Probability Theory and Statistics Notebook I

Zhang Ruiyang

Contents

Preface **ii**

I Probability Theory **1**

1 Fundamentals **2**

2 Random Variables **3**

3 Common Distributions **4**

3.1 Bernoulli Trial 4

3.2 Binomial Distribution 5

3.3 Poisson Distribution 6

3.4 Normal Distribution 9

3.5 Other Distributions 15

3.6 Summary 18

Preface

Part I

Probability Theory

Chapter 1

Fundamentals

Chapter 2

Random Variables

Chapter 3

Common Distributions

In the previous Chapter, we have talked a lot about properties of random variables. They may sound boring and dull if we do not have solid examples of what random variables can be. In this Chapter, we will be discussing some of the more common types of random variables and distributions.

3.1 Bernoulli Trial

A **Bernoulli trial**, named after Jacob Bernoulli[1], is a random experiment with exactly two possible outcomes and constant probability.

An example of a Bernoulli trial is a coin tossing, with only two possible outcomes head and tail. Since there are finite many (2, to be more precise) possible outcomes, the distribution is discrete.

Normally, we will denote the 'success' outcome as 1 and 'failure' as 0. Also, we usually denote the chance of 'success' as p, the other being $1-p$ for obvious reason. This variable p is the **parameter** of the distribution. A parameter[2] is a characteristic that can help defining a distribution or a function.

Knowing the nature of the distribution, we would want to find out the PMF of it. A random variable that follows Bernoulli trials is called Bernoulli random variable. Thus, for a Bernoulli random variable X, we have

$$P(X = x) = p^x(1-p)^{1-x}$$

[1]The Bernoulli family has many great Mathematicians and you will see a lot of things named after Bernoulli in Mathematics.

[2]Parameter is an important concept in Statistics. A large proportion of Statistics deals with finding the parameter(s) of a set of data. Frequentists and Bayesian Statisticians understand parameter differently too. We also have a branch of Statistics called non-parametric which is not based solely on parametrized families of probability distributions.

where $x = 0, 1$ and $0 \leq p \leq 1$.

When $x = 1$, $P(X = 1) = p(1-p)^0 = p$. When $x = 0$, $P(X = 0) = p^0(1-p)^1 = 1-p$.

Having the PMF will help us obtain the expectation and variance. For the same random variable X, we have

$$\mathrm{E}[X] = 0 \times (1-p) + 1 \times p = p$$

and

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = p \times 1^2 - p^2 = p(1-p).$$

The last property we will be discussing for a Bernoulli trial is its MGF. Since we know

$$P(X = x) = \begin{cases} p \text{ if } x = 1 \\ 1-p \text{ if } x = 0 \\ 0 \text{ otherwise} \end{cases}$$

and

$$M_X(t) = \mathrm{E}[e^{tX}] = \sum_n P(X = n)e^{tn},$$

we get

$$\begin{aligned} M_X(t) &= P(X = 0)e^0 + P(X = 1)e^t \\ &= (1-p) + pe^t. \end{aligned}$$

3.2 Binomial Distribution

Bernoulli trials deals with single-time experiment with two possible outcomes. This is quite limited, and we would like to go further from this model. The most immediate thing we can change is by changing the number of times of the experiment from 1 to n. This gives us **binomial distribution**, discovered by Jacob Bernoulli in his work entitled *Ars Conjectandi*.

Bernoulli trials have a single parameter, the probability of success p, and binomial distribution has one more parameter, the number of trials n. Everything else, like 0 and 1 for success and failure, p for success rate and $1-p$ for failure rate, are defined identically for a typical binomial distribution. We will sometimes denote such a distribution as $\mathrm{Binom}(n, p)$. This type of distribution, like a Bernoulli trial, is discrete.

The PMF of a binomial distribution can be found by referring to binomial theorem. Binomial theorem states that

$$(x+y)^n = \binom{n}{0}x^n y^0 + \binom{n}{1}x^{n-1}y^1 + \cdots + \binom{n}{n-1}x^1 y^{n-1} + \binom{n}{n}x^0 y^n = \sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}$$

where the value of n can be any real number.

We can substitute x and y with p and $1-p$, which change the statement into

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

This way, we get the PMF of a binomial distribution with random variable as X, n trials and success rate as p as

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

Having the PMF will help us to obtain the expectation and variance. For the same random variable X, we have

$$\mathrm{E}[X] = \sum_{i=1}^{n} \mathrm{E}[X_i] = np$$

and

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p).$$

where each X_i is an identical and independent Bernoulli random variable.

The fact that a binomial distribution is a compilation of several Bernoulli trials help us to get its MGF. Since

$$M_X(t) = \mathrm{E}[e^{tX}] = \mathrm{E}[e^{t(nX_i)}],$$

we get

$$\begin{aligned} M_X(t) &= (\mathrm{E}[e^{t(X_i)}])^n \\ &= ((1-p) + pe^t)^n. \end{aligned}$$

3.3 Poisson Distribution

Binomial distribution involves n Bernoulli trials each with success rate p. These are useful in many cases, but it can definitely be extended. The probability of an event occurring does not always have to be constant throughout. For example, the probability of crime occurring in a neighbourhood may be constant at first, but it will not have the same probability as the total number of crime increases. Many factors and reasons may cause this phenomenon to occur but we will not try to figure out why in this notebook. What we do see here is that there are definitely examples of things that have less and less chances of happening when the number of times increase. These examples will be modelled by the **Poisson distribution**, named after the French Mathematician Siméon Poisson[3].

[3]Fun fact, 'poisson' means fish in French. It has nothing to do with the content here of course.

Given $\text{Binom}(n, p)$ with random variable X, we have PMF

$$\text{Binom}_k(n, p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $0 \leq k \leq n$.

Since we know the probability will decrease as number of trials increase, we might as well set the probability of success as α/n where α is a constant. Making that change to the statement of PMF, we get

$$\text{Binom}_k(n, \alpha/n) = \binom{n}{k} (\frac{\alpha}{n})^k (1 - \frac{\alpha}{n})^{n-k}$$

This is not that nice of an equation. We need to find ways to make it more elegant. Let us first set n to a very big number and approaches infinity. Looking at the case when $k = 0$, we get

$$\lim_{n\to\infty} \text{Binom}_0(n) = \lim_{n\to\infty} (1 - \frac{\alpha}{n})^n.$$

Note that the Taylor series $\ln(1-x) = -\sum_{n=1}^{\infty} x^n/n$, we will get

$$\ln((1 - \frac{\alpha}{n})^n) = n \ln(1 - \frac{\alpha}{n}) = n(-\sum_{n=1}^{\infty} \frac{x^n}{n}) = -\alpha - \frac{\alpha^2}{2n} - \cdots .$$

When $n \to \infty$, the terms on the right, starting from the second one, will be equal to 0. By making both sides of the equation the power to e, we have

$$\lim_{n\to\infty} \text{Binom}_0(n) = \lim_{n\to\infty} (1 - \frac{\alpha}{n})^n = e^{-\alpha}.$$

This gives us the one value of the full PMF. This is certainly not enough, so we try to find the other ones. We happen to realise that two consecutive probabilities Binom_k have a ratio, which can be found by the following statement:

$$\lim_{n\to\infty} \frac{\text{Binom}_{k+1}(n)}{\text{Binom}_k(n)} = \lim_{n\to\infty} \frac{n-k}{k+1} (\frac{\alpha}{n})(1 - \frac{\alpha}{n})^{-1} = \frac{\alpha}{k+1}.$$

That is all we need to find the PMF of a Poisson distribution. We have

$$\lim_{n\to\infty} \text{Binom}_k(n) = \lim_{n\to\infty} \text{Binom}_0(n) \cdot \prod_{j=0}^{k} (\frac{\alpha}{j+1}) = e^{-\alpha} \cdot \frac{\alpha^k}{1 \cdot 2 \cdots k} = \frac{e^{-\alpha}}{k!} \alpha^k = \pi_k(\alpha)$$

where $\pi_k(\alpha)$ represents the probability of the event to happen for k number of times with the average number of event occurrence as α for a Poisson Distribution.

Having the PMF will help us to obtain the expectation and variance. For the same random variable X, we have

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{k\geq 0} k\frac{1}{k!}\lambda^k e^{-k} \\
&= \lambda e^{-\lambda}\sum_{k\geq 1}\frac{1}{(k-1)!}\lambda^{k-1} \\
&= \lambda e^{-\lambda}\sum_{k-1\geq 0}\frac{1}{(k-1)!}\lambda^{k-1} \\
&= \lambda e^{-\lambda}e^{\lambda} \\
&= \lambda
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\
&= \sum_{k\geq 0} k^2\frac{1}{k!}\lambda^k e^{-k} - \lambda^2 \\
&= \lambda e^{-\lambda}\sum_{k\geq 1} k\frac{1}{(k-1)!}\lambda^{k-1} - \lambda^2 \\
&= \lambda e^{-\lambda}\Big(\sum_{k\geq 1}(k-1)\frac{1}{(k-1)!}\lambda^{k-1} + \sum_{k\geq 1}\frac{1}{(k-1)!}\lambda^{k-1}\Big) - \lambda^2 \\
&= \lambda e^{-\lambda}\Big(\lambda\sum_{k\geq 2}(k-2)\frac{1}{(k-2)!}\lambda^{k-1} + e^{\lambda}\Big) - \lambda^2 \\
&= \lambda e^{-\lambda}(\lambda e^{\lambda} + e^{\lambda}) - \lambda^2 \\
&= \lambda^2 + \lambda - \lambda^2 \\
&= \lambda.
\end{aligned}$$

Knowing the PMF can help us find the MGF. Since

$$M_X(t) = \mathrm{E}[e^{tX}] = \sum_{n=0}^{\infty} e^{tn}P(X=n),$$

we get

$$\begin{aligned}
M_X(t) &= \sum_{n=0}^{\infty} e^{tn}\frac{e^{-\alpha}}{n!}\alpha^n \\
&= e^{-\alpha}\sum_{n=0}^{\infty} e^{tn}\frac{(\lambda e^t)^n}{n!} \\
&= e^{-\alpha}e^{\alpha e^t} \\
&= e^{\alpha(e^t-1)}
\end{aligned}$$

We will sometimes use λ to replace α in the above equation.

3.4 Normal Distribution

Earlier on, we have learned about the statement of the PMF of a binomial distribution. The PMF contains a combinatoric operation $\binom{n}{k}$. This is hard to compute when n gets bigger, especially at times when modern computing tools were not available. This leads Mathematicians to find ways to approximate the statement. The Mathematician who made progress on that is de Moivre.

At the time of de Moivre, another Mathematician James Stirling came up with a formula that approximates factorial. The **Stirling formula** says that

$$n! \approx \sqrt{2\pi n}(\frac{n}{e})^n.$$

This formula was, in fact, first thought off by de Moivre but was completed by Stirling and then used by de Moivre in his work.

Another related question to the approximation of factorial is that of finding the probability of binomial random variable X to take values within a certain distance d from the mean, in Mathematical expression it is

$$P(|X - np| \leq d).$$

I will present de Moivre's approach of solving these problems below.

Let us first consider a binomial distribution $b(n, p)$ where n is even and p is $\frac{1}{2}$. The PMF of that is $b(n, 1/2, i) = \binom{n}{i}(1/2)^n$. We will denote $b(n, 1/2, i)$ by $b(i)$ here.

Using Stirling's formula, we have

$$\begin{aligned}
b(\frac{n}{2}) &= \binom{n}{\frac{n}{2}} \cdot (\frac{1}{2})^n \\
&= \frac{n!}{(\frac{n}{2})!(\frac{n}{2})!} \cdot (\frac{1}{2})^n \\
&\approx \frac{\sqrt{2\pi n}(\frac{n}{e})^n}{[(\sqrt{\pi n}(\frac{n/2}{e})^{n/2})]^2} \cdot (\frac{1}{2})^n \\
&= \frac{\sqrt{2\pi n}(\frac{n}{e})^n}{\pi n(\frac{n/2}{e})^n} \cdot (\frac{1}{2})^n \\
&= \frac{\sqrt{n} \cdot n^n}{(n/2)(n/2)^n} \times 2^{-n} \times \frac{1}{\sqrt{2\pi}} \\
&= \frac{n^{n+1/2-1-n}}{2^{-n-1}} \times 2^{-n} \times \frac{1}{\sqrt{2\pi}} \\
&= 2 \times n^{-\frac{1}{2}} \times \frac{1}{\sqrt{2\pi}} \\
&= \sqrt{\frac{2}{n\pi}}.
\end{aligned}$$

Knowing the probability at $n/2$ is not enough. We would also want to know the probability at some distance away from the center, which is $n/2 + d$. It is slightly troublesome to find an approximation for that directly, but we can find the ratio of that probability and the one we have computed just now.

Before doing the full computation, we notice

$$\begin{aligned}
(\frac{n}{2})! &\approx \sqrt{\pi n} \cdot (\frac{n}{2e})^{n/2} \\
(\frac{n}{2}+d)! &\approx \sqrt{\pi n + 2d\pi} \cdot (\frac{\frac{n}{2}+d}{e})^{n/2+d} \\
(\frac{n}{2}-d)! &\approx \sqrt{\pi n - 2d\pi} \cdot (\frac{\frac{n}{2}-d}{e})^{n/2-d} \\
[(\frac{n}{2})!]^2 &\approx \pi n \cdot (\frac{n}{2e})^n \\
(\frac{n}{2}+d)! \cdot (\frac{n}{2}-d)! &\approx \sqrt{\pi^2 n^2 - 4d^2\pi^2} \cdot e^{-n} \cdot (\frac{n}{2}+d)^{n/2+d} (\frac{n}{2}-d)^{n/2-d} \\
&= (n/2+d)^{n/2+d+1/2} \cdot (n/2-d)^{n/2-d+1/2} \cdot 2\pi e^{-n}.
\end{aligned}$$

So, the ratio is

$$\begin{aligned}
\text{Ratio} &\approx n^{n+1} \cdot 2^{-n-1} \cdot (n/2+d)^{-n/2-d-1/2} \cdot (n/2-d)^{-n/2+d-1/2} \\
&= (\frac{n}{2})^{n+1} \cdot (n/2+d)^{-n/2-d-1/2} \cdot 2^{-n} \cdot (n/2-d)^{-n/2+d-1/2} \\
&= (1+\frac{2d}{n})^{-n/2-d-1/2} \cdot (1-\frac{2d}{n})^{-n/2+d-1/2} \\
&= (1-\frac{2d}{n})^{2d} \cdot (1-\frac{4d^2}{n^2})^{-n/2+d-1/2} \\
&\approx (1-\frac{2d}{n})^{2d} \cdot [(1-\frac{1}{n^2/(4d^2)})^{n^2/(4d^2)}]^{-2d^2/n} \\
&= (1-\frac{2d}{n})^{2d} \cdot e^{-2d^2/n} \\
&\approx e^{-2d^2/n}.
\end{aligned}$$

Thus, we get

$$b(\frac{n}{2}+d) \approx \frac{2}{\sqrt{2n\pi}} \cdot e^{-2d^2/n}.$$

Using what we have obtained so far, it is easy to get

$$\begin{aligned} P(|\ \frac{X}{n} - \frac{1}{2}\ | \leq \frac{c}{\sqrt{n}}) &= \sum_{-c\sqrt{n} \leq i \leq c\sqrt{n}} b(\frac{n}{2} + i) \\ &\approx \sum_{-c\sqrt{n} \leq i \leq c\sqrt{n}} \frac{2}{\sqrt{2n\pi}} \cdot e^{-2i^2/n} \\ &= \sum_{-2c \leq \frac{2i}{\sqrt{n}} \leq 2c} \frac{1}{\sqrt{2\pi}} \cdot \frac{2}{\sqrt{n}} \cdot e^{-\frac{1}{2}(\frac{2i}{\sqrt{n}})^2} \\ &\approx \int_{-2c}^{2c} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \end{aligned}$$

which we call as the PDF of a normal distribution with mean 0 and variance 1. We get the conclusion that the limit of a binomial distribution as n increases is a **normal distribution**.

This is only for the special case of $p = 1/2$, and de Moivre did some more work for the cases when that assumption is not there. Later on, Pierre-Simon Laplace worked on that problem further and generalised that approximation to any p. This is the first time Mathematicians obtain the formula of a normal distribution PDF, as a limit case of binomial distribution. This theorem is therefore named after these two contributors, as **de Moivre-Laplace central limit theorem**. The statement is the following:

For any two constants a and b, $-infty < a < b < +\infty$, we have

$$\lim_{n \to \infty} P(a < \frac{X_n - np}{\sqrt{np(1-p)}} < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

We will be studying more of limits and central limit theorems in the next Chapter when we talk about limit theorems.

–

De Moivre's work stops here. If you look up the proper statement of PDF of a normal distribution, you will realise that it is close but not identical to the one above. This is because de Moivre's work did not discover this gem of probability fully, which is also why the distribution is never named after him.

The next big achievement in the study of normal distribution was made from studying the error distribution in the 18th century when scientists tried to get rid of the observation errors in astronomical data. Observation errors in the data had been hard to deal with, and scientists commonly used arithmetic mean to get rid of the error and improve the accuracy. Although everyone used arithmetic mean, there was no proof of why that method is so effective at that time. Also, since errors have their own distribution, is there any correlation between the error distribution and arithmetic mean?

The first two major work on this problem were completed by Thomas Simpson and Laplace. They made several contributions but they failed to provide a very good solution.

And there we have the entrance of Gauss. Gauss' involvement in Astronomy began by correctly predicting the time and place of the occurrence of Ceres in 1801. He did not publish his working right afterwards the prediction, probably because he was not too confident with his method at that moment of time. As he formalised it a few years later in 1809, he published it and it was a derivation of normal distribution from the distribution of error using least square method.[4] I will briefly present his work below. However, some of the concepts have yet to be properly defined and discussed at this time of the notebook, so the readers can come back to this at a later time.

The assumption Gauss made was that since the arithmetic mean was such a great tool, it must be the same as the maximum likelihood estimator, MLE, of the error distribution.

Let as have the real value of the parameter as θ and $x_1, x_2, \ldots, x_n$ be n independent measurements, each with an error $e_i = x_i - \theta$. Assume the density function of the error e_i is $f(e)$, the joint probability of the n errors is denoted by

$$\begin{aligned} L(\theta) &= L(\theta; x_1, \cdots, x_n) \\ &= f(e_1) \cdots f(e_n) \\ &= f(x_1 - \theta) \cdots f(x_n - \theta). \end{aligned}$$

To find the MLE, let

$$\frac{d \log L(\theta)}{d\theta} = 0.$$

We get

$$\sum_{i=1}^{n} \frac{f'(x_i - \theta)}{f(x_i - \theta)} = 0.$$

If $g(x) = \frac{f'(x)}{f(x)}$, the above equation becomes

$$\sum_{i=1}^{n} g(x_i - \theta) = 0.$$

Based on the assumption, the value of θ should be the arithmetic mean $\bar{x}$. This means we will have

$$\sum_{i=1}^{n} g(x_i - \bar{x}) = 0.$$

[4] Gauss' derivation of normal distribution one of the many possible derivations. We only present the first one here. The interested readers can refer to Chapter 7 of the book "Probability Theory: The Logic of Science" by E. T. Jaynes which present more possible derivations.

When $n = 2$, we have $g(x_1-\bar{x})+g(x_2-\bar{x}) = 0$. Since $x_1-\bar{x} = -(x_2-\bar{x})$ as $\bar{x} = (x_1+x_2)/2$, we get

$$g(x) = g(-x).$$

When $n = m + 1$, $x_1 = \cdots = x_m = -x$, $x_{m+1} = mx$ and $\bar{x} = 0$, we get $\sum_{i=1}^{n} g(x_i - \bar{x}) = mg(-x) + g(mx) = 0$, which implies

$$mg(x) = g(mx).$$

The only function that satisfies the above two properties is that of $g(x) = cx$ for some constant c. This gives us

$$f(x) = \sqrt{\frac{\alpha}{2\pi}} e^{-\frac{1}{2}\alpha(x-\theta)^2}$$

where α is a positive constant.

This is very close to the definition of a normal distribution PDF, which is why the normal distribution is sometimes known as Gaussian distribution to recognise of Gauss' work. Another name of normal distribution, although seldom used nowadays, is Laplace distribution as Laplace contributed to the work too and expanded it to central limit theorem. Back then, French Mathematicians addressed the distribution as Laplace distribution as Laplace is French, and German Mathematicians addressed it as Gaussian distribution as Gauss is German. Mathematicians from other countries called it, by the neutral name, Laplace-Gaussian distribution.

The confusing naming was settled by the recommendation of Henri Poincaré and the publicising of Statistician Karl Pearson[5], which is the reason why we normally call it as "normal distribution".

–

The general formula of the PDF of a normal distribution with mean μ and variance σ^2 is

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

The expectation and variance need not to be found since they are required parameters in order to define the normal distribution.

The MGF can be found using the PDF. We will be using the standardised normal distribution which is the one with mean 0 and variance 1. The random variable is usually

[5]His given name was originally "Carl", but he changed the spelling to "Karl" due to his fascination for socialism and work of Karl Marx.

denoted by Z if it is normalised.

$$\begin{aligned}
M_X(t) &= \mathrm{E}[e^{zt}] \\
&= \int_{-\infty}^{\infty} e^{zt} f(z) dz \\
&= \int_{-\infty}^{\infty} e^{zt} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \int_{-\infty}^{\infty} e^{zt - z^2/2} \frac{1}{\sqrt{2\pi}} dz \\
&= \int_{-\infty}^{\infty} e^{-(z-t)^2/2} e^{t^2/2} \frac{1}{\sqrt{2\pi}} dz \\
&= e^{t^2/2} \int_{-\infty}^{\infty} e^{-(z-t)^2/2} \frac{1}{\sqrt{2\pi}} dz \\
&= e^{t^2/2}
\end{aligned}$$

as the expression under the integral is the PDF of a normal distribution with mean t and variance 1.

For the MGF of a general normal distribution $N(\mu, \sigma^2)$, we have

$$\begin{aligned}
M_X(t) &= \mathrm{E}[e^{xt}] \\
&= \int_{-\infty}^{\infty} e^{xt} f(x) dz \\
&= \int_{-\infty}^{\infty} e^{xt} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} dx
\end{aligned}$$

Define $z = \frac{x-\mu}{\sigma}$, which implies $x = z\sigma + \mu$.

Substituting it back the equation, we get

$$\begin{aligned}
M_X(t) &= e^{\mu t} \int_{-\infty}^{\infty} e^{z\sigma t - z^2/2} \frac{1}{\sqrt{2\pi\sigma^2}} \left| \frac{dx}{dz} \right| dz \\
&= e^{\mu t} \int_{-\infty}^{\infty} e^{z\sigma t - z^2/2} \frac{1}{\sqrt{2\pi}} dz \\
&= e^{\mu t + (\sigma^2 t^2)/2}
\end{aligned}$$

since $dx/dz = \sigma$ and by replacing t with σt of the final line of the MGF derivation of standardised normal distribution above.

–

An useful property of normal distribution is that when we have multiple independent normal random variables, the summation of them become a new normal random variable.

Let X_j be independent random variables with distribution $N(\mu_j, {\sigma_j}^2)$. Then, $X_1 + \cdots + X_n$ has the normal distribution $N(\sum_{j=1}^n \mu_j, \sum_{j=1}^n {\sigma_j}^2)$.

This property can be shown using MGF. When $n = 2$, we have

$$\begin{aligned} M_{X_1+X_2}(t) &= M_{X_1}(t) \cdot M_{X_2}(t) \\ &= e^{\mu_1 t + ({\sigma_1}^2 t^2)/2} \cdot e^{\mu_2 t + ({\sigma_2}^2 t^2)/2} \\ &= e^{(\mu_1+\mu_2)t + ({\sigma_1}^2 + {\sigma_2}^2)t^2/2}, \end{aligned}$$

which is the moment generating function of $N(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$. This process with induction can prove the statement above.

3.5 Other Distributions

There are many types of distributions other than the ones we have already mentioned earlier. We will be going to cover some of them here and highlight their relationships.

–

Sometimes, an event will happen after a period of time regardless of when we start the observation. For example, the time until a given radioactive particle decays will be independent of when you start with counting the time. This property is known as 'memorylessness' or 'Markov property' which we will study further in Chapter 6. If we have a CDF $F(X)$ and its tail distribution $\bar{F}(X) = 1 - F(X)$ for random variable X, it will be memoryless if we have

$$P(X > s + t | X > t) = P(X > s)$$

or

$$\frac{\bar{F}(s+t)}{\bar{F}(t)} = \bar{F}(s).$$

Not a lot of probability distributions have that memoryless property. In fact, there are only two such distributions - **exponential distribution** for the continuous case and **geometric distribution** for the discrete case.

Let us focus on the continuous case here. Notice we have the equation $\frac{\bar{F}(s+t)}{\bar{F}(t)} = \bar{F}(s)$, we can replace $\bar{F}(X)$ by $g(x)$ and rewrite the equation into

$$g(s + t) = g(s)g(t)$$

after some simple manipulation.

Now, the task is to find all such g for the above equality to hold. Note that when $s = t = 1$, we have $g(2) = [g(1)]^2$, and when $s = t = 1/2$, we have $g(1) = [g(1/2)]^2$ and $g(1/2) = [g(1)]^{1/2}$. By extension of these two equalities, we get

$$\begin{aligned} g(a) &= [g(1)]^a \\ &= e^{\ln g(1) a} \\ &= e^{-\lambda a} \end{aligned}$$

where $\lambda = -\ln g(1)$.

This is the complementary CDF, and the CDF is $F_X(x) = 1 - e^{-\lambda x}$ for positive x. By differentiating that, we get PDF $f(x) = \lambda e^{-\lambda x}$.

Using the PDF, we can get its MGF. We have

$$M_X(t) = \mathrm{E}[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t}.$$

We can get the expectation and variance of exponential distribution by differentiating the MGF. So,

$$\mathrm{E}[X] = M_X'(0) = \frac{1}{\lambda}$$

and

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = M_X''(0) - (M_X'(0))^2 = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}.$$

–

Another way to get exponential distribution is to use Poisson distribution. Exponential distribution describes the time until an unlikely event to happen. An unlikely event's occurrence can be modelled by a Poisson distribution, and we can divide the time into a series of discrete time with length 1.

First, for a Poisson random variable X with expectation λ, we have

$$P(X = 0) = e^{-\lambda}$$

and if it occurs for t times non-stop, its probability is$e^{-\lambda t}$. This gives us the probability $P(T > t)$ where T stands for the first time the unlikely event occurs. It is a tail distribution, the corresponding CDF is

$$F_T(t) = 1 - P(T > t) = 1 - e^{-\lambda t}$$

and the PDF obtained after differentiation is

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

which is same as the function we obtained from the other method.

–

The value of the above derivation is that we can extend it further to obtain the **gamma distribution**.

Instead of the first occurrence of an event, let us find the k-th occurrence of that event. With each of 0-th to $k-1$-th occurrence, the corresponding probability from the Poisson distribution will change accordingly. The CDF will then be

$$\begin{aligned} F_T(t) &= P(T \leq t) \\ &= 1 - P(T > t) \\ &= 1 - P(0, 1, \dots k-1 \text{ events in } [0, t]) \\ &= 1 - \sum_{x=0}^{k-1} \frac{(\lambda t)^x e^{-\lambda t}}{x!} \end{aligned}$$

and by differentiating, we get the PDF

$$\begin{aligned} f_T(t) &= \frac{d}{dt}(1 - \sum_{x=0}^{k-1} \frac{(\lambda t)^x e^{-\lambda t}}{x!}) \\ &= \frac{d}{dt}(1 - e^{-\lambda t} - \sum_{x=1}^{k-1} \frac{(\lambda t)^x e^{-\lambda t}}{x!}) \\ &= \lambda e^{-\lambda t} - \sum_{x=1}^{k-1} \frac{1}{x!}(x \cdot (\lambda t)^{x-1} \cdot \lambda \cdot e^{-\lambda t} - \lambda \cdot (\lambda t)^x \cdot e^{-\lambda t}) \\ &= \lambda e^{-\lambda t} - \lambda \cdot e^{-\lambda t} \sum_{x=1}^{k-1} \frac{1}{x!}(x \cdot (\lambda t)^{x-1} - (\lambda t)^x) \\ &= \lambda e^{-\lambda t} + \lambda \cdot e^{-\lambda t} \sum_{x=1}^{k-1} \Big(\frac{(\lambda t)^x}{x!} - \frac{(\lambda t)^{x-1}}{(x-1)!} \Big) \\ &= \lambda e^{-\lambda t} + \lambda \cdot e^{-\lambda t} \Big(\frac{(\lambda t)^{k-1}}{(k-1)!} - 1 \Big) \\ &= \frac{\lambda \cdot e^{-\lambda t} (\lambda t)^{k-1}}{(k-1)!} \\ &= \frac{1}{\Gamma(k)(1/\lambda)^k} e^{-\frac{t}{(1/\lambda)}} (t)^{k-1} \end{aligned}$$

where $\Gamma(r)$ is the gamma function with $\Gamma(r) = \int_0^\infty z^{r-1} e^{-z} dz$, and the two parameters are scale $1/\lambda$ and shape k.

We will state without working for the following properties: the MGF of it is $(1 - t/\lambda)^{-k}$, the expectation is k/λ and the variance is k/λ^2.

3.6 Summary

	PMF / PDF	E[X]	Var[X]	MGF
Bernoulli(p)	$p^x(1-p)^{1-x}$	p	$p(1-p)$	$(1-p)+pe^t$
Binomial(n,p)	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$	$[(1-p)+pe^t]^n$
Poisson(λ)	$(e^{-\lambda}\lambda^x)/(x!)$	λ	λ	$e^{\lambda(e^t-1)}$
Normal(μ,σ^2)	$(1/(\sqrt{2\pi}\sigma))(e^{(x-\mu^2)/(2\sigma^2)})$	μ	σ^2	$e^{\mu t+(\sigma^2t^2)/2}$
Exp(λ)	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$	$\lambda/(\lambda-t)$
Gamma(k,θ)	$\frac{1}{\Gamma(k)(\theta)^k}e^{-\frac{t}{(\theta)}}(t)^{k-1}$	$k\theta$	$k\theta^2$	$(1-\theta t)^{-k}$