C_2^0 C_3^1}{36} = \frac{12}{36}, \quad P(R=1,B=1) = \frac{C_4^1 C_2^1 C_3^0}{36} = \frac{8}{36}, \quad P(R=1,B=2) = 0,$$
$$P(R=2,B=0) = \frac{C_4^2 C_2^0 C_3^0}{36} = \frac{6}{36}, P(R=2,B=1) = 0, \quad P(R=2,B=2) = 0.$$

ii) The marginal distribution of the number of red balls is

$$P(R=0) = \sum_{b=0}^{2} P(R=0,B=b) = P(R=0,B=0) + P(R=0,B=1) + P(R=0,B=2) = \frac{3+6+1}{36} = \frac{10}{36},$$
$$P(R=1) = \sum_{b=0}^{2} P(R=1,B=b) = P(R=1,B=0) + P(R=1,B=1) + P(R=1,B=2) = \frac{12+8+0}{36} = \frac{20}{36}.$$
$$P(R=2) = \sum_{b=0}^{2} P(R=2,B=b) = P(R=2,B=0) + P(R=2,B=1) + P(R=2,B=2) = \frac{6+0+0}{36} = \frac{6}{36}.$$

The marginal distribution of the number of blue balls is

$$P(B=0)=\sum_{r=0}^{2}P(R=r,B=0)=P(R=0,B=0)+P(R=1,B=0)+P(R=2,B=0)=\frac{3+12+6}{36}=\frac{21}{36},$$
$$P(B=1)=\sum_{r=0}^{2}P(R=r,B=1)=P(R=0,B=1)+P(R=1,B=1)+P(R=2,B=1)=\frac{6+8+0}{36}=\frac{14}{36}.$$
$$P(B=2)=\sum_{r=0}^{2}P(R=r,B=2)=P(R=0,B=2)+P(R=1,B=2)+P(R=2,B=2)=\frac{1+0+0}{36}=\frac{1}{36}.$$

iii) The expected number of red balls is $E[R]=\sum_{r=0}^{2}rP(R=r)=0*\frac{10}{36}+1*\frac{20}{36}+2*\frac{6}{36}=\frac{32}{36}$.
The expected number of blue balls is $E[B]=\sum_{b=0}^{2}bP(B=b)=0*\frac{21}{36}+1*\frac{14}{36}+2*\frac{1}{36}=\frac{16}{36}$.

Important note 9.1. If X and Y are independent, then the joint probability mass function of X and Y has a particularly simple form. In the case of independence, $P(X=x,Y=y)=P(X=x)P(Y=y)$, for all x and y. The joint distribution is the product of the marginal distributions.

Example 9.4. Angel rolls a die four times and flips a coin twice. Let X be the number of ones she gets on the die. Let Y be the number of heads she gets on the coin. Assume die rolls are independent of coin flips.
(i) Find the joint probability mass function.
(ii) Find the probability that Angel gets the same number of ones on the die as heads on the coin.

Answer 9.4. i) The joint probability mass function of X and Y is

$$P(X=x,Y=y)=P(X=x)P(Y=y)=C_4^x(1/6)^x(5/6)^{4-x}C_2^y(1/2)^y(1/2)^{2-y}=C_4^x(1/6)^x(5/6)^{4-x}C_2^y(1/2)^2,$$
$$\forall\ 0\le x\le 4, 0\le y\le 2.$$

ii) The probability the number of ones is equal to to the number of heads is

$$\begin{aligned}&P(X=Y)=P(X=0,Y=0)+P(X=1,Y=1)+P(X=2,Y=2)\\&=C_4^0(1/6)^0(5/6)^{4-0}C_2^0(1/2)^2+C_4^1(1/6)^1(5/6)^{4-1}C_2^1(1/2)^2+C_4^2(1/6)^2(5/6)^{4-2}C_2^2(1/2)^2\\&=(5/6)^4*1/4+4*1/6*(5/6)^3*2*1/4+6*1/36*25/36*1/4=0.3424.\end{aligned}$$

If X and Y are independent random variables, then knowledge of whether or not X occurs gives no information about whether or not Y occurs. It follows that if f and g are functions, then $f(X)$ gives no information about whether or not $g(Y)$ occurs and hence $f(X)$ and $g(Y)$ are independent random variables.

Theorem 8. Let X and Y be independent random variables. Then for any functions f and g, $E[f(X)g(Y)]=E[f(X)]E[g(Y)]$.

Sums of random variables figure prominently in probability and statistics. To find probabilities of the form $P(X+Y=k)$, observe that $X+Y=k$ if and only if $X=i$ and $Y=k-i$ for some i. This gives

$$P(X+Y=k)=P(\cup_i\{X=i,Y=k-i\})=\sum_i P(X=i,Y=k-i).$$

Example 9.5. A nationwide survey collected data on TV usage in the United States The distribution of U.S. households by number of TVs per household is given in Table **??**. If two households are selected at random, find the probability that there are two TVs in both houses combined.

Table 2: Distribution of U.S. households by number of TVs.

TVs	0	1	2	3	4	5
Proportion of households	0.01	0.21	0.33	0.23	0.13	0.09

Answer 9.5. Let T_1 and T_2 be the number of TVs in the two households, respectively. It is reasonable to assume that T_1 and T_2 are independent.

The desired probability is

$$\begin{aligned} P(T_1 + T_2 = 2) &= P(T_1 = 0, T_2 = 2) + P(T_1 = 1, T_2 = 1) + P(T_1 = 2, T_2 = 0) \\ &= P(T_1 = 0)P(T_2 = 2) + P(T_1 = 1)P(T_2 = 1) + P(T_1 = 2)P(T_2 = 0) \\ &= 0.01 * 0.33 + 0.21^2 + 0.33 * 0.01 = 0.051. \end{aligned}$$

10 Conditional probability

The simplest stochastic process is a sequence of i.i.d random variables. Such sequences are often used to model random samples in statistics. However, most real-world system exhibit some type of dependency between variables, and an independent sequence is often an unrealistic model. Thus the study of stochastic processes really begins with conditional probability - conditional distributions and conditional expectation.
Sixty students were asked, "Would you rather be attacked by a big bear or swarming bees?" Their answers, along with gender, are collected in Table **??**, a common way to present data for two variables, in this case gender and attack preference. The table includes row and column totals, called marginals or marginal totals, and the overall total surveyed. The table of counts is the basis for creating a probability model for selecting a student at

Table 3: Contingency table

	Big bear	Swarming bees	
Female	27	9	36
Male	10	14	24
Total	37	23	60

random and asking their gender and attack preference. The sample space consists of the four possible responses $\Omega = \{(F, B), (F, S), (M, B), (M, S)\}$, where M is male, F is female, B is big bear, and S is swarming bees. The probability function is constructed from the contingency table so that the probability of each outcome is the corresponding proportion of responses. That is illustrated in Table **??**. Some questions of interest are:

Table 4: Probability function

	Big bear	Swarming bees
Female	27/60	9/60
Male	10/60	14/60

1. What is the probability that a student is male and would rather be attacked by a big bear?
2. What is the probability that a male student would rather be attacked by a big bear?

These questions are worded similarly but ask different things. The proportion of students who are male and prefer a big bear is $10/60 = 0.167$. That is, $P(M \cap B) = 0.167$. But the proportion of male students who prefer a big bear is $10/24 = 0.4167$ since there are 24 males and 10 of them prefer a big bear.

The second probability is an example of a conditional probability. In a conditional probability, some information about the outcome of the random experiment is known - in this case that the selected student is male. The probability is conditional on that knowledge.

For events A and B, the conditional probability of A given that B occurs is written $P(A|B)$. We also read this as "the probability of A conditional on B." Hence, the probability the student would rather be attacked by a big bear conditional on being male is $P(B|M) = 0.4167$.

The probability of preferring a big bear conditional on being male is computed from the table by taking the number of students who are both male and prefer a big bear as a proportion of the total number of males. That is, $P(B|M) = \dfrac{\text{Number of males who prefer big bear}}{\text{Number of males}}$. Dividing numerator and denominator by the total number of students, this is equivalent to $P(B|M) = \dfrac{P(B \cap M)}{P(M)}$. This suggests how to define the general conditional probability $P(A|B)$.

Definition 10.1. (conditional probability). For events A and B, the conditional probability of A given B is $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$, defined for $P(B) > 0$.

Example 10.1. In a population, 60% of the people have brown hair (H), 40% have brown eyes (E), and 30% have both (H and E). The probability that someone has brown eyes given that they have brown hair is $P(E|H) = \dfrac{P(EH)}{P(H)} = \dfrac{0.3}{0.6} = 0.5$.

Example 10.2. A subject in an experiment is given three tries to complete a task. On the first try, the probability of success is 0.30. If they fail, the chance of success on the second attempt is 0.50. And if they fail that, the chance of success on the third try is 0.65. What is the probability that they complete the task?

Answer 10.1. Let S_1, S_2, S_3 denote the events that the task is completed on the first, second, and third tries, respectively. The desired probability is $P(S_1 \cup S_2 \cup S_3) = 1 - P(S_1^c S_2^c S_3^c) = 1 - P(S_1^c)P(S_2^c|S_1^c)P(S_3^c|S_1^c S_2^c) = 1 - (0.70)(0.50)(0.35) = 0.8775$.

Important note 10.1. (Tree diagrams.) Tree diagrams are useful tools for computing probabilities. They often arise when events can be ordered sequentially (first one thing happens, then the next). They are also great visual aids that decompose a problem into smaller logical units. Probabilities are written on the branches of the tree, and outcomes are written at the end of each branch.

Example 10.3. Two dice are rolled. What is the probability that the first die is a 2 given that the sum of the dice is 7?

Answer 10.2. We use random variables to notate the problem. Let X_1 and X_2 be the outcomes of the first and second die, respectively. Then the sum of the dice is $X_1 + X_2$. The problem asks for $P(X_1 = 2|X_1 + X_2 = 7)$. Using the conditional probability definition, we obtain:

$$P(X_1 = 2|X_1 + X_2 = 7) = \frac{P(X_1 = 2, X_1 + X_2 = 7)}{P(X_1 + X_2 = 7)} = \frac{P(X_1 = 2, X_2 = 5)}{6/36} = \frac{1/36}{6/36} = \frac{1}{6}.$$

Example 10.4. John flips three coins. The probability of getting all heads is $(1/2)^3 = 1/8$. Suppose Amy peeks and sees that the first coin came up heads. For Amy, what is the probability that John gets all heads?

Answer 10.3. We have $P(HHH|H) = \dfrac{P(HHH \cap H)}{P(H)} = \dfrac{P(HHH)}{P(H)} = \dfrac{1/8}{1/2} = 1/4$.

An alternative way is to just consider the result of the two remaining coins. The probability to get two heads in flipping two coins is of course 1/4.

Important note 10.2. A common mistake when first working with conditional probability is to write $P(A|B) = P(A)/P(B)$. In general, this is just wrong. However, in the special case when A implies B, then it is correct since $A \cap B = A$ and thus $P(A|B) = \dfrac{P(A \cap B)}{P(B)} P(A)/P(B)$.

If the sample space can be partitioned into a collection of disjoint events $B_1, B_2, ..., B_k$, then A can be expressed as the disjoint union:

$$A = (A \cap B_1) \cup ... \cup (A \cap B_k).$$

10.1 INDEPENDENCE AND DEPENDENCE

The probability that you get an A in your math class is probably dependent on how much you study. But it probably is not dependent on the color of your roommate's hair. Intuitively, your grade and your roommate's hair color are independent events.

On the other hand, you are more likely to get an A in math class if you study hard. Most likely,

$$P(\text{A in math class}|\text{Roommate is a red head}) = P(\text{A in math class})$$

while $P(\text{A in math class}|\text{Study hard}) > P(\text{A in math class})$. This suggests the definition of independent events.

Definition 10.2. (Independent events.) Events A and B are independent if $P(A|B) = P(A)$. Equivalently, A and B are independent if $P(A \cap B) = P(A)P(B)$. Events that are not independent are said to be dependent.

Example 10.5. A card is drawn from a standard deck. Let A be the event that it is a spade. Let B be the event that it is an ace. Then $P(A|B) = \dfrac{P(AB)}{P(B)} = \dfrac{P(\text{Ace of Spades})}{P(\text{Ace})} = \dfrac{1/52}{1/13} = \dfrac{1}{4} = P(A)$. The two events are independent. This example illustrates that independence is not the same as mutually exclusive. Beginning students sometimes confuse the two. The events A and B are not mutually exclusive since $AB = \{\text{Ace of Spades}\} = \emptyset$. One can tell if two events are mutually exclusive by looking at the Venn diagram. Independence is more subtle. A Venn diagram alone will not identify independence. Knowledge of the probabilities $P(A)$, $P(B)$, and $P(AB)$ is required.

Important note 10.3. If A and B are independent, then B and A are independent. And thus $P(B|A) = P(B)$.

Definition 10.3. A collection of events is independent if for every finite subgroup $A_1, ..., A_k$, we have $P(A_1...A_k) = P(A_1)...P(A_k)$.

Definition 10.4. (Two independent σ-algebras)
Two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ are said to be independent if $P(A \cap B) = P(A) \cap P(B), \quad \forall A \in \mathcal{F}, \ B \in \mathcal{G}$. In words, two σ-algebras are independent, if any two events, one from each σ-algebra, are independent.

Two random variables X and Y are independent if and only if two corresponding σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent. A random variable X is said to be independent to a σ-algebra $\mathcal{F}$ if two σ-algebras $\sigma(X)$ and $\mathcal{F}$ are independent.

10.1.1 Sampling with and without replacement

Independence is often associated with sampling with replacement. For instance, a bowl contains 10 balls of different colors, including red and green. Pick two balls. Let R_1 be the event that the first ball is red. Let G_2 be the event that the second ball is green. If we sample with replacement, then $P(G_2|R_1) = 1/10 = P(G_2)$, and the events are independent. After the first ball is picked, it is returned to the bowl and the second selection is

made as if nothing changed.

On the other hand, sampling without replacement gives $P(G_2|R_1) = 1/9$, since once the first red ball is picked there are nine balls remaining. For $P(G_2)$ appeal to symmetry. The second ball is equally likely to be any of the 10 colors. Thus, $P(G_2) = 1/10$. The events are not independent.

Note that for finding $P(G_2)$ one can also condition on whether or not the first ball selected is green, giving $P(G_2) = P(G_2|G_1)P(G_1) + P(G_2|G_1^c)P(G_1^c) = 0 * 1/10 + 1/9 * 9/10 = 1/10$.

Typically, when sampling with replacement, successive outcomes are independent events. When sampling without replacement, they are not independent. However, when the population size is very big (when the number of balls in the bowl is large), the actual numerical probabilities resulting from the two sampling schemes are practically the same. As a mind stretch, you can think of sampling without replacement from a bowl of size n, and then let $n \to \infty$. Sampling without replacement from an "infinite bowl" gives sampling with replacement!

In statistical surveying, while many practical sampling schemes from large populations are done without replacement, the analysis is often done with replacement to exploit the computational advantages of working with independence.

10.2 NEW INFORMATION CHANGES THE SAMPLE SPACE

In Example **??**, the fact that Amy's probability of getting three heads is different from John's highlights the fact that probability is not an intrinsic "physical" property of a random experiment. In fact, the probability may change based on information and context.

When we ask what is the probability of getting all heads in three coin tosses, implicit in that question is that you have not seen the outcome of the experiment. If you see the outcome, then you know that either all heads came up or they did not, so the probability is either 1 or 0. On the other hand, when some part of the experiment is observed, then that partial information becomes relevant in the probability calculation.

Partial information about the outcome of a random experiment actually changes the set of possible outcomes, that is, it changes the sample space of the original experiment and reduces it based on new information. For the three coin tosses, and before Amy peeks, the sample space is $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$. But after she looks and sees that the first coin is heads, the sample space reduces to $\Omega = \{HHH, HHT, HTH, HTT\}$. The resulting conditional probability is a probability function computed on the restricted sample space.

10.3 Law of total probability

If conditional probabilities of the form $P(A|B_i)$ are known, then the law of total probability can be used to find $P(A)$.

Theorem 9. (Law of total probability). Let $B_1, ..., B_k$ be a sequence of events that partition the sample space. That is, the B_i are mutually exclusive (disjoint) and their union is equal to Ω. Then, for any event A, we have

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i).$$

Example 10.6. According to the Howard Hughes Medical Institute, about 7% of men and 0.4% of women are colorblind - either cannot distinguish red from green or see red and green differently from most people. In the

United State, about 49% of the poplulation is male and 51% female. A person is selected at random. What is the probability they are colorblind.

Answer 10.4. As you contemplate answering this question, you might find yourself saying, "Well, it depends - on whether you are male or female." The problem provides conditional information based on sex but the question asks for an unconditional probability.

The event C = {Colorblind} can be decomposed into the disjoint union {Colorblind} = {Colorblind and Male}∪ {Colorblind and Female}.
Let C, M, F denote the events that a random person is colorblind, male and female, respectively. By the law of total probability we have:

$$P(C) = P(C|M)P(M) + P(C|F)P(F) = 0.07 * 0.49 + 0.004 * 0.51 = 0.03634 = 3.634\%$$

Example 10.7. A company's policyholders are 20% under the age of 25 (group 1), 30% between 25 and 39 (group 2), and 50% over the age of 40 (group 3). An insurance company predicts that the likelihood that a person in group 1 or 2 or 3 will have an auto accident during the next year is 11%, or 3%, or 2%, respectively.What is the probability that a random policyholder will have an auto accident next year?

Answer 10.5. Denote the three age groups by G_1, G_2, and G_3, respectively. Let A be the event of having an auto accident. Conditioning on age group, the law of total probability gives $P(A) = P(A|G_1)P(G_1) + P(A|G_2)P(G_2) + P(A|G_3)P(G_3) = (0.11)(0.20) + (0.03)(0.30) + (0.02)(0.50) = 0.041$.

Example 10.8. In a standard deck of cards, the probability that the suit of a random card is hearts is $13/52 = 1/4$. Assume that a standard dect has one card missing. A card is picked from the deck. Find the probability that it is a heart.

Answer 10.6. Assume that the missing card can be any of the 52 cards picked uniformly at random. Let M denote the event that the missing card is a heart, and the complement event is M^c. Let H denote the event that the card that is picked from the deck is a heart. By the law of total probability, we have:

$$P(H) = P(H|M)P(M) + P(H|M^c)P(M^c) = \frac{12}{51}\frac{1}{4} + \frac{13}{51}\frac{3}{4} = \frac{1}{4}.$$

The result can also be obtained by appealing to symmetry. Since all cards are equally likely, and all four suits are equally likely, the argument by summetry gives that the desired probability is $1/4$.

Example 10.9. (Gambler's ruin). The gambler's ruin problem was discussed by mathematicians Blaise pascal and Pierre Fermat in 1656. A gambler starts with k dollars. On each play a fair coin is tossed and the gambler wins \$1 if heads occurs, or loses \$1 if tails occurs. The gambler stops when he reaches n ($n > k$) or loses all his money. Find the probability that the gambler will eventually lose.

Answer 10.7. We now make two observations. First, the gambler will eventually stop playing, either by reaching n or by reaching 0. One might argue that the gambler could play forever. However, it can be shown that that event occurs with probability 0. Second, assume that after, say, 100 wagers, the gambler's capital returns to $\$k$. Then, the probability of eventually wining $\$n$ is the same as it was initially. The memoryless character of the process means that the probability of winning $\$n$ or losing all of his money only depends on how much capital the gambler has, and not how many previous wagers the gambler made.

Let p_k denote the probability of reaching n when the gambler's fortune is k. What is the gambler's status if heads is tossed? Their fortune increases to $k + 1$ and the probability of winning is the same as it would be if

the gambler had started the game with $k+1$. Similarly, if tails is tossed and the gambler's fortune decreases to $k-1$. Hence, we have the following relationship: $p_k = \frac{1}{2}p_{k+1} + \frac{1}{2}p_{k-1}$ or

$$p_{k+1} - p_k = p_k - p_{k-1}, \quad \forall k = 1, 2, ..., n-1, \tag{5}$$

with $p_0 = 0$ and $p_n = 1$. Unwinding the recurrence gives

$$p_k - p_{k-1} = p_{k-1} - p_{k-2} = ... = p_1 - p_0 = p_1, \quad \forall k = 1, ..., n.$$

We have $p_2 - p_1 = p_1$ thus $p_2 = 2p_1$. Also $p_3 - p_2 = p_1$ thus $p_3 = 3p_1$. More generally, $p_k = kp_1, \quad \forall k = 1, ..., n$.

Summing both sides of the equation (**??**), we obtain: $\sum_{k=1}^{n-1}(p_{k+1} - p_k) = \sum_{k=1}^{n-1}(p_k - p_{k-1})$. Both sum telescope to $p_n - p_1 = p_{n-1} - p_0$, which gives $1 - p_1 = (n-1)p_1$ or $p_1 = 1/n$. Thus $p_k = k/n, \quad \forall k = 0, ..., n$.

We now can conclude that the probability that the gambler starts with $\$k$ will win $\$n$ is k/n. Hence, the probability of the gambler's ruin is $1 - k/n$.

10.4 Bayes' rule

It should be clear from many previous examples that in general $P(A|B) = P(B|A)$. The probability that someone uses hard drugs given that they smoke marijuana (fairly low) is not equal to the probability that they smoke marijuana given that they use hard drugs (fairly high—no pun intended). When conditional probabilities arise in real-world problems, they can be confusing and subject to misinterpretation. Data may often be given in the form $P(A|B)$, but what is really desired is the "inverse probability" $P(B|A)$. Bayes formula, also known as Bayes theorem, is a simple but remarkably powerful result for treating such conditional probability problems.

Sometimes, we need to find a conditional probability of the form $P(B|A)$, but what is given in the problem are reverse probabilities of the form $P(A|B)$ and $P(A|B^c)$. Bayes' rule provides a method for inverting the conditional probability.

Bayes' rule is a consequence of the law of total probability and the definition of conditional probability, as

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

Given a countable sequence of events $B_1, B_2, ...$, which partition the sample space, a more general form of Bayes' rule is

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}, \quad \forall i = 1, 2, ...$$

Example 10.10. (Diagnostic tests.) Diagnostic tests are commonly used to determine the likelihood of disease. Results are never certain with the possibility of false positives and false negatives. Confusion about conditional probability can lead to erroneous conclusions about the efficacy of a particular test.

Suppose a rare disease affects 1% of the population. A hypothetical blood test to detect the disease seems to be relatively accurate. On the one hand, the test has a 99% sensitivity, which means that if someone has the disease the chance that the test result is "positive" is 0.99. This also means that there is a 1% chance of error, called the false-negative rate. On the other hand, the test has a 90% specificity, which means that if someone does not have the disease the test will be "negative" 9 times out of 10. That is, there is a 10% false-positive rate.

The terms "sensitivity" and "specificity" are used by epidemiologists, public health workers who study the distribution patterns of disease and health events. A major tool in their arsenal is probability. Suppose a random person gets tested, and the test comes back positive. What is the probability that they actually have the disease?

Answer 10.8. Before proceeding, you might want to test your intuition and guess the answer without doing any computations. Is the probability of having the disease close to $10, 50$, or 90%? Many people, even experienced doctors, when asked this question assume that the test is fairly accurate and give a high estimate for the probability of disease. Let D be the event that a person has the disease. Let S be the event that the test comes back positive. The problem is asking for $P(D|S)$.

The data for this problem, however, are of the form $P(S|D)$ and $P(S|D^c)$. The 99% sensitivity rate means $P(S|D) = 0.99$. And the false-negative rate gives $P(S^c|D) = 0.01$. The 90% specificity rate means that $P(S|D^c) = 0.90$. And the false-positive rate gives $P(S^c|D^c) = 0.10$. We are also told that $P(D) = 0.01$. The data in the problem are probabilities that are conditional on having or not having the disease. But the problem is asking for a conditional probability given the outcome of the test. In order to solve the problem, we need to invert the conditional probability $P(D|S)$ to use the available information. By Bayes formula, $P(D|S) = \dfrac{P(S|D)P(D)}{P(S|D)P(D) + P(S|D^c)P(D^c)} = \dfrac{0.99(0.01)}{0.99(0.01) + 0.10(0.99)} = 0.091$. The chance of actually having the disease after testing positive is less than 10%. What looked like a fairly accurate diagnostic test is virtually worthless for deciding if someone has the disease.

While the final result might be perplexing, even paradoxical, the key to understanding it is the very low 1% probability of having the disease. Most people do not have the disease. Even though the diagnostic test has a low false-positive rate, the low rate applied to a large population of people who do not have the disease results in a lot of people with false positives. Imagine a hypothetical town of $10,000$. About 100 people (1%) will have the disease. If everyone takes the diagnostic test, about 99 people would test positive and one person would test negative. On the other hand, about 9900 people do not have the disease (99%). And if everyone takes the test, 8910 of them (90%) will test negative. But 990 (10%) will test positive. This means that about $99 + 990 = 1089$ people test positive. And of those, $99/1089 = 0.091$ have the disease.

Example 10.11. According to the Howard Hughes Medical Institute, about 7% of men and 0.4% of women are colorblind - either cannot distinguish red from green or see red and green differently from most people. In the United State, about 49% of the population is male and 51% female. Suppose a person is color blind. What is the probability they are male?

Answer 10.9. The problem asks for $P(M|C)$, and again we must invert the conditional probability in order to use the given data, which is conditional on sex, not color blindness. By Bayes formula,

$$P(M|C) = \frac{P(C|M)P(M)}{P(C|M)P(M) + P(C|F)P(F)} = \frac{(0.07)(0.49)}{(0.07)(0.49) + (0.004)(0.51)} = 0.9438.$$

Example 10.12. At a college, 5% of the students are math majors. Of the math majors, 10% are double majors. Collegewide, 20% of the students are double majors. What is the probability that a double major is a math major?

Answer 10.10. Let D and M denote being a double major and math major, respectively. From the assumption of the problem, we have $P(M) = 0.05$, $P(D) = 0.2$, and $P(D|M) = 0.1$. The event that a double major is a math major is $M|D$. Thus the desire probability $P(M|D) = \dfrac{P(M \cap D)}{P(D)} = \dfrac{P(M)P(D|M)}{P(D)} = \dfrac{0.05 * 0.1}{0.2} = 0.025$.

Example 10.13. The use of polygraphs (lie detectors) is controversial, and many scientists feel that they should be banned. On the contrary, some polygraph advocates claim that they are mostly accurate. We now

show an example that this claim is not really true.

Assume that one person in a company of 100 employees is a thief. To find the thief the company will administer a polygraph test to all its employees. The lie detector has the property that if a subject is a liar, there is a 95% probability that the polygraph will detect that they are lying. However, if the subject is telling the truth, there is 10% chance the polygraph will report a false positive and assert that the subject is lying.

Assume that a random employee is given the polygraph test and asked whether they are the thief. The employee says "no", and the lie detector reports that they are lying. Find the probability that the employee is in fact lying.

Answer 10.11. Let L denote the event that the employee is a liar. Let D denote the event that the lie detector reports that the employee is a liar. The desired probability is $P(L|D)$. By Bayes' rule, we have:

$$P(L|D) = \frac{P(D|L)P(L)}{P(D|L)P(L) + P(D|L^c)P(L^c)} = \frac{0.95 * 0.01}{0.95 * 0.01 + 0.1 * 0.99} = 0.088.$$

There is less than 10% chance that the employee is in fact the thief.

Important note 10.4. Many people, when first given this problem and asked to guess the answer, choose a probability closer to 90%. The mistake is a consequence of confusing the conditional probabilities $P(L|D)$ and $P(D|L)$, the probability that the individual is a liar, given that the polygraph says they are, with the probability that the polygraph says they are lying, given that they are a liar. Since the population of truth tellers is relatively big, the number of false positives - truth teller whom the lie detector falsely records as being a liar - is also significant. In this case, about 10% if 99, or about 10 employees will be false positives. Assuming that the lie detector correctly identifies the thief as lying, there will be about 11 employees who are identified as liars by the polygraph. The probability that one of them chosen at random is infact the thief is only about 1/11.

10.4.1 Bayesian statistics

Bayes formula is intimately connected to the field of Bayesian statistics. **Statistical inference uses data to infer knowledge about an unknown parameter in a population**. For instance, 100 fish are caught and measured to estimate the mean length of all the fish in a lake. The 100 fish measurements are the sampled data, and the mean length of all the fish in the lake is the unknown parameter. In Bayesian statistics, the unknown population parameter is considered random and the tools of probability are used to make probabilistic estimates of the parameter. One conditions on the data in order to compute $P(Parameter|Data)$. For example, suppose your friend has three coins: one is fair, one is two-headed, and one is two-tailed. A coin is picked uniformly at random. It is tossed and comes up heads. Which coin is it?

In a Bayesian context, the type of coin is the unknown parameter. The outcome of the coin toss-heads in this case-is the data. Let $C = 1, 2$, or 3, depending upon whether the coin is fair, two-headed, or two-tailed, respectively. Let H denote heads. For $c = 1, 2, 3$, Bayes formula gives

$$P(Parameter|Data) = P(C = c|H) = \frac{P(H|C = c)P(C = c)}{P(H)} = \frac{P(H|C = c)}{3P(H)}.$$

By the law of total probability, $P(H) = P(H|C = 1)P(C = 1) + P(H|C = 2)P(C = 2) + P(H|C = 3)P(C =$

$3) = \frac{1}{2}\frac{1}{3} + 1\frac{1}{3} + 0\frac{1}{3} = \frac{1}{2}$. This gives

$$P(C = c|H) = \frac{2P(H|C = c)}{3} = \begin{cases} 1/3, \text{ if the coin is fair } (c = 1) \\ 2/3, \text{ if the coin is two-headed } (c = 2) \\ 0, \text{ if the coin is two-tailed } (c = 3). \end{cases}$$

In Bayesian statistics, this probability distribution is called the posterior distribution of the parameter (coin) given the data. A "best guess" of your friend's coin is that it is two-headed. It is twice as likely to be two-headed than it is to be fair.

Example 10.14. Bertrand's box paradox. The French mathematician Joseph Louis François Bertrand posed the following problem in 1889. There are three boxes. One box contains two gold coins; one box contains two silver coins; and one box contains one gold and one silver coin. A box is picked uniformly at random. A coin is picked from the box and it is gold. What is the probability that the other coin in the box is also gold?

Answer 10.12. The correct answer is 2/3. Many people feel the answer should be 1/2, according to the following logic: The gold coin must have come from one of two boxes that are equally likely, either the gold-gold box or the gold-silver box. Thus, the gold-gold box is chosen half the time. The fallacy is that once we know the coin is gold, the two boxes are not equally likely. There are three gold coins. Two of them come from the gold-gold box, and one from the gold-silver box. If the second coin is gold, it must have come from the gold-gold box and the resulting probability is two out of three.

Here is a conditional probability analysis. Let G_1 and G_2 denote that the first coin and second coin chosen are gold, respectively. Then, $P(G_2|G_1) = P(G_2G_1)/P(G1)$. The numerator is equal to the probability of picking the gold-gold box, which is 1/3. By conditioning on which box was chosen,

$$P(G_1) = \frac{1}{3}\Big(P(G_1|\text{gold-gold}) + P(G_1|\text{silver-silver}) + P(G_1|\text{gold-silver})\Big) = \frac{1}{3}\big(1 + 0 + 1/2\big) = 1/2.$$

The desired probability is $P(G_2|G_1) = (1/3)/(1/2) = 2/3$.

10.5 Conditional distributions

The distribution of a random variable X refers to the set of values of X and their corresponding probabilities. The distribution of a random variable is specified with either a probability mass function (pmf), if X is discrete, or a probability density function (pdf) if X is continuous.

For more than one variable, there is a joint distribution, specified by either a joint pmf or pdf. If X and Y are discrete random variables, their joint probability mass function is $P(X = x, Y = y)$, considered a function of x and y. If X and Y are continuous, the joint density function $f(x, y)$ satisfies

$$P(X \le x, Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)dtds, \quad \forall x, y \in \mathbb{R}.$$

For jointly distributed random variables X and Y, the conditional distribution of Y given $X = x$ is specified by either a conditional pmf or a conditional pdf.

10.5.1 Discrete case

Definition 10.5. (CONDITIONAL PROBABILITY MASS FUNCTION). If X and Y are jointly distributed discrete random variables, then the conditional probability mass function of Y given $X = x$ is $P(Y = y|X =$

$x) = \dfrac{P(X = x, Y = y)}{P(X = x)}$ defined when $P(X = x) > 0$.

The conditional pmf is a function of y, where x is treated as fixed.

Example 10.15. Max chooses an integer X uniformly at random between 1 and 100. If $X = x$, Mary then chooses an integer Y uniformly at random between 1 and x. Find the conditional pmf of Y given $X = x$.

Answer 10.13. By the structure of this two-stage random experiment, the conditional distribution of Y given $X = x$ is uniform on $\{1, ..., x\}$. Thus the conditional pmf is $P(Y = y|X = x) = \dfrac{1}{x}, \ \ \forall y = 1, ..., x.$

Important note 10.5. Note that the conditional pmf is a probability function. For fixed x, the probability $P(Y = y|X = x)$ are nonnegative and sum to 1 as

$$\sum_y P(Y = y|X = x) = \sum_y \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x)}{P(X = x)} = 1.$$

Example 10.16. The joint probability mass function of X and Y is $P(X = x, Y = y) = \dfrac{x + y}{18}, \ \ \forall x, y = 0, 1, 2.$ Find the conditional probability mass function of Y given $X = x$.

Answer 10.14. The marginal distribution of X is

$$P(X = x) = \sum_{y=0}^{2} P(X = x, Y = y) = \frac{x}{18} + \frac{x + 1}{18} + \frac{x + 2}{18} = \frac{x + 1}{6}, \ \ \forall x = 0, 1, 2.$$

The conditional probability mass function is

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{(x + y)/18}{(x + 1)/6} = \frac{x + y}{3(x + 1)}, \ \ \forall y = 0, 1, 2.$$

Example 10.17. A bag contains 2 red, 3 blue, and 4 white balls. Three balls are picked from the bag (sampling without replacement). Let B be the number of blue balls picked. Let R be the number of red balls picked. Find the conditional pmf of B given $R = 1$.

Answer 10.15. We have

$$P(B = b|R = 1) = \frac{P(B = b, R = 1)}{P(R = 1)} = \frac{C_3^b C_2^1 C_4^{3-b-1}/C_9^3}{C_2^1 C_7^2/C_9^3} = \frac{C_3^b C_4^{2-b}}{21} = \begin{cases} 2/7, & \text{if } \ b = 0, \\ 4/7, & \text{if } \ b = 1, \\ 1/7, & \text{if } \ b = 2. \end{cases}$$

Example 10.18. During the day, Sam receives text message and phone calls. The numbers of each are independent Poisson random variables with parameters λ and μ, respectively. If Sam receives n texts and phone calls during the day, find the conditional distribution of the number of texts he receives.

Answer 10.16. Let T be the number of texts Sam receives. Let C be the number of phone calls. From the assumption, $T \sim Pois(\lambda), C \sim Pois(\mu)$. As T and C are independent, $T + C \sim Pois(\lambda + \mu)$. For $0 \le t \le n$, the conditional probability mass function of T given $T + C = n$ is

$$P(T = t|T + C = n) = \frac{P(T = t, T + C = n)}{P(T + C = n)} = \frac{P(T = t, C = n - t)}{e^{-(\lambda+\mu)}\dfrac{(\lambda + \mu)^n}{n!}} = \frac{P(T = t)P(C = n - t)}{e^{-(\lambda+\mu)}\dfrac{(\lambda + \mu)^n}{n!}}$$

$$= \frac{e^{-\lambda}\dfrac{\lambda^t}{t!}e^{-\mu}\dfrac{\mu^{(n-t)}}{(n - t)!}}{e^{-(\lambda+\mu)}\dfrac{(\lambda + \mu)^n}{n!}} = \frac{n!}{t!(n - t)!}\frac{\lambda^t \mu^{(n-t)}}{(\lambda + \mu)^n} = C_n^t (\frac{\lambda}{\lambda + \mu})^t (\frac{\mu}{\lambda + \mu})^{(n-t)}.$$

The conditional distribution of T given $T + C = n$ is binomial with parameters n and $p = \lambda/(\lambda + \mu)$.

10.5.2 Continuous case

For continuous random variable X and Y, the conditional density function of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)},$$

where f_X is the marginal density function of X, which is computed by the formula: $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$. The conditional density is a function of y, where x is treated as fixed.

Conditional densities are used to compute conditional probabilities. For any subset D of $\mathbb{R}$, we have

$$P(Y \in D|X = x) = \int_D f_{Y|X}(y|x)dy.$$

Example 10.19. Random variables X and Y have joint density function $f(x,y) = e^{-x^2y}$, for $x > 1, y > 0$. Find and describe the conditional distribution of Y given $X = x$.

Answer 10.17. The marginal density function of X is $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^{\infty} e^{-x^2y}dy = -\frac{1}{x^2}e^{-x^2y}\Big|_0^{\infty} = \frac{1}{x^2}$. The conditional distribution of Y given $X = x$ is $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = x^2e^{-x^2y}$.

In the conditional density function, x^2 is treated as a constant. Thus, this is the density of an exponential distribution with parameter x^2. That is, the conditional distribution of Y given $X = x$ is exponential with parameter x^2.

Example 10.20. Random variables X and Y have joint density $f(x,y) = \frac{y}{4x}$, for $0 < y < x < 4$. Find $P(Y < 1|X = x)$ and $P(Y < 1|X = 2)$.

Answer 10.18. We have $P(Y < 1|X = x) = \int_{-\infty}^{1} f_{Y|X}(y|x)dy$, where $f_{Y|X}(y|x)$ is the conditional density function of Y given $X = x$.

The marginal density function of X is computed as $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^x \frac{y}{4x}dy = \frac{x}{8}, \quad \forall\ 0 < x < 4$. The conditional density is then $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{y/4x}{x/8} = \frac{2y}{x^2} \quad \forall\ 0 < y < x < 4$. The desired probability is

$$P(Y < 1|X = x) = \int_0^{\min(x,1)} \frac{2y}{x^2}dy = \frac{y^2}{x^2}\Big|_0^{\min(x,1)} = \frac{(\min(x,1))^2}{x^2} = \begin{cases} 1 & \text{if } x \le 1 \\ \frac{1}{x^2} & \text{if } x > 1 \end{cases}.$$

As a result, $P(Y < 1|X = 2) = \frac{1}{4}$.

Example 10.21. Random variables X and Y have joint density function $f(x,y) = e^{-x}, \quad \forall 0 < y < x < \infty$. Find $P(Y < 2|X = 5)$.

Answer 10.19. The desired probability is $P(Y < 2|X = 5) = \int_0^2 f_{Y|X}(y|5)dy$.

To find the conditional density function $f_{Y|X}(y|x)$, we first find the marginal density function:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^x e^{-x}dy = xe^{-x}, \quad \forall x > 0.$$

We now can find the conditional density function as follows:

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{e^{-x}}{xe^{-x}} = \frac{1}{x}, \quad \forall 0 < y < x.$$

From this, we deduce that $f_{Y|X}(y|5) = \frac{1}{5}$. As a result, we obtain:

$$P(Y < 2|X = 5) = \int_0^2 f_{Y|X}(y|5)dy = \int_0^2 \frac{1}{5}dy = \frac{2}{5}.$$

Example 10.22. Tom picks a real random number X uniformly distributed on $(0, 1)$. Tom then shows his number x to Marisa who then picks a real random number Y uniformly distributed on $(0, x)$.

- Compute the conditional distribution of Y given $X = x$.
- Compute the joint distribution of X and Y.
- Compute the marginal density function of Y.

Answer 10.20.

- The conditional distribution of Y given $X = x$ is given directly in the statement of the problem. This distribution is uniform on $(0, x)$. Thus $f_{Y|X}(y|x) = \frac{1}{x}, \quad \forall\ 0 < y < x$.
- For the joint density function, we obtain: $f(x, y) = f_X(x)f_{Y|X}(y|x) = (1/1) * (1/x) = 1/x, \quad \forall\ 0 < y < x < 1$.
- To find the marginal density function of Y, integrate out the x variable in the joint density function. This gives

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)dx = \int_y^1 \frac{1}{x}dx = -\ln y, \quad \forall 0 < y < 1.$$

Example 10.23. The time between successive tsunamis in the Caribbean is modeled with an exponential distribution. The parameter value of the exponential distribution is unknown and is itself modeled as a random variable Λ uniformly distributed on $(0, 1)$. Suppose the time X between the last two consecutive tsunamis was 2 years. Find the conditional density of Λ given $X = 2$.

Answer 10.21. As Λ is a uniformly distributed random variable on $(0, 1)$, the marginal density function of Λ is $f_\Lambda(\lambda) = 1$ for all $\lambda \in (0, 1)$. In addition, given the value of $\Lambda = \lambda$, X is an exponential random variable with parameter λ. That means $f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}, \ \forall\ x > 0, \lambda \in (0, 1)$. The joint density function of X and Λ is $f(x, \lambda) = f_\Lambda(\lambda)f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}, \ \forall\ x > 0, \lambda \in (0, 1)$. The marginal density function of X can now be computed as $f_X(x) = \int_{-\infty}^{\infty} f(x, \lambda)d\lambda = \int_0^1 \lambda e^{-\lambda x}d\lambda = \frac{1}{x^2}(1 - e^{-x} - xe^{-x})$.

The conditional density of Λ given $X = x$ is computed as: $f_{\Lambda|X}(\lambda|x) = \frac{f(\lambda, x)}{f_X(x)} = \frac{x^2\lambda e^{-\lambda x}}{1 - e^{-x} - xe^{-x}}$. As a result, we obtain: $f_{\Lambda|X}(\lambda|2) = \frac{2^2\lambda e^{-2\lambda}}{1 - e^{-2} - 2e^{-2}} = \frac{4\lambda e^{-2\lambda}}{1 - 3e^{-2}}$.

11 Conditional expectations

11.1 Conditional expectation given on a partition event

A conditional expectation is an expectation computed with respect to a conditional probability distribution. Write $E(Y|X = x)$ for the conditional expectation of Y given $X = x$.

Definition 11.1. (Conditional expectation of Y given $X = x$)

$$E(Y|X=x) = \begin{cases} \sum\limits_{y} yP(Y=y|X=x), \text{if X is discrete} \\ \int\limits_{-\infty}^{\infty} yf_{Y|X}(y|x)dy, \text{if X is continuous} \end{cases}$$

It is clear that $E(Y|X = x)$ is a function of x, i.e., the result depends on the value of x.

Example 11.1. A school cafeteria has two registers. Let X and Y denote the number of students in line at the first and second registers, respectively. The joint probability mass function of X and Y is specified by the following joint distribution given Table **??**. Find the expected number of people in line at the second register if

Table 5: Joint distribution table

X/Y	0	1	2	3	4
0	0.15	0.14	0.03	0	0
1	0.14	0.12	0.06	0.01	0
2	0.03	0.06	0.1	0.03	0.02
3	0	0.01	0.03	0.02	0.01
4	0	0	0.02	0.01	0.01

there is one person at the first register.

Answer 11.1. The problem asks for $E(Y|X = 1)$. We have:

$$E(Y|X=1) = \sum_{y=0}^{4} yP(Y=y|X=1) = \sum_{y=0}^{4} y\frac{P(X=1,Y=y)}{P(X=1)} = \frac{\sum\limits_{y=0}^{4} yP(X=1,Y=y)}{P(X=1)}.$$

Using the law of total probability, we obtain:

$$\begin{aligned} P(X=1) &= P(X=1,Y=0)+P(X=1,Y=1)+P(X=1,Y=2)+P(X=1,Y=3)+P(X=1,Y=4) \\ &= 0.14+0.12+0.06+0.01+0=0.33. \end{aligned}$$

Also

$$\begin{aligned} \sum_{y=0}^{4} yP(X=1,Y=y) &= P(X=1,Y=1)+2P(X=1,Y=2)+3P(X=1,Y=3)+4P(X=1,Y=4) \\ &= 0.12+2*0.06+3*0.01+4*0=0.27. \end{aligned}$$

As a result, we deduce that $E(Y|X = 1) = \dfrac{0.27}{0.33} = 0.818$. In words, it is expected that there is less than 1 person in the second cafeteria.

Example 11.2. Assume that X and Y have joint density function $f(x, y) = \dfrac{2}{xy}, \quad \forall\ 1 < y < x < e$. Find $E(Y|X = x)$.

Answer 11.2. It should be noted here that for each fixed x, Y receives values in the range $(1, x)$. The desired conditional expectation is thus $E(Y|X = x) = \int\limits_1^x yf_{Y|X}(y|x)dy = \int\limits_1^x y\dfrac{f(x,y)}{f_X(x)}dy$. The marginal density function of random variable X is computed by $f_X(x) = \int\limits_{-\infty}^{\infty} f(x,y)dy = \int\limits_1^x \dfrac{2}{xy}dy = \dfrac{2\ln x}{x}, \quad \forall\ 1 < x < e$. Thus $\dfrac{f(x,y)}{f_X(x)} = \dfrac{1}{y\ln x}, \ \forall\ 1 < y < x$. As a result, we obtain: $E(Y|X = x) = \int\limits_1^x y\dfrac{1}{y\ln x}dy = \dfrac{x-1}{\ln x}$.

Important note 11.1. 1. If X and Y are independent then $E(Y|X=x)=E(Y)$.

2. If g is a function then $E(g(X)|X=x)=g(x)$.

3. If g is a function then

$$E(g(Y)|X=x)=\begin{cases}\sum\limits_y g(y)P(Y=y|X=x), \text{if X is discrete}\\ \int_{-\infty}^{\infty} g(y)f_{Y|X}(y|x)dy, \text{if X is continuous}\end{cases}$$

11.2 Conditional expectation given on a general event

Remind that if X and Y are jointly distributed discrete random variables then

$$E(Y|X=x)=\sum_y yP(Y=y|X=x),$$

where $X=x$ is a partition event.

We now extend this definition to a general event A, i.e.,

$$E(Y|A)=\sum_y yP(Y=y|A)=\sum_y y\frac{P(\{Y=y\}\cap A)}{P(A)}=\frac{\sum\limits_y yP(\{Y=y\}\cap A)}{P(A)}=\frac{E(YI_A)}{P(A)},$$

where I_A, the indicator random variable for A, is defined by $I_A=\begin{cases}1, & \text{if } A \text{ occurs}\\ 0, & \text{if } A \text{ does not occurs.}\end{cases}$

Let $A_1,...,A_k$ be a sequence of events that partition the sample space. Observe that $I_{A_1}+I_{A_2}+...+I_{A_k}=1$ as every outcome $\omega\in\Omega$ is contained in exactly one of the A_i's. It follows that $Y=\sum\limits_{i=1}^k YI_{A_i}$. Taking expectation both sides we obtain: $E(Y)=E(\sum\limits_{i=1}^k YI_{A_i})=\sum\limits_{i=1}^k E(YI_{A_i})=\sum\limits_{i=1}^k E(Y|A_i)P(A_i)$. This is the law of total expectation, which looks very similar to the law of total probability.

Important note 11.2. We now have two ways to compute the unconditional expectation $E(Y)$, with Y is a discrete random variable on the sample space Ω. First, compute from the definition $E(Y)=\sum\limits_{i=1}^n y_iP(Y=y_i)$. Second, compute through the conditional expectation of partition events: $E(Y)=\sum\limits_{j=1}^m E(Y|A_j)P(A_j)$.

Example 11.3. A fair coin is flipped repeatedly. Find the expected number of flips needed to get two heads in a row.

Answer 11.3. Let Y be the number of flips needed. Consider three events: i) T, the first flip is tails; ii) HT, the first flip is heads and the second flip is tails; iii) HH, the first two flips are heads. The events T, HT, HH partition the sample space. By the law of total expectation, we have:

$$E(Y)=E(Y|T)P(T)+E(Y|HT)P(HT)+E(Y|HH)P(HH)=\frac{1}{2}E(Y|T)+\frac{1}{4}E(Y|HT)+\frac{1}{4}2$$

Consider $E(Y|T)$. Assume that the first flip is tails. Since successive flips are independent, after the first tails we start over waiting for two heads in a row. Since one flip has been used, it follows that $E(Y|T)=1+E(Y)$.

Similarly, after first heads and then tails, we start over again, have used up two coins tosses. Thus $E(Y|HT) = 2 + E(Y)$. As a result, we obtain:

$$E(Y) = \frac{1}{2}(1 + E(Y)) + \frac{1}{4}(2 + E(Y)) + \frac{1}{2} \rightarrow E(Y) = 6.$$

In words, we expect after 6 tosses, we will have two heads in a row.

Example 11.4. The time that Joe spends talking on the phone is exponentially distributed with mean 5 minutes (on average each call takes 5 minutes). What is the expected length of his phone call if he talks for more than 2 minutes?

Answer 11.4. Let Y be the length of Joe's phone call. With $A = \{Y > 2\}$, the desired conditional expectation is

$$E(Y|A) = \frac{E(YI_A)}{P(A)} = \frac{1}{P(Y > 2)} \int_{-\infty}^{\infty} yI_{y>2}(y)\frac{1}{5}e^{-y/5}dy = \frac{1}{P(Y > 2)} \int_{2}^{\infty} y\frac{1}{5}e^{-y/5}dy = \frac{1}{e^{-2/5}}7e^{-2/5} = 7.$$

Note that the solution can be obtained using the memoryless property of the exponential distribution. The conditional distribution of Y given $Y > 2$ is equal to the distribution of $2+Z$, where Z has the same distribution as Y. Thus

$$E(Y|Y > 2) = E(2 + Z) = 2 + E(Z) = 7.$$

11.2.1 CONDITIONAL EXPECTATION ON A FILTRATION

We consider a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$. If X is $\mathcal{G}$-measurable, then the information in $\mathcal{G}$ is sufficient to determine the value of X. If X is independent of $\mathcal{G}$, then the information in $\mathcal{G}$ provides no help in determining the value of X. In the intermediate case, we can use the information in $\mathcal{G}$ to estimate but not precisely evaluate X. The conditional expectation of X given $\mathcal{G}$ is such an estimate. The conditional expectation of a random variable over a σ-algebra $\mathcal{G}$ is itself a random variable, and this random variable is the best prediction about X, given the available information in $\mathcal{G}$.

Definition 11.2. Definition of conditional expectation. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let X be a random variable that is either nonnegative or integrable. The conditional expectation of X given $\mathcal{G}$, denoted $E[X|G]$ is any random variable that satisfies:

i) (Measurability) $E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable, and

ii) $\int_A E[X|\mathcal{G}](\omega)dP(\omega) = \int_A X(\omega)dP(\omega), \quad \forall A \in \mathcal{G}$ If $\mathcal{G}$ is the σ-algebra generated by some other random variable W (i.e., $\mathcal{G} = \sigma(W)$), we generally write $E[X|W]$ rather than $E[X|\sigma(W)]$.

Property (i) guarantees that the information in $\mathcal{G}$ is enough to determine $E[X|\mathcal{G}]$.

Property (ii) ensures that $E[X|G]$ is indeed an estimate of X. It gives the same averages as X over all the sets in $\mathcal{G}$.

The existence of a conditional expectation $E[X|\mathcal{G}]$ is proved by the Radon-Nikodym Theorem. There might be many conditional expectations $E[X|\mathcal{G}]$ but they will almost agree surely, i.e., the set of $\omega \in \Omega$ for which the random variables are different has zero probability.
Properties of conditional expectation $E[X|\mathcal{G}]$.

1. Linearity of conditional expectations. If X and Y are integrable random variables and c_1 and c_2 are constants, then $E[c_1X + c_2Y|\mathcal{G}] = c_1E[X|\mathcal{G}] + c_2E[Y|\mathcal{G}]$. Expectation of a sum is equal to a sum of expectations.

2. Taking out what is known. If X,Y, XY are integrable random variables, and X is $\mathcal{G}$-measurable, then $E[XY|\mathcal{G}] = XE[Y|\mathcal{G}]$. X is known under the available information in G and it is taken out in the process of determining $E[XY|\mathcal{G}]$.

3. Iterated conditioning (Tower property). If $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$ ($\mathcal{H}$ contains less information in $\mathcal{G}$) and X is an integrable random variable, then $E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{H}]$. Noted that $E[X|\mathcal{G}]$ is the best estimate of X given the information in $\mathcal{G}$ and $E[E[X|\mathcal{G}]|\mathcal{H}]$ is the best estimate of $E[X|\mathcal{G}]$ given the information in $\mathcal{H}$. In particular, if $\mathcal{H}$ is $\mathcal{F}_0 = \{\emptyset, \Omega\}$ the we have $E\big[E[X|\mathcal{G}]|\mathcal{F}_0\big] = E[X|\mathcal{F}_0] = E[X]$.

4. Independence. Suppose random variables $X_1, X_2, ..., X_k$ are $\mathcal{G}$-measurable and the random variables $Y_1, Y_2, ..., Y_n$ are independent of $\mathcal{G}$. Let $f(x_1, x_2, ..., x_k, y_1, ..., y_n)$ be a function of the dummy variables $x_1, x_2, ..., x_k, y_1, y_2, ..., y_n$ and define $g(x_1, ..., x_k) = Ef(x_1, ..., x_k, Y_1, ..., Y_n)$. Then $E[f(X_1, ..., X_k, Y_1, ..., Y_n)|\mathcal{G}] = g(X_1, ..., X_k)$. In particular, if X is independent of $\mathcal{G}$ then $E[X|G] = E[X]$.

5. If $\phi(x)$ is a convex function of a dummy variable x and X is integrable, then $E[\phi(x)|\mathcal{G}] \geq \phi(E[X|\mathcal{G}])$.

11.2.2 CONDITIONAL EXPECTATION ON A RANDOM VARIABLE

Let X and Y are random variable. The condition expectation of Y given X, denoted by $E(Y|X)$, is the conditional expectation of Y on the natural filtration $\sigma(X)$ of the random variable X. The conditional expectation $E(Y|X)$ has three important properties:

1. $E(Y|X)$ is a random variable.

2. $E(Y|X)$ is a function of X.

3. $E(Y|X)$ is equal to $E(Y|X = x)$ whenever $X = x$, i.e., $E(Y|X = x) = g(x)$, $\forall x$, then $E(Y|X) = g(X)$.

Example 11.5. Let X be uniformly distributed on $(0, 1)$. If $X = x$, a second number Y is picked uniformly on $(0, x)$. Find $E(Y|X)$.

Answer 11.5. The conditional distribution of Y given $X = x$ is uniform on $(0, x)$, for $0 < x < 1$. It follows that $E(Y|X = x) = x/2$. Since this is true for all x, $E(Y|X) = X/2$.

Important note 11.3. Since $E(Y|X)$ is a random variable with some probability distribution, it makes sense to take its expectation with respect to that distribution. This leads to one of the most important results in probability, the tower property of the conditional expectation. For random variables X and Y, $E(Y) = E(E(Y|X))$.

Example 11.6. Angel will harvest a random number N tomatoes in her garden, where N has a Poisson distribution with parameter λ. Each tomato is checked for defects. The chance that a tomato has defects is p. Defects are independent from tomato to tomato. Find the expected number of tomatoes with defects.

Answer 11.6. Let X be the number of tomatoes with defects. The conditional distribution of X given $N = n$ is binomial with parameters n and p. This gives $E(X|N = n) = np$. Since this is true for all n, $E(X|N) = Np$. By the tower property of conditional expectation, we have

$$E(X) = E(E(X|N)) = E(Np) = pE(N) = p\lambda.$$

Example 11.7. When Katie goes to the gym, she will either run, bicycle, or row. She will choose one of the aerobic activities with respective probabilities $0.5, 0.3$, and 0.2. And having chosen an activity the amount of time (in minutes) she spends exercising is exponentially distributed with respective parameters $0.05, 0.025$, and 0.01. Find the expectation of Katie's exercise time.

Answer 11.7. Let T be Katie's exercise time, and let A be a random variable that takes values $1, 2$, and 3 corresponding to her choice of running, bicycling, and rowing. The conditional distribution of exercise time given her choice is exponentially distributed. Thus, the conditional expectation of exercise time given her choice is the reciprocal of the corresponding parameter value. By the tower property of conditional expectation, we have: $E[T] = E[E[T|A]] = \sum_{a=1}^{3} E[T|A = a]P(A = a) = \frac{1}{0.05} * 0.5 + \frac{1}{0.025} * 0.3 + \frac{1}{0.01} * 0.2 = 42$ (minutes).

Example 11.8. (At the gym, continued). When Tyler goes to the gym, he has similar habits as Katie, except, whenever he chooses rowing he only rows for 10 minutes (the exercising time is not exponentially distributed with parameter $0, 01$ as Katie), stops to take a drink of water, and starts all over, choosing one of the three activities at random as if he had just walked in the door. Find the expectation of Tyler's exercise time.

Answer 11.8. The conditional expectation of Tyler's exercise time given that he chooses running or bicycling is the same as Katie's. The conditional expectation given that he picks rowing is $E[T|A = 3] = 10 + E[T]$ since after 10 minutes of rowing, Tyler's subsequent exercise time has the same distribution as if he had just walked into the gym and started anew. His expected exercise time is

$$E(T) = E[E[T|A]] = \sum_{a=1}^{3} E[T|A = a]P(A = a) = \frac{1}{0.05} * 0.5 + \frac{1}{0.025} * 0.3 + (10 + E[T]) * 0.2.$$

Solving the above equation, we obtain $E[T] = 30$ (minutes).

Example 11.9. Ellen's insurance will pay for a medical expense subject to a \$100 deductible. Assume that the amount of the expense is exponentially distributed with mean \$500. Find the expectation and standard deviation of the payout.

Answer 11.9. Let M be the amount of the medical expense and let X be the insurance company's payout. Then $X = \max(M - 100, 0)$ where M is exponentially distributed with parameter $1/500$. The conditional expectation of X given M, $E(X|M)$, is a function of M, denoted by $g(M)$. In particular, we have

$$g(x) = E(X|M = x) = E(\max(M - 100, 0)|M = x) = E(\max(x - 100, 0)|M = x) = \max(x - 100, 0).$$

To find the expected payment, we use the tower property of conditional expectation as follows:

$$\begin{aligned} E(X) &= E(E(X|M)) = E(g(M)) = \int_0^\infty g(x)\lambda e^{-\lambda x}dx = \int_0^\infty \max(x - 100, 0)\lambda e^{-\lambda x} \\ &= \int_{100}^\infty (x - 100)\lambda e^{-\lambda x} = 500e^{-1/5} \approx 409.365. \end{aligned}$$

For standard deviation, we first find

$$\begin{aligned} E(X^2) &= E(E(X^2|M)) = E(h(M)) = \int_0^\infty h(x)\lambda e^{-\lambda x}dx = \int_0^\infty \max(x - 100, 0)^2\lambda e^{-\lambda x} \\ &= \int_{100}^\infty (x - 100)^2\lambda e^{-\lambda x} = 500000e^{-1/5} \approx 409365. \end{aligned}$$

This gives $SD(X) = \sqrt{E(X^2) - E(X)^2} = \$491.72$.

Example 11.10. (Random sum of random variables). A stochastic model for the cost of damage from traffic accidents is given by **?**]. Let X_k be the amount of damage from an individual's kth traffic accident. Assume $X_1, X_2, ...$ is an i.i.d sequence with mean μ. The number of accidents N for an indvidual driver is random variable with mean λ. It is assumed that the number of accidents is independent of the amount of damages for each accident. That is N is independent of X_k. For an individual driver, find expected value of the total cost of damages.

Answer 11.10. Let T be the total cost of damages. Then $T = X_1 + ... + X_n = \sum_{k=1}^{N} X_k$.

The number of summands is random. The random variable T is a random sum of random variables. By the tower property, we have $E(T) = E(E(T|N))$.

In addition,

$$E(T|N=n) = E(\sum_{k=1}^{N} X_k|N=n) = E(\sum_{k=1}^{n} X_k|N=n) = E(\sum_{k=1}^{n} X_k) = \sum_{k=1}^{n} E(X_k) = n\mu,$$

where the third equality is because N is independent of X_k's. Since the final equality hods for all $n \in \mathbb{N}$, we have $E(T|N) = N\mu$. Thus $E(T) = E(E(T|N)) = E(N\mu) = \mu E(N) = \mu\lambda$.

The result is intuitive. The expected total cost of damage is the product of the expected number of accidents and the expected cost per accident.

11.3 COMPUTING PROBABILITIES BY CONDITIONING

When we first introduced indicator variables, we showed that probabilities can actually be treated as expectations, since for any event A, $P(A) = E[I_A]$, where I_A is an indicator random variable. That means $I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$. Applying the tower property of expectation gives $P(A) = E[I_A] = E[E[I_A|X]]$. If X is continuous with density function f_X, then $P(A) = E[E[I_A|X]] = \int_{-\infty}^{\infty} E[I_A|X=x]f_X(x)dx = P(A|X=x)f_X(x)dx$. This is the continuous version of the law of total probability, a powerful tool for finding probabilities by conditioning.

For instance, consider $P(X < Y)$, where X and Y are continuous. Treating $\{X < Y\}$ as the event A, and conditioning on Y, gives $P(X < Y) = \int_{-\infty}^{\infty} P(X < Y|Y = y)f_Y(y)dy = \int_{-\infty}^{\infty} P(X < y|Y = y)f_Y(y)dy$. If X and Y are independent, $P(X < Y) = \int_{-\infty}^{\infty} P(X < y|Y = y)f_Y(y)dy = \int_{-\infty}^{\infty} P(X < y)f_Y(y)dy = \int_{-\infty}^{\infty} F_X(x)f_Y(y)dy$, where $F_X(x)$ is the cumulative distribution function of X.

Example 11.11. The times F and Z that Lief and Liz arrive to class have exponential distributions with respective parameters λ_F and λ_Z. If their arrival times are independent, the probability that Liz arrives before Lief is

$$\begin{aligned} P(Z < F) &= \int_{-\infty}^{\infty} F_Z(y)f_F(y)dy = \int_{-\infty}^{\infty} (1 - e^{-\lambda_Z y})\lambda_F e^{-\lambda_F y}dy = \int_{-\infty}^{\infty} \lambda_F e^{-\lambda_F y}dy - \int_{-\infty}^{\infty} \lambda_F e^{-(\lambda_F+\lambda_Z)y}dy \\ &= 1 - \frac{\lambda_F}{\lambda_F + \lambda_Z} \int_{-\infty}^{\infty} (\lambda_F + \lambda_Z)e^{-(\lambda_F+\lambda_Z)y}dy = 1 - \frac{\lambda_F}{\lambda_F + \lambda_Z} \end{aligned}$$

Example 11.12. The density function of X is $f_X(x) = xe^{-x}$, for $x > 0$. Given $X = x$, Y is uniformly distributed on $(0, x)$. Find $P(Y < 2)$.

Answer 11.11. As given $X = x$, Y is uniformly distributed on $(0, x)$, the conditional distribution of Y given $X = x$ is $f_{Y|X}(y|x) = \dfrac{1}{x}$. We also have

$$\begin{aligned}
P(Y<2) &= E(I_{Y<2}) = E(E(I_{Y<2}|X)) = \int_{-\infty}^{\infty} E(I_{Y<2}|X=x)f_X(x)dx = \int_{-\infty}^{\infty} P(Y<2|X=x)f_X(x)dx \\
&= \int_0^{\infty}\left[\int_0^{\min(x,2)} f_{Y|X}(y|x)dy\right]xe^{-x}dx \text{ (as } 0\le y\le x, y\le 2) = \int_0^{\infty}\left[\int_0^{\min(x,2)} \frac{1}{x}dy\right]xe^{-x}dx \\
&= \int_0^{\infty}\frac{\min(x,2)}{x}xe^{-x}dx = \int_0^{2}\min(x,2)e^{-x}dx + \int_2^{\infty}\min(x,2)e^{-x}dx = \int_0^{2}xe^{-x}dx + \int_2^{\infty}2e^{-x}dx \\
&= -(3e^{-2}-1) + 2e^{-2} = 1 - e^{-2} = 0.8647.
\end{aligned}$$

Example 11.13. Bob's insurance will pay for a medical expense subject to a \$100 deductible. Suppose the amount of the expense is exponentially distributed with parameter λ. Find the distribution of the amount of the insurance company's payment.

Answer 11.12. Let M be the amount of the medical expense and let X be the insurance company's payout. Then $X = \max(M - 100, 0)$ where M is exponentially distributed with parameter λ. The problem is to find the distribution of the continuous random variable X. In particular, we find the cumulative distribution of X,i.e., find $P(X \le x)$ for each $x > 0$ as $X \ge 0$.

$$\begin{aligned}
&P(X\le x) = E(I_{X\le x}) = E(E(I_{X\le x}|M)) = \int_{-\infty}^{\infty} E(I_{X\le x}|M=m)f_M(m)dm \\
&= \int_0^{\infty} P(X\le x|M=m)\lambda e^{-\lambda m}dm = \int_0^{\infty} P(\max(M-100,0)\le x|M=m)\lambda e^{-\lambda m}dm \\
&= \int_0^{100} P(\max(M-100,0)\le x|M=m)\lambda e^{-\lambda m}dm + \int_{100}^{\infty} P(\max(M-100,0)\le x|M=m)\lambda e^{-\lambda m}dm \\
&= \int_0^{100} P(0\le x)\lambda e^{-\lambda m}dm + \int_{100}^{\infty} P(M\le 100+x|M=m)\lambda e^{-\lambda m}dm \\
&= \int_0^{100} \lambda e^{-\lambda m}dm + \int_{100}^{100+x} P(M\le 100+x|M=m)\lambda e^{-\lambda m}dm + \int_{100+x}^{\infty} P(M\le 100+x|M=m)\lambda e^{-\lambda m}dm \\
&= \int_0^{100} \lambda e^{-\lambda m}dm + \int_{100}^{100+x} \lambda e^{-\lambda m}dm + \int_{100+x}^{\infty} 0*\lambda e^{-\lambda m}dm = 1 - e^{-(100+x)\lambda}
\end{aligned}$$

11.4 Exercise

1. The number of tornadoes T and earthquakes E over a month's time in a particular region is independent and has a Poisson distribution with parameters four and two, respectively.
   (a) Find the joint pmf of T and E.
   (b) What is the probability of no tornadoes and no earthquakes in that region next month?
   (c) What is the probability of at least two tornadoes and at least two earthquakes?
   (d) What is the probability of at least two tornadoes or at most two earthquakes?

2. The joint pmf of X and Y is $P(X = x, Y = y) = \dfrac{x+1}{12}$, for $x = 0, 1$ and $y = 0, 1, 2, 3$. Find the marginal distributions of X and Y. Describe their distributions qualitatively. That is, identify their distributions as one of the known distributions you have worked with (e.g., Bernoulli, binomial, Poisson, or uniform).

3. A joint probability mass function is given by

   $P(X = 1, Y = 1) = 1/8, \quad P(X = 1, Y = 2) = 1/4, \quad P(X = 2, Y = 1) = 1/8, \quad P(X = 2, Y = 2) = 1/2.$

   (a) Find the marginal distributions of X and Y.
   (b) Are X and Y independent?
   (c) Compute $P(XY \leq 3)$.
   (d) Compute $P(X + Y > 2)$.

4. A bag contains one red, two blue, three green, and four yellow balls. A sample of three balls is taken without replacement. Let B be the number of blue balls and Y the number of yellow balls in the sample.
   (a) Find the joint probability table.
   (b) Find Cov(B, Y).

5. * The joint probability mass function of X and Y is $P(X = x, Y = y) = \dfrac{1}{e^2 y!(x - y)!}, x = 0, 1, ..., y = 0, 1, ..., x$.
   (a) Find the conditional distribution of Y given $X = x$.
   (b) Describe the distribution in terms of distributions that you know.
   (c) Without doing any calculations, find $E[Y|X = x]$ and $V[Y|X = x]$.

6. Let X be a random variable such that $P(X = k) = k/10$, for $k = 1, 2, 3, 4$. Let Y be a random variable with the same distribution as X. Suppose X and Y are independent. Find $P(X + Y = k)$, for $k = 2, ..., 8$.