Probability Theory and Statistics Notebook I

Zhang Ruiyang

Contents

Preface

Part I

Probability Theory

Chapter 1

Fundamentals

Chapter 2

Random Variables

2.1 Overview

Let us start this Chapter with a scenario. Assuming we are working for the town government and we have collected all the medical data of people in town. That will be a lot of information. Among them, some are similar and can be categorised together. Blood types, for example, can be grouped together. However, sorting the data is not enough and it will be great if we can learn something new from it. If we want to find the average blood type, although it may sound like a silly task, we cannot do that since the blood types are not numerical values. This issue can be solved if we convert these blood types into numbers and compute the average that way.

This scenario demonstrates why we need **random variables**. The information we have collected is the sample space. The process of assigning a real number to each possible outcomes is the defining of a random variable - blood type in the case of this scenario. Random variable, mathematically, is a real-valued function that maps the sample space to real numbers.

There are other associated properties of a random variable. First, random variables are either **discrete** or **continuous**. Also, we can find functions that represent the distribution of random variables - **density functions**, **mass functions**, and **probability functions**. Some more, we can get a rough idea of the random variables by computing things like **expectation**, **variance**, and **moment**. Moreover, condition and independence mentioned in the previous Chapter can be applied to random variables too.

Some textbooks like to separate the discussion of discrete random variables and continuous random variable into two Chapters. I, on the other hand, prefer to put them side by side since the underlying idea is the same.

2.2 Functions for Probabilities

As we have mentioned before, a random variable is a mapping. A mapping has domain and range. The range of a random variable determines whether it is discrete or continuous. If the range is finite or at most countably infinite, for example numbers 1 to 6 for the value of a dice roll, it is a **discrete random variable** (DRV). If the range takes uncountably many values, for example an interval from -1 to 1, then it is a **continuous random variable** (CRV).

A random variable has many possible values and each has a different chance of occurring. To describe the chances for each possible value of the random variable to be taken, we have **probability mass function** (PMF), and **probability density function** (PDF), for discrete case and continuous case respectively. They describe the probability of occurrence of each possible values. We also have **cumulative distribution function** (CDF). It describes the probability of occurrence within an interval of values. Sometimes, we will call it as the **probability function**.

2.2.1 PMF and PDF

For a discrete random variable X, its probability mass function is denoted by the function f_X. In particular, if x is any possible value of X, the probability mass of x, denoted by $f_X(x)$, is the probability of the event $\{X = x\}$. So we have $f_X(x) = P(\{X = x\})$. Normally, we will use a capital letter to represent a random variable and the corresponding lower case letter to represent a possible value of the random variable. A property of PMF is that $\sum_x f_X(x) = 1$, which follows from the normalisation axiom of probability.

For a continuous random variable, its probability density function is also denoted by f_X. Everything is the same as PMF except for the normalisation property. For a CRV, we use integration instead of summation as there are uncountably many possible values of X. So, we have $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

If we have a linear function of random variable X instead of plainly the variable, the PMF/PDF of it can be computed as well. If we have a linear function $aX + b$ of random variable X as a new variable Y, each possible outcome of this new random variable will have the sample probability mass/density as the corresponding outcome of X. For example, outcome $ax + b$ of Y will have the same probability as the outcome x of X.

So far, we have been working with one random variable, but there can be more. If we have two variables, the PMF / PDF will become **joint PMF**. For random variables X and Y, the joint PMF is defined by

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

for all pairs of numerical values (x, y) that X and Y can take. Joint PMF can be understood as the probability of the intersection of two events $X = x$ and $Y = y$.

Associated with the joint PMF, we have the **marginal PMF**s for each of the random variables. We have

$$f_X(x) = \sum_y f_{X,Y}(x, y)$$

and

$$f_Y(y) = \sum_x f_{X,Y}(x, y).$$

Basically, the marginal PMF is trying to look at the probability of only one of the two random variables.

In Chapter 1, we have talked about conditional probability and how that is a law of probability too. So, it is only natural for us to see how we can incorporate conditional probability into PMFs.

The conditional PMF of X given an event A with positive probability is defined by

$$f_{X|A}(x) = P(X = x|A)$$

since $X = x$ is an event and A is its condition in this case. This can be understood easily if we are familiar with conditional probabilities. A property of conditional PMF is that is satisfies normalisation, meaning

$$\sum_x f_{X|A}(x) = 1.$$

Multiplication rule for conditional probability can be encompassed using conditional PMF. The conditional PMF of X given $Y = y$ is related to the joint PMF by

$$f_{X,Y}(x, y) = f_Y(y) f_{X|Y}(x|y).$$

This will become obvious once we see the original statement of multiplication rule $P(A \cap B) = P(B)P(A|B)$.

If we combine the definition of marginal PMF and that of multiplication rule for PMFs, we will have the following statement

$$f_X(x) = \sum_y f_Y(y) f_{X|Y}(x|y)$$

which tells us that we can calculate the marginal PMF using conditional PMF of X given Y.

As before, conditional probability tells us about the partial information an event contributes to another. If there is no partial information, then these two events are independent events. The same logic can be used here for conditional PMFs.

When we have

$$f_{X|A}(x) = f_X(x)$$

for all x, we will say X is independent of the event A. This means A is independent to all events $\{X = x\}$.

For two random variables X and Y, they will be independent for for all possible pairs (x, y), the events $\{X = x\}$ and $\{Y = y\}$ are independent, or we can have the equality

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

for all x, y.

2.2.2 CDF

So far in this Chapter, our concepts are all divided into two categories - discrete and continuous. Mathematicians like generalisation and simplicity. So it is natural for us to find ways to combine these two, which is by **cumulative distribution function**, or CDF.

The CDF of a random variable X is denoted by F_X and provides the probability $P(X \leq x)$. In particular, for every x we have

$$F_X(x) = P(X \leq x) = \begin{cases} \sum_{k \leq x} f_X(k) \text{ in the discrete case} \\ \int_{-\infty}^{x} f_X(t)dt \text{ in the continuous case} \end{cases}$$

We can reverse the above process to get PMF/PDF from CDF. In the discrete case, if X takes integer values, we have

$$f_X(k) = F_X(k) - F_X(k-1)$$

for all integers k. For the continuous case, we have

$$f_X(k) = \frac{dF_X(x)}{dx}$$

when the CDF is a differentiable function.

There are several properties of a CDF that we should take note of. The CDF F_X of a random variable X is defined by

$$F_X(x) = P(X \leq x)$$

for all x, and it has the following properties:

- F_X is monotonically nondecreasing:

$$\text{if } x \leq y, \text{ then } F_X(x) \leq F_X(y)$$

- $F_X(x)$ tends to 0 as $x \to -\infty$, and to 1 as $x \to \infty$.

- F_X has a piecewise constant and staircase-like form when X is discrete.
- F_X has a continuously varying form when X is continuous.

Sometimes, it is useful to study the opposite question and ask how often the random variable is above a particular level. This is called the **complementary cumulative distribution function** or simply the **tail distribution**. It is defined as

$$\bar{F}_X(x) = P(X > x) = 1 - F_X(x).$$

2.3 Expectation, Variance, and Moments

When we want to study random variables, we would like to know some big-picture properties of it. For example, we want to know what is the most common value this random variable will take, or what might be the average value. Some of these properties - mainly expectation, variance, standard deviation and moments - will be studied in this section.

2.3.1 Expectation

We will start with expectation. To study a concept, it is always beneficial to learn about its history. This allows us to understand the motivation behind the discovery or the creation of a concept, which either brings us clarity or helps us understand the concept. I would like to call this approach as "Motivation-Driven Learning"[1] although I doubt I am the first one to raise such a concept. However, since this is my notebook, I get the freedom to write whatever I like.

The origin of expectation originated in the middle of the 17th century from the study of the so-called problem of points, which seeks to divide the stakes in a fair way between two players who have to end their game before it's properly finished. This problem had been debated for centuries, and many conflicting proposals and solutions had been suggested over the years.

When the problem was posed in 1654 to Blaise Pascal by French writer and amateur Mathematician Chevalier de Méré, Méré claimed that this problem couldn't be solved and that it showed just how flawed mathematics was when it came to its application to the real world. Pascal, being a Mathematician, was provoked and determined to solve the problem once and for all.

He began to discuss the problem in a now famous series of letters to Pierre de Fermat. Soon enough they both independently came up with a solution. They solved the problem in different computational ways but their results were identical because their computations were based on the same fundamental principle. The principle is that the value of a future

[1] Inspired by the "Data-Driven" trend in the technology industry

gain should be directly proportional to the chance of getting it. This principle seemed to have come naturally to both of them. They were very pleased by the fact that they had found essentially the same solution and this in turn made them absolutely convinced they had solved the problem conclusively; however, they did not publish their findings. They only informed a small circle of mutual scientific friends in Paris about it.

Three years later, in 1657, a Dutch mathematician Christiaan Huygens, who had just visited Paris, published a treatise *De ratiociniis in ludo aleæ* on probability theory. In this book he considered the problem of points and presented a solution based on the same principle as the solutions of Pascal and Fermat. Huygens also extended the concept of expectation by adding rules for how to calculate expectations in more complicated situations than the original problem (e.g., for three or more players). In this sense this book can be seen as the first successful attempt at laying down the foundations of the theory of probability.[2]

Story time is over. We should head back to the Maths.

The **expected value** , or **mean**, of a random variable is the weighted sum of its value using its PMF or PDF. It represents what the average value the random variable is.

When we have a discrete random variable X with a finite number of outcomes $x_1, x_2, \ldots, x_k$ with probabilities $f_1, f_2, \ldots, f_k$ respectively, its expectation $\mathrm{E}[X]$ will be

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i f_i.$$

The countably infinite version of this statement will be similar, thus it is omitted here.

When we have a continuous random variable X with PDF $f_X(x)$, its expectation will be

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

We should note that expected value is often denoted by μ.

To make explanation more convenient here, we will talk about the continuous case by default. The readers can fill in the gaps of the discrete case pretty easily as these two are essentially the same.

Sometimes, instead of plainly the random variable, we have a function of it. If we have a random variable X with given PDF f_X, a real-valued function $Y = g(x)$ is also a random variable. The expectation of Y, or the function $g(x)$, is

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

This is because the probability of x in X is identical to the probability of $g(x)$ in Y.

[2]This story is copied entirely from Wikipedia.

A special case is when $g(x) = ax + b$ where a and b are two real constant. This gives us the statement

$$\mathrm{E}[Y] = \mathrm{E}[aX + b] = a\mathrm{E}[X] + b.$$

We know this is true since a multiplication of the integrand will cause the same multiplication to the whole value, also $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Before we finish with expectation, it is important to know that a probability can be represented by an expectation. If we want to find the probability of event A, we can construct an indicator function $I_A(x)$ that satisfies

$$I_A = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ otherwise} \end{cases}$$

and we get

$$P(A) = \mathrm{E}[I_A].$$

2.3.2 Variance and Standard Deviation

Expectation tells us the average value, but it is not enough. We may also want to know how values deviate from the mean. This is helpful and tells us more about the nature of the data. For example, a set of data that is scattered may have the same mean with another set of data that clustered over two extremes. They are different yet the difference cannot be learned from mean. To help with that problem, we introduce variance and standard deviation.

Variance and standard deviation are two closely related concepts. Historically speaking, standard deviation is introduced before variance. Although that is the case, it is more natural to learn variance before standard deviation.

Variance measures how far a set of numbers are spread out from their average value. Mathematically speaking, it is the expectation of the distance between a random variable from its mean. We will use the square of the difference between a possible value of random variable and the mean s the distance. The formula of random variable X, denoted by $\mathrm{Var}(X)$, with mean μ will be

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

We can represent the variance using an integration, which is

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu^2)^2 f_X(x)dx.$$

We can replace the random variable with a linear function of it. For $Y = aX + b$, its variance is

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

since multiplication gets squared for variance and the constant term are cancelled out by the subtraction. Another way to look at it is that variance measures the deviation of data and it will not change if we shift all the data with the same degree.

When we have, not one, but two random variables, we would like to find out their relationship using **covariance**. Covariance is a measure of the joint variability of two random variables and it is positive when they tend to show similar behaviour.

To properly define it, given two random variables X and Y, the covariance of the two, denoted by $\mathrm{Cov}(X, Y)$, is

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

Notice that if X and Y are independent, $\mathrm{Cov}(X, Y) = 0$ as $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Some properties of covariance will be stated below. The proofs will be omitted as they are quite straight forward.

$$\begin{aligned}\mathrm{Cov}(X, X) &= \mathrm{Var}(X) \\ \mathrm{Cov}(X, Y) &= \mathrm{Cov}(Y, X) \\ \mathrm{Cov}(cX, Y) &= c\mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y + Z) &= \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)\end{aligned}$$

Standard deviation is a measure to tell us how far away a data point is from the mean. It is calculated by square rooting the variance, and it is often denoted by σ. For random variable X, we have $\sigma^2(X) = \mathrm{Var}(X)$.

2.3.3 Moment

Moments of a function are quantitative measures related to the shape of the function's graph. Here, we will be studying the moments of a probability distribution.

The definition of n-th **moment** of a real-valued continuous function $f(x)$ of a real variable about a value c is

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx.$$

When we take certain special values of n and c, we can get some very useful properties. There are three main kinds of moments - raw, central and standardised.

Raw moment is a moment of a probability distribution of a random variable about 0, meaning the c in the equation will be 0. So the n-th raw moment of random variable X is

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

The first raw moment is the expected value μ, since the statement is identical to that of expectation.

Central moment is a moment of a probability distribution of a random variable about the random variable's mean, meaning the c in the equation will be set to μ. So the n-th central moment of random variable X is

$$\mu_n = \int_{-\infty}^{\infty} (x - \mu)^n f(x) dx.$$

The zeroth central moment μ_0 is 1, due to normalisation axiom of probability. The first central moment is 0. The second central moment is variance.

Standardised moment of a probability distribution is a moment that is normalised, typically a central moment divided by an expression of the standard deviation. So the n-th standardised moment of random variable X is

$$\tilde{\mu}_n = \frac{\mu_n}{\sigma^n} = \frac{\int_{-\infty}^{\infty} (x - \mu)^n f(x) dx}{(\sqrt{\mathrm{E}[(X - \mu)^2]})^n}.$$

The first standardised moment is 0. The second standardised moment is 1. The third standardised moment is **skewness**. The fourth standardised moment is **kurtosis**.

Skewness is a measure of asymmetry of the probability distribution. A positive skew means the right tail is longer. A negative skew means the left tail is longer.

Kurtosis[3], originated by Karl Pearson[4] from Pearson distribution, is a measure of the "tailedness" of the probability distribution of a real-valued random variable[5].

It will be a cool idea to combine all the possible kinds of moments, and Taylor series of e^{tX} comes into rescue. We have

$$e^{tX} = 1 + tX + \frac{t^2 X^2}{2!} + \frac{t^3 X^3}{3!} + \cdots.$$

If we take the expectation to both sides of the above equation, we will call this equation as the **moment generating function**, or MGF, of random variable X and get

$$M_X(t) = \mathrm{E}[e^{tX}] = 1 + t\mathrm{E}[X] + \frac{t^2 \mathrm{E}[X^2]}{2!} + \frac{t^3 \mathrm{E}[X^3]}{3!} + \cdots$$

for $t \in \mathbb{R}$, which contains all the n-th moments.

[3] from Greek *kurtos*.

[4] founder of the first Statistics department in the world, at UCL - where I am attending for my undergraduate degree.

[5] I do not fully understand this at this time. I may come back and explain more.

–

A useful properties of MGF is that we can take a certain derivative of it to get a certain moment, which is

$$\mathrm{E}[X^n] = \frac{d^n M_X(t)}{dt^n}\Big|_{t=0} = M_X^{(n)}(0)$$

where $M_X^{(n)}$ is the n-th derivative of the MGF.

–

Another thing to take note of regarding MGF is that two random variables with the same MGF will have the same distribution function. This means, for random variable X and Y, if

$$M_X(t) = M_Y(t)$$

for all t on some interval for MGF to exist, we must have

$$F_X(a) = F_Y(a)$$

for all $a \in \mathbb{R}$.

This is due to the fact that the probability distribution of a random variable is uniquely determined by its generating function, as we could have probably guessed from the construction of MGF using moments.

We can expand this. Suppose that $X_1, \ldots, X_k$ are independent random variables, then

$$\begin{aligned} M_{X_1+\cdots+X_k}(t) &= \mathrm{E}[e^{(X_1+\cdots+X_k)t}] \\ &= \mathrm{E}[e^{X_1 t} \cdots e^{X_k t}] \\ &= \mathrm{E}[e^{X_1 t}] \cdots \mathrm{E}[e^{X_k t}] \\ &= M_{X_1}(t) \cdots M_{X_k}(t). \end{aligned}$$