Chapter 2

Random Variables

2.1 Introduction to random variables

In the previous chapter, we have been working with *events*, which is a conceptualization of real world outcomes occurred with probabilities. In this chapter, we introduce a much more powerful conceptualization that deals with uncertain outcomes — random variables, which is the foundation of all probability and statistical studies.

In high school, all mathematical models come with certainty. For example, the falling time of any object from height h down to the earth is: $t = \sqrt{\frac{2h}{g}}$, where g is the gravity constant. The outcome is *deterministic*. The variables that enter into the equation either have unknown values or known certain values. Errors are possible only due to frictions or measurement errors.

But many real world processes come naturally with uncertainty. Think about the temperature tomorrow, or the stock market returns. We can only make predictions with probabilities. Yes, you may argue this uncertainly is due to incomplete information. If we have all the knowledge regarding the climate, we can predict exactly the temperature. But given the imperfection of the human knowledge, the only feasible option is to build this uncertainly into our mathematical models. Random variable is core concept and the Swiss knife that we use to deal with uncertainties mathematically.

Informally, a random variable differs from a normal variable as it is "random".

> A random variable is a variable whose value is uncertain, but comes with probabilities.

A random variable, say X, is never associated with a certain value, such as $X = 1$, or $X = 2$. It could be any of these values, but with different probabilities, e.g. $P(X = 1) = 0.2$, $P(X = 2) = 0.4$.

Definition 2.1. Given an experiment with sample space S, a **random variable** is a function from the sample space S to the real numbers $\mathbb{R}$.

As an example, flipping a coin twice, let X be the number of heads. Then $X(\cdot)$ is a functions that maps events in $\{HH, HT, TH, TT\}$ into real numbers. In our case, the mapping goes like

$$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0.$$

X is therefore an encoding of events in the sample space into real numbers. We could, of course, have different encodings. Conder the random variable Y as the number of tails. Then we have $Y = 2 - X$.

$$Y(HH) = 0, Y(HT) = 1, Y(TH) = 2, Y(TT) = 2.$$

We could also define Z as the number heads in the 1st toss only. The encoding goes like

$$Z(HH) = 1, Z(HT) = 1, Z(TH) = 0, Z(TT) = 0.$$

We have listed three ways of "encoding" the same experiment as random variables. All of them are valid random variables, but they map the outcomes into different numbers. We can say that, a random variable is a numeric "summary" of an aspect of an experiment.

Remark. We usually use capital letters, such as X, Y, Z, to denote random variables. We use small letters, such as x, y, z, to denote specific values. $P(X = x)$ means the probability of X taking the value x. "$X = x$" is an event. In the example above, $X = 2$ corresponds to the event HH. Note that we don't write $P(X)$. It is meaningless if X takes no value.

Definition 2.2. Let X be a random variable. The **distribution** of X is the collection of all probabilities of the form $P(X \in C)$ for all sets C of real numbers such that $\{X \in C\}$ is an event.

A **distribution** specifies the probabilities associated with all values of a random variable. In the above example, the distribution of X is given by

$$P(X=0)=\frac{1}{4}, P(X=1)=\frac{1}{2}, P(X=2)=\frac{1}{4}.$$

The distribution of Y is given by

$$P(Y=0)=\frac{1}{4}, P(Y=1)=\frac{1}{2}, P(Y=2)=\frac{1}{4}.$$

The distribution of Z is given by

$$P(Z=0)=\frac{1}{2}, P(Z=1)=\frac{1}{2}.$$

You may have noted that the probabilities in a distribution always sums up to 1, as all possible events constitute the entire sample space.

Example 2.1. Roll two fair 6-sided dice. Let $T=X+Y$ be the total of the two rolls, where X and Y are the individual rolls. Find the distribution for T.

2.2 Discrete and continuous random variables

2.2.1 Discrete distributions

Definition 2.3. We say X is a **discrete random variable** if X can take only a finite number k of different values $x_1, \ldots, x_k$ or, at most, an infinite sequence of countable different values $x_1, x_2, \ldots$.

The finite or countably infinite set of values x such that $P(X=x)>0$ is called the **support** of X.

Definition 2.4. If a random variable X has a discrete distribution, the **probability mass function** (PMF, sometimes also known as **probability function**, or **frequency function**) of X is defined as the function p such that

$$p(x)=P(X=x)$$

where $p(x)\geq 0$ for all possible values of x and $\sum_{\text{all } x} p(x)=1$.

Remark. $p(x)$ differs from the probability function $P(\cdot)$. $p(x)$ is a real-valued function. We can manipulate it as normal real-valued functions. Some textbooks prefer to use $f(x)$. In this book, we use $p(x)$ to distinguish it from the probability density function for continuous random variables. Sometimes, it is convinient to add a subscript, $p_X(x)$, to specify this is the PMF for random variable X.

Remark. The PMF $p(x)$ of a random variable X must satisfy the following criteria:

- Nonnegative: $p(x) \geq 0$ for all possible values of x;
- Sums to 1: $\sum_{\text{all } x} p(x) = 1$.

There are different ways to represent a PMF. We can (1) list all the possible values and their associated probabilities; (2) write a formula for the PMF; or (3) visualize it in a graph. In our example of two coins, the PMF can be written as

$$p_X(x) = \begin{cases} \frac{1}{4} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1 \\ \frac{1}{4} & \text{if } x = 2 \end{cases}.$$

2.2.2 Continuous distributions

Definition 2.5. We say a random variable X has a **continuous distribution** if the possible values of X takes the form of a continuum.

Definition 2.6. For a continuous random variable X, the **probability density function** (PDF) of X is a real-valued function f such that

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

where $f(x) \geq 0$ for all x and $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Continuous random variables are usually measurements. Examples include height, weight, temperature, the amount of money and so on.

The probability of a continuous random variable is not defined at specific values. Instead, it is represented by **the area under a curve** of the PDF. The probability of observing any single value is equal to 0, since the number of values assumed by the random variable is infinite. For continuous distributions, $P(a < x < b) = P(a \leq x \leq b) = P(a \leq x < b) = P(a < x \leq b)$.

More on continuous distributions will be discussed in later chapters.

2.2.3 Cumulative distribution function

Unlike PMF or PDF, a cumulative distribution function can be defined for both discrete and continuous random variables.

Definition 2.7. The **cumulative distribution function** (CDF) of a random variable X is the function F given by $F(x) = P(X \leq x)$.

For discrete random variables, $F(x) = \sum_{k \leq x} p(k)$.

For continuous random variables, $F(x) = \int_{-\infty}^{x} f(t)dt$. We thus have $\frac{dF(x)}{dx} = f(x)$.

Like PMF and PDF, CDF gives the full distribution of a random variable. Given the CDF, we can figure out any probability distribution of the random variable. For example, $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$.

Theorem 2.1. Any CDF has the following properties:

- $P(X > x) = 1 - F(x)$
- $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$
- Increasing: if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
- Right-continuous: for any a, $F(a) = \lim_{x \to a+} F(x)$.
- $F(x) \to 0$ as $x \to -\infty$; $F(x) \to 1$ as $x \to +\infty$.

2.3 Practical examples

In this section, we showcase some examples of how we apply random variables to model real-world scenarios. Anything can be a random variable, the height of a persion, the number of students in a class, etc. We define something as a random variable not because it is random in nature, but because we don't have enough information give a definite answer.

Example 2.2 (Height of a person)**.** What is the average height of a man? While the height of an individual is a specific, measurable figure, asking a broad question like this leads us into the realm of probability and statistics. For instance, imagine you've just landed in a new country and are curious about the average height of its residents. This question cannot be answered definitively without further context; instead, it requires a probabilistic approach.

Let's define H as a random variable representing the height of a person in that country. This variable H is governed by an unknown distribution that reflects the varying heights within the population. To estimate this distribution, we can collect a sample of individuals from the population. By analyzing this sample, we can gain insights into the average height and the variability of heights in the population. This example underscores the use of random variables when we want to answer general inquires about the attributes of a population.

Example 2.3 (Average temperature)**.** If you are tasked with making a weather forecast — specifically, to determine the average temperature for a day in September — how would you approach this question? Given the constraints of limited resources, such as not having the means to establish climate stations globally or run complex climate models using supercomputers, your best option is to treat this problem as one involving a random variable.

Let's denote T as the random variable representing the average temperature in September. To answer the question, you would first analyze historical temperature data to estimate the distribution of T. By examining this historical distribution, you might discover that the average temperature typically falls between 20 and 25 degrees Celsius for most September days.

While this estimate is far from perfect, it represents the best inference you can make given the constraints. It is certainly more reliable than making a completely random guess. This example underscores the use of random variables when we our knowledge of a complex phenomenon is constrainted.

It's important to recognize, though, that the historical data you are using does not encapsulate the "true distribution" of temperatures; it serves only as an approximation. If certain temperature values are absent from historical records, this does not imply that their probabilities are zero. Those values may simply be missing from the data collection, highlighting the limitations of your dataset. Thus, while your estimate provides useful insights, it is essential to approach it with caution, acknowledging that the actual distribution may differ.

Example 2.4 (Stock returns)**.** Suppose you're considering investing in the stock market and want to identify which stock might yield the greatest return. Predicting stock returns is notoriously challenging; if it were easy, everyone would be wealthy. However, this doesn't mean we cannot make informed decisions. The financial market operates as an information marketplace — having more information gives you a competitive edge.

If you had perfect information about who would buy or sell which stock at specific times, you could potentially predict price movements with high accuracy. Unfortunately, in reality, we often face uncertainty. To navigate this uncertainty, we can model stock returns as random variables. For instance, let X_j represent the monthly return of stock j. By collecting and analyzing the distribution of X_j over the past few years, we can gain insights into the stock's historical performance.

However, relying solely on past returns to guide investment decisions is not a good strategy. One major reason is that historical performance does not guarantee future outcomes. Market conditions, company performance, and economic factors can change dramatically, making past returns an unreliable predictor.

In this example, we highlight random variable as a technique to model uncertainty, and also acknowledge the limitation of statistics. Statistics may or may not be useful without an understanding of the subject matter.

Summary. Let's summarize what we've covered regarding random variables.

1. A random variable serves as a numerical representation of a specific aspect of an experiment or a random phenomenon. It allows us to quantify outcomes in a meaningful way, enabling analysis and interpretation of the results. For example, in a coin toss, we might define a random variable to represent the number of heads observed in a series of flips.

2. We typically model situations as random variables because we often lack sufficient information to draw definitive conclusions. In these instances, probability provides a framework for making educated guesses about uncertain outcomes. It acts as a compromise, allowing us to express our uncertainty mathematically and make decisions based on incomplete information. This is particularly useful in fields like finance, economics, and social sciences, where uncertainties are inherent.

3. Generally, we do not have access to the true distribution of a random variable, which is why we rely on finite samples, often derived from historical records, to approximate this distribution. By analyzing past data, we can estimate the probabilities associated with different outcomes. However, it's important to note that these approximations are subject to sampling variability and may not capture the entire complexity of the underlying phenomenon. Thus, understanding the limitations of our data and the potential for biases is crucial when making inferences based on random variables.

2.4 Bernoulli distribution

We introduce some "named distributions" from now on. These distributions are named because they provide standardized models for common "patterns" of random processes.

Definition 2.8. A random variable X is said to have the **Bernoulli distribution**, denoted as $X\sim$ Bern(p), if X has only two possible values, 0 and 1, and $P(X=1)=p$, $P(X=0)=1-p$.

The PMF of a Bernoulli random variable X is given by

$$p_X(x)=\begin{cases} p & \text{if } x=1,\\ 1-p & \text{if } x=0.\end{cases}$$

This can also be expressed as

$$p_X(x)=p^x(1-p)^{1-x} \text{ for } x\in\{0,1\}\,.$$

Example 2.5. Flip a coin once. Let X be the number of heads up. Then $X\sim$ Bern(p). If the coin is fair, we have $p=0.5$.

The Bernoulli distribution is widely used because it provides a simple yet powerful framework for modeling binary outcomes, where events can be classified as success or failure (**Bernoulli trial**). This versatility allows it to be applied across a wide range of fields and scenarios.

One key reason for its popularity is that many real-world phenomena can be distilled into binary outcomes. For instance, in quality control, a product can either pass or fail inspection; in healthcare, a treatment may either be effective or ineffective; and in marketing, a consumer may either purchase a product or not. Because nearly any situation involving two possible outcomes can be framed in terms of success and failure, the Bernoulli distribution becomes a natural choice for analysis.

Usually, we apply an **indicator variable** for binary outcomes. An indicator variable assigns a value of 1 to represent the occurrence of a specific event (success) and a value of 0 to indicate that the event did not happen (failure). This binary representation allows us to convert any event into a random variable, which can then be analyzed with Berboulli distribution.

Bernoulli distribution serves as the foundation for more complex models, such as the binomial distribution, which deals with multiple independent trials. This hierarchical structure makes it easier to build upon and develop more sophisticated statistical methods. Its simplicity also facilitates calculations and interpretations, making it accessible for researchers and practitioners alike.

2.5 Binomial distribution

Definition 2.9. Suppose $X_1, X_2, \dots, X_n$ are independent and identical $\text{Bern}(p)$ distributions. Let X be the total number of successes of the n independent trials. That is, $X = X_1 + X_2 + \cdots + X_n$. Then X has the **Binomial distribution**, $X \sim \text{Bin}(n, p)$.

The probability mass function of X directly follows from the combination theory:

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

This is a valid PMF because, by the Binomial theorem, we have

$$\sum_{k=0}^{n} p_X(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1.$$

Example 2.6. In the previous example of tossing two coins, we compute the distribution of X by counting the equally likely outcomes in an event. We can get the same result by realizing it is a Binomial distribution. $X \sim \text{Bin}(2, 1/2)$.

Since each coin tossing is an independent Bernoulli trial. The probabilities come directly from the PMF.

$$\begin{aligned} P(X=0) = p_X(0) &= \binom{2}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^2 = \frac{1}{4}; \\ P(X=1) = p_X(1) &= \binom{2}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1 = \frac{1}{2}; \\ P(X=2) = p_X(2) &= \binom{2}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^0 = \frac{1}{4}. \end{aligned}$$

Utilizing the Binomial distribution also allows us to generalize the problem. Suppose we are tossing n coins, we want to find the probability of getting k heads. It is almost impossible to count all the possible outcomes, but the answer immediately follows from the Binomial PMF.

Example 2.7. The Binomial distribution is often used to model the probability that a certain number of "successes" occur during a certain number of trials. Here is an example. Suppose it is known that 5% of adults who take a certain medication experience negative side effects. We want to find the probability that a certain number of patients in a random sample of 100 will experience negative side effects. Let X be the number patients that experience negative side effects, it follows that $X \sim \text{Bin}(100, 0.05)$.

Example 2.8. Let $X \sim Bin(n,p)$ and $Y \sim Bin(m,p)$ be two independent Binomail random variables. Show that $X+Y \sim Bin(n+m,p)$.

Proof. By the definition of the Binomial distribution, X is the number of successes in n independent trials, and Y is the number of successes in m independent trials. Therefore, $X+Y$ is the number of successes in $n+m$ independent trials, which is exactly $Bin(n+m,p)$.

We can also prove it using indicator variables. $X = \sum_{i=1}^{n} X_i$ where $X_i \sim Bern(p)$; $Y = \sum_{j=1}^{m} Y_j$ where $Y_j \sim Bern(p)$. Therefore, $X+Y = \sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j = \sum_{k=1}^{n+m} Z_k$. Since X_i and Y_j are identical Bernoulli random variables, $Z_k = X_k$ for $k = 1, \ldots, n$; $Z_k = Y_{k-n}$ for $k = n+1, \ldots, n+m$.

Another way is to leverage the PMF:

$$\begin{aligned}
P(X+Y=k) &= \sum_{i+j=k} P(X=i)P(Y=j)\\
&= \sum_{i+j=k} \binom{n}{i} p^i(1-p)^{n-i} \binom{m}{j} p^j(1-p)^{m-j}\\
&= \sum_{i+j=k} \binom{n}{i}\binom{m}{j} p^{i+j}(1-p)^{m+n-i-j}\\
&= p^k(1-p)^{m+n-k}\sum_{i=0}^{k}\binom{n}{i}\binom{m}{k-i}\\
&= p^k(1-p)^{m+n-k}\binom{n+m}{k}.
\end{aligned}$$

The last step: $\binom{n+m}{k} = \sum_{i=0}^{k}\binom{n}{i}\binom{m}{k-i}$ is known as the Vandermonde's identity.

Example 2.9. Let's explore an example that appears to be Binomial but is, in fact, not a Binomial distribution. Given a 5-card hand. Find the distribution of the number of aces.

Solution. Let X be the number of aces. It is tempting to say $X \sim Bin(5, p)$. But this not correct. Because having one ace is NOT independent from having another ace. We need to use the classical approach:

$$P(X=k) = \frac{C_4^k C_{48}^{5-k}}{C_{52}^5}.$$

This example leads to a named distribution that is closed related to Binomial — Hypergeometric distribution.

2.6 Hypergeometric distribution

Suppose we have a box filled with w white and b black balls. We draw n balls out of the box with replacement. Let X be the number of white balls. Then $X \sim Bin(n, w/(w+b))$. Since the draws are independent Bernoulli trials, each with probability $w/(w+b)$ of success. If we instead sample without replacement, then the number of white balls follows a **Hypergeometric distribution**. We denote this by $X \sim \text{HGeom}(w, b, n)$.

Theorem 2.2. *If $X \sim HGeom(w, b, n)$, then the PMF of X is*

$$p_X(k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n-k \leq b$, and $p_X(k) = 0$ otherwise.

Example 2.10. Let's redo the ace-card exercise with Hypergeometric distribution. In a 5-card hand, the number of aces in the hand has the HGeom$(4, 48, 5)$ distribution, which can be seen by thinking of the aces as white balls and the non-aces as black balls. The probability of having exactly three aces is $\frac{\binom{4}{4}\binom{48}{2}}{\binom{52}{5}} = 0.0017\%$.

The Binomial and Hypergeometric distributions are often confused. Both are discrete distributions taking on integer values between 0 and n for some n, and both can be interpreted as the number of successes in n Bernoulli trials. However, a crucial part of the Binomial story is that the Bernoulli trials involved are independent. The Bernoulli trials in the Hypergeometric story are dependent, since the sampling is done without replacement.

2.7 Uniform distribution

Definition 2.10. Let $a \leq b$ be integers. Suppose that the value of a random variable X is equally likely to be each of the integers $a, \ldots, b$. Then we say that X has the **discrete uniform distribution** on the integers $a, \ldots, b$. We denote it as $X \sim DUnif(a, \ldots, b)$.

The PMF of $X \sim DUnif(a, \ldots, b)$ is given by

$$p(x) = \begin{cases} \frac{1}{b-a+1} & \text{for } x = a, \ldots, b \\ 0 & \text{otherwise} \end{cases}.$$

Example 2.11. Let X be a random number from 1,2,...,100. Then $X \sim DUnif(1, ..., 100)$. And $P(X = k) = 1/100$ for any $k = 1, ..., 100$.

The uniform distribution can be defined in discrete cases, but its continuous form is more well-known.

Definition 2.11. Let a and b be two real numbers such that $a < b$. Let X be a random variable such that $a \leq X \leq b$ and, for every subinterval interval of

$[a, b]$, the probability that X belongs to that subinterval is proportional to the length of that subinterval. Then we say X has the **uniform distribution** on the interval $[a, b]$. We denote it as $X \sim Unif(a, b)$.

The PDF of $X \sim Unif(a, b)$ is given by

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

We verify this is a valid PDF because

$$\int_a^b f(x)dx = \int_a^b \frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b dx = 1;$$

or the area of the rectangle surrounded by $x = a, x = b$ and $f(x) = \frac{1}{b-a}$ is 1.

2.8 Functions of random variables

Functions of random variables are also random variables. If X is a random variable, then X^2, e^X and $\sin(X)$ are also random variables.

Definition 2.12. For an experiment with sample space S, a random variable X, and a function $g : \mathbb{R} \to \mathbb{R}$. $g(X)$ is the random variable that maps s to $g(X(s))$ for all $s \in S$.

Theorem 2.3. *Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$. If $g(X)$ is a one-to-one function. Then the support of $g(X)$ is the set of all y such that $x = g^{-1}(y)$ is in the support of X. The PMF of $g(X)$ is*

$$P(g(X) = y) = P(g(X) = g(x)) = P(X = x).$$

Theorem 2.4. *Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$. Then the support of $g(X)$ is the set of all y such that $g(x) = y$ for at least one x in the support of X. The PMF of $g(X)$ is*

$$P(g(X) = y) = \sum_{x:g(x)=y} P(X = x).$$

Example 2.12. Let X be a discrete random variable with $p_X(k) = \frac{1}{5}$ for $k = -1, 0, 1, 2, 3$. Let $Y = 2|X|$. Find the range and PMF of Y.

Definition 2.13. Give an experiment with sample space S, if X, Y are random variables that map $s \in S$ to $X(s)$ and $Y(s)$, then $g(X, Y)$ is the random variable that maps s to $g(X(s), Y(s))$ for all $s \in S$.

Example 2.13. We roll two fair 6-sided dice. Let X be the number on the first die, and Y be the number on the second die. Find the distribution of $\max(X, Y)$.

Solution: We just show how to compute one case in the distribution, other cases are similar.

$$\begin{aligned} P(\max(X,Y)=5) &= P(X=5, Y\le 4) + P(X\le 4, Y=5) + P(X=5, Y=5) \\ &= 2P(X=5, Y\le 4) + 1/36 \\ &= 2(4/36) + 1/36 = 9/36. \end{aligned}$$

2.9 Independence of random variables

Definition 2.14. Random variables X and Y are **independent** if

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$$

for all $x, y \in \mathbb{R}$.

In the discrete case, this is equivalent to the condition

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all possible values of x, y.

Example 2.14. Rolling two fair dice, X is the number on the first die, Y is the number on the second die, then $X + Y$ is not independent of $X - Y$.

Solution: It suffices to show one counter-example that does not follow the multiplication rule.

$$P(X + Y = 12, X - Y = 1) = 0$$

since this is not possible. However,

$$P(X + Y = 12)P(X - Y = 1) = \frac{1}{36} \times \frac{5}{36}.$$

Definition 2.15. Random variables $X_1, \dots, X_n$ are **independent** if

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n)$$

for all $x_1, \dots, x_n \in \mathbb{R}$.

Comparing this to the criteria for independence for n events, in which we require independence to hold for every pair, triplet, quadruplet and so on. It might look strange at first this only requires one condition. But in fact, this is equally, if not more, demanding as the criteria for events. As we require this to hold for all values of $x_1, \dots, x_n$. This entails pairwise independence, as we can rule out irrelevant variables by setting it to infinity:

$$P(X_1 \leq x_1, X_2 \leq x_2, X_3 \leq \infty, \dots) = P(X_1 \leq x_1)P(X_2 \leq x_2)$$

since $P(X_i \leq \infty) = 1$.

Theorem 2.5. *If X and Y are independent, then any function of X is independent of any function of Y.*

Definition 2.16. If a given number of random variables are independent and have the same distribution, we call them **independent and identically distributed**, or **i.i.d** for short.

- Independent and identically distributed (X, Y independent die rolls)
- Independent and not identically distributed (X: die roll; Y: coin flip)
- Dependent and identically distributed (X: number of Heads; Y: number of Tails)
- Dependent and not identically distributed (X: economic growth; Y: presidential election)

Definition 2.17. Random variables X and Y are **conditionally independent** given Z if

$$P(X \leq x, Y \leq y | Z = z) = P(X \leq x | Z = z)P(Y \leq y | Z = z)$$

for all $x, y \in \mathbb{R}$ and all z in the support of Z.

For discrete random variables, the equivalent definition is to require

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z).$$

Definition 2.18. For any discrete random variable X and Z, the function of x for a fixed z:

$$p_{X|Z}(x|z) = P(X = x|Z = z)$$

is called the **conditional PMF** of X given $Z = z$.

Application: seller ratings*

This example involves multiple types of discrete distributions. The technique used to solve this problem aligns with Bayesian inference, which is beyond the scope of this course. However, it remains an interesting case. The procedure illustrates the process of statistical modeling: we begin with an assumption and a proposed statistical model, then update it with new data. Finally, we draw inferences based on the model, typically addressing the question we aim to answer. You are not required to understand everything in this example. Nonetheless, it helps to develop a mindset of statistical inference early in the study.

Suppose you are shopping a product online. There are three sellers with the following ratings:

- Seller 1: 100% positive out of 10 reviews
- Seller 2: 96% positive out of 50 reviews
- Seller 3: 93% positive out of 200 reviews

Which seller is likely to give the best service?

The problem is intriguing because it is obvious that higher ratings do not necessarily means higher satisfaction. We have to weight in the number of reviews. The more reviews, the more trustworthy the ratings are. Let $X_j^{(i)}$ be a random variable that means consumer j is satisfied with seller i, where $i \in \{1, 2, 3\}$. Assume $X_j^{(i)}$ follows a Bernoulli distribution:

$$X_j^{(i)} = \begin{cases} 1 & \text{satisfaction with probability } \theta_i \\ 0 & \text{otherwise} \end{cases}$$

where θ_i is an unknown parameter of seller i that captures their "genuine" satisfaction rate. We assume the consumers independently write their ratings. The overall positive rate of seller i is therefore $R_i = \frac{1}{n_i}\sum_j X_j^{(i)}$ where n_i is the total number of reviews. We want to infer the value of θ_i from their observed positive rate R_i. From now on we drop the seller index i to simply the notation since it is symmetric for all sellers.

Because we have no prior knowledge about θ. We assume that θ takes any value from $[0, 1]$ equally likely, i.e. $\theta \sim \text{Unif}(0, 1)$. Assuming each X_j is independent and identical, then

$$S = X_1 + X_2 + \cdots + X_n$$

follows the Binomial distribution with PMF:

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Our goal is to find: $p(\theta|k)$. Recall that the Bayes' rule allows us to invert the conditional probability:

$$p(\theta|k) = \frac{p(k|\theta)p(\theta)}{p(k)} = \frac{p(k|\theta)p(\theta)}{\int_{-\infty}^{\infty} p(k|\theta)p(\theta)d\theta}$$

Since $\theta \sim \text{Unif}(0, 1)$, we have

$$p(\theta) = \begin{cases} 1 & \text{if } \theta \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

We now focus on $\theta \in [0, 1]$, since the probability is 0 otherwise. Substitute in the PMF of the Binomial distribution,

$$p(\theta|k) = \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}}{\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta}$$

The hard part is to evaluate the integral. We state without proof (this is known as the Beta function, which we will prove in later chapters):

$$\int_0^1 \theta^k(1-\theta)^{n-k} = \frac{k!(n-k)!}{(n+1)!}$$

Therefore,

$$p(\theta|k) = \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}$$

Now suppose you are the next customer. The probability that you would be satisfied is

$$\begin{aligned} P(X_{n+1}=1|S=k) &= \int_0^1 P(x_{n+1}=1|\theta)p(\theta|k)d\theta \\ &= \int_0^1 \theta \times \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}d\theta \\ &= \frac{(n+1)!}{k!(n-k)!}\int_0^1 \theta^{k+1}(1-\theta)^{(n+1)-(k+1)}d\theta \\ &= \frac{(n+1)!}{k!(n-k)!} \times \frac{(k+1)!(n-k)!}{(n+2)!} \\ &= \frac{k+1}{n+2}. \end{aligned}$$

Now we substitute the ratings for the three sellers:

- Seller 1: $n = 10, k = 10$
- Seller 2: $n = 50, k = 48$
- Seller 3: $n = 200, k = 186$

The probabilities that you would be satisfied with each seller are: 92%, 94%, 93%. The result is known as the **Laplace's rule of succession**. The rule of thumb is, pretending we have too more reviews: one is positive, the other is negative. Compute the satisfaction rate as $\frac{k+1}{n+2}$.

How to choose a distribution?

The fact is, we never know the "true" distribution of a real-world problem. When building a probability model, the distribution is typically *assumed* based on the nature of the data and the problem at hand. This assumption is crucial because the probability distribution determines how the random variable behaves, including its likelihood of taking specific values. Typically, this process involves:

1. Choosing the distribution: Based on the characteristics of the real-world situation or data, you assume an appropriate probability distribution. For example, if you're modeling the number of successes in a fixed number of independent trials, you might assume a Binomial distribution.

2. Assumptions behind the distribution: Every distribution has underlying assumptions. For example, a Binomial distribution assumes independent trials with two possible outcomes (success/failure) and a constant probability of success.

3. Fitting the model: Once you assume a distribution, you use data to estimate parameters of the distribution (e.g., mean, variance, or rate parameters), which allows you to make probabilistic predictions and inferences.

It is important to stress that the data we have collected from real events does not directly reveal the **Data Generating Process (DGP)**, which is the true underlying process that produces the data. Instead, when we assume a distribution, we are essentially making a hypothesis about what that DGP might be. The actual relationship between the assumed distribution and the data is one of approximation and testing, rather than perfect correspondence.

The assumed distribution is a *theoretical model* that we believe could explain the underlying patterns in the data. The data is a finite set of observations, which is only a sample from the potential infinite population or DGP. The data is influenced by noise, randomness, and sample size, so it doesn't always clearly show the true DGP. When we assume a distribution, we're making an educated guess about the DGP based on the nature of the problem, properties of the data, and sometimes prior knowledge or experience.

Data alone, especially from a finite sample, does not directly tell us what the DGP is. Instead, we infer the DGP by fitting models to the data and assessing how well they describe it. Since data is inherently noisy and finite, different models may fit the data well, meaning that multiple distributions could seem plausible based on the data alone. That's why we use goodness-of-fit tests, residual analysis, and model comparison to narrow down our choices.

If the data pattern conflicts with the assumed distribution, it might suggest that the assumption be wrong, and we should revisit our model. However, some degree of mismatch can be due to sample noise, outliers, or oversimplification, and may not always mean the assumption is entirely incorrect.

The workflow of probability modeling

The above example is a good illustration of how we do probability modeling. Here we summarize it into several key steps.

1. Understanding of the problem and data exploration: The typical workflow of probability modeling begins with a clear understanding of the problem we are trying to solve. This involves identifying the objective of the model, determining which quantities or events need to be modeled as random variables. This also involves gathering relevant data, if available, or understanding the kind of data we will be working with.

2. Assumption of probability distribution: Based on the nature of the data and the problem at hand, choose a candidate distribution. For discrete data, this could be distributions like Bernoulli or Binomial. For continuous data, it might be Normal or Uniform distributions.

3. Parameter estimation: The candidate distribution usually involves unknown parameters. In most of the applications, we are interested in estimating these parameters. In our example, we update the parameter with the Bayes' rule. But there are other estimation methods available, such as Maximum Likelihood Estimation (MLE). Estimation quantifies the model and provides specific estimates based on the data.

4. Model fit and evaluation: We skip this step in our example. But normally, we need to evaluate how well the assumed distribution fits the data. This involves performing goodness-of-fit tests or graphical diagnostics. If the assumed distribution doesn't fit the data well, the model might need to be refined.

5. Simulation or inference: After refining the model, we can run simulations or make inferences. If the model is meant to simulate real-world processes, we can now generate new data based on the probability distribution and its parameters. We may also use the model to predict future outcomes or estimate probabilities of specific events.