3
Probability Distributions

Figure 3.1: A cartoon capturing the essence of the Heisenberg's uncertainty principle: $\Delta x \Delta p \geq \frac{\hbar}{2}$; where Δx is the uncertainty in position, Δp is the uncertainty in momentum (or velocity), and $\hbar$ is the Planck's constant.

DISTRIBUTIONS ARE generalizations of mathematical functions from a purely technical standpoint. But perhaps it is most pertinent to begin by asking a more utilitarian question. Why should we study distributions? Specifically, why should we study *probability* distributions? One of the motivations stems from a practical limitation of experimental measurements that is underlined by the uncertainty principle postulated by Werner Heisenberg (**cf.** Figure 3.1). The very fabric of reality and the structure of scientific laws that govern our ability to understand physical phenomena demand a probabilistic (statistical) approach. Our inability to make infinite-precision measurements of data necessitates the consideration of averages over many measurements, and under similar conditions, as a more reliable strategy to affix experimental values to unknowns with reasonable accuracy. Typically, the operation of averaging is equivalent to performing an integral of the form $\overline{u} := \int_{\mathfrak{D}} u(x)dx$ over the domain $\mathfrak{D}$. The function $u(x)$ is a mathematical representation of a quantity of interest whose average we are interested in calculating over $\mathfrak{D}$. This simple expression for the average is computed by integrating uniformly over $\mathfrak{D}$. More generally, the average may be computed by using an appropriate weight function $f(x)$ and integrating as $\int_{\mathfrak{D}} u(x)f(x)dx$. Here $f(x)$ serves as the distribution function over $\mathfrak{D}$ that appropriately weighs $u(x)$ during the averaging process.

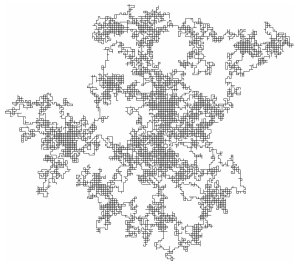

Figure 3.2: Signature of a two dimensional Brownian motion. Like in the one dimensional case, the expected value of the two dimensional random velocity vector $\langle \mathbf{U} \rangle \equiv \mathbf{0}$.

As an example, consider a large collection of particles moving in one dimensional space (the particles are free to move either to the left or to the right with any speed). The velocity of the i^{th} particle is given by $u^{(i)}$. Let us suppose that we want to find the average velocity of the flow. One option is to find an ensemble average of the velocities of many such particles. This is equivalent to computing $\overline{u} := \int_{\mathfrak{D}} u(x)dx$ if the velocity profile of the overall flow, $u(x)$, is known to us over the domain $\mathfrak{D}$. Alternatively, if the velocity distribution function of the particles $f(u)$ is known, then we can compute $\overline{u} \equiv \langle U \rangle = \int_{\mathfrak{U}} uf(u)du$ where $\mathfrak{U}$ is the appropriate domain in the velocity space and $f(u)$ prescribes the probability distribution of U.[1] Two interesting cases are noteworthy here: (i) if the aforementioned particles follow a Brownian motion (**cf.** Figure 3.2), then $f(u)$ is the Maxwell-Boltzmann (symmetric and bell-shaped) distribution and $\overline{u} = 0$ as expected, and (ii) if the flow is unidirectional and every particle moves with constant speed u_0, then clearly $\overline{u} = u_0$. The latter result may be expressed as a weighted integral by using the Dirac delta distribution as follows: $\overline{u} = \int_{\mathfrak{U}} u\delta(u - u_0)du = u_0$. In order to facilitate a deeper appreciation of the practical utility of distribution functions, let us delve a little more on this example of a collection of particles moving with velocity u_0. A gas in thermal equilibrium comprises a collection of such parti-

[1] $\langle U \rangle \equiv E(U)$ is the expected value of the random velocity variable U and is analogous to the ensemble average.

cles. Let us say we want to deduce the notion of temperature of the gas. This may be done by associating the concept of temperature to the average kinetic energy of the particles. If all the particles have an identical velocity u_0, then the average kinetic energy is simply $\frac{1}{2}mu_0^2$. However, even in a steady and uniform flow, it is unlikely that every particle constituting the flow will move identically with velocity u_0. Even though the majority of the particles may have the velocity u_0, there are likely going to be exceptional particles with slightly higher or lower velocities. So instead of applying the Dirac delta distribution, a more appropriate weight function may be $f(u)$ with a distribution profile akin to the one shown in Figure 3.3. So in order to estimate the average kinetic energy of the particles, the distribution $f(u)$ must be employed to probe the kinetic energy law $\psi(u) = \frac{1}{2}mu^2$ across the spectrum of particle velocities. Thence, the average kinetic energy is $\int_{-\infty}^{\infty} \psi(u)f(u)du$ which now has a clear mathematical interpretation in the sense of the distribution $f(u)$.

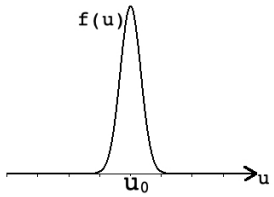

Figure 3.3: Distribution of velocity of particles in a uniform flow shows that a majority of the particles move with a velocity u_0 although there may be some small deviations.

Thus, statistical averages (and *moments*)[2] rely on the probability distribution functions of the relevant variables and are commonly used in developing many statistical models of practical importance such as the kinetic theory of gases. We must now begin a formal study of such distributions that lie at the heart of all statistical analysis.

[2] In this chapter, we will encounter terms such as $E(X^r)$, $r \in \mathbb{I}^+$, that are known as the moments of the random variable X.

3.1 Chapter objectives

The chapter objectives are listed as follows.

1. Students will learn the notion of probability mass function, probability density function, and cumulative distribution function.
2. Students will learn to compute expected values and higher order statistical moments using probability distribution functions.
3. Students will learn the concept of moment generating functions and use it to deduce the distribution of a random variable.
4. Students will learn the concept of joint probability distributions of multiple random variables. They will learn to deduce the marginal probability distribution from the joint probability distribution.
5. Students will study different types of discrete and continuous random variables. They will learn how to appropriately characterize a given phenomenon using one or many of these named probability models.
6. Students will learn to design and analyse an application project from the actuarial sciences (e.g., insurance model) by using a certain compound probability distribution model.

3.2 Chapter project: Predicting insurance claim aggregates during a policy period

3.2.1 Epilogue: Modeling insurance claims using a compound probability distribution

A certain insurance company is interested in predicting the total aggregate of all claims made during a fixed policy period from a portfolio of insurance products. Such an exercise will enable the company to make an assessment of its financial risks while charting out product launch schedules for the upcoming financial year.

Figure 3.4: Claims are made to an insurance manager at the end of a policy period.

A consultant to the company designs the following mathematical model to accomplish this task. Consider that the firm expects a certain number (N_j) of claims, from amongst its clients, during a fixed period j. Since there is no reason for this number N_j to be deterministically computable,[3] it is reasonable to assume N_j to be a random variable. Now there are N_j of these claims, each claim amount is independent of the other and is also independent of N_j. This is also reasonable because each claim is made by a different client acting independent of the other. Further, each claim amount is also a random number which possibly corresponds to a common probability distribution. Let the claim amount by the i^{th} client be denoted by X_i. X_i corresponds to a probability distribution function $F_X(x)$. The aggregate claim for the policy period j under consideration is also a random quantity $Y_j = X_1 + X_2 + \cdots X_{N_j} = \sum_{i=1}^{N_j} X_i$ that obeys a compound probability distribution. Based on this model, a quantity of interest to the insurance firm is $E(Y_j)$ that you as the consultant will have to estimate in this project.

Moreover, consider there are four policy periods in a given financial year. The total premium collected at the beginning of the year by the insurance firm is $\$\, m$. Let λ_j be the rate at which claims are received per policy period j. Now consider $Z = \sum_{k=1}^{4} Y_k$ is the aggregate claim at the end of the 4^{th} policy period (year end). The company incurs a loss if $Z > \$\, m$. In this project, you will simulate a certain compound stochastic process in Matlab and compute the associated risk for the insurance firm in terms of a probability $P(Z > \$\, m)$. Concurrently, you will learn about a composite stochastic model known as the *compound Poisson process* that is used by insurance companies to assess their risks.

Before we urge the readers to work on this aforementioned project, let us first learn some of the essential and fundamental elements of probability distributions.

[3] A multitude of external factors may determine the value of N_j. The complex inter-relationship between these factors may further enhance the uncertainty in knowing what the exact value of N_j might be.

3.3 *Geometrical interpretation of integration with respect to a distribution function*

We may recall from our elementary calculus course that the *Riemann integral* $\int_{\mathfrak{D}} u(x)dx$ with respect to the independent variable x may be interpreted as area under the curve $u(x)$ (**cf.** Figure 3.5). This Riemann integral can be approximated by the Riemann sum $\sum_{x_i \in \mathfrak{D}} u(x_i^*)\Delta x_i$, where $x_i^* \in [x_{i-1}, x_i]$, $\forall i = 1, 2, \cdots, n$ and the x_is are generally equally spaced n nodes (grid points) in the domain of integration $\mathfrak{D}$ along the x axis. Simply put, this sum is an aggregate of areas of thin rectangular strips of height $u(x_i^*)$ and width Δx_i. The Riemann sum is equal to the integral $\int_{\mathfrak{D}} u(x)dx$ in the limit $\Delta x_i \to 0$ (for all i) if and when this limit exists. This computation relies on the fact that all *observable* x values[4] are equally important and equally likely to be encountered in a practical situation. Hence the function $u(x_i^*)$ is multiplied by the same unit scalar for all values of x_i^* in the summand. However, this need not always be the case as has been explained in the example discussed in the introductory paragraphs of this chapter. Consider that the importance factors of the different x_i^* values are illustrated by the profile $f(x)$ defined over all the x_i^*s in $\mathfrak{D}$ (**cf.** Figure 3.6). This relative importance of certain x_i^*s over the others necessitates scaling the $u(x_i^*)$s appropriately by the weight factors $f(x_i^*)$s. The accurate representation of this case entails that we now have a modified integral of the form $\int_{\mathfrak{D}} u(x)f(x)dx$ which may be approximated by the sum $\sum_{x_i \in \mathfrak{D}} f(x_i^*)u(x_i^*)\Delta x_i$.

It is important to note that the profile of the importance factors of the observables as prescribed by the distribution function $f(x)$ will be different for different practical applications. In fact, much of this chapter and the subsequent models that we will discuss in this book will highlight this fact emphatically. Further, since the averaging procedure of

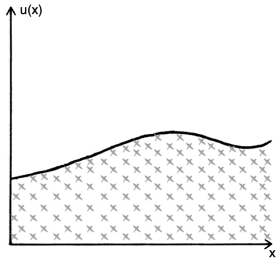

Figure 3.5: Profile of the function $u(x)$ along x. The shaded area under the curve $u(x)$ is given by $\int u(x)dx$.

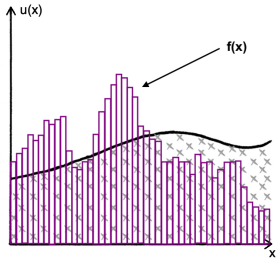

Figure 3.6: Profile of the weight function $f(x)$ demonstrates the relative importance of the observables x in $\mathfrak{D}$.

[4] We are now using a terminology for x that will serve as a bridge from calculus to probability vocabulary.

the function $u(x)$ is being performed under the *guidance* of the distribution function $f(x)$, it is more appropriate to underscore this fact by using a modified notation for the integral, viz., $\int_{\mathfrak{D}} u(x)dF(x)$. In the context of probability theory, there is a unique relation between the cumulative distribution function $F(x)$ and the probability density function $f(x)$ that we will address in one of the subsequent sections of this chapter $\left(f(x) = \frac{dF(x)}{dx}\right)$. Thus far in our discussion here we have been referring to the probability density function $f(x)$ as the distribution function in the general sense of computing averages of functions. Going forward in subsequent sections of this chapter, and in the following chapters, we will make this distinction explicit in most scenarios.

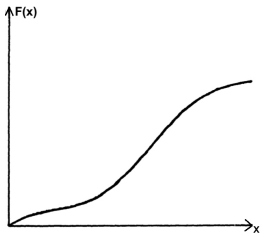

Figure 3.7: Distribution profile of the observables x is prescribed by some function $F(x)$.

Before we dive further into the conceptual elements of probability distributions, it may be useful to shed light on the geometrical meaning of the integral $\int_{\mathfrak{D}} u(x)dF(x)$ which is widely known as the *Riemann–Stieltjes integral*. For this purpose, we will consider a distribution profile of the observables prescribed by some function $F(x)$ as shown in Figure 3.7. Since the function being averaged ($u(x)$) has an *independent* existence compared to the distribution profile of the observables ($F(x)$), each of x, u, and f can be represented along an independent axis in a three dimensional representational space. Further, since $F(x)$ and $u(x)$ have independent origin and existence, the profile of a sheet traced by $u(x)$ along x, and protruding out of the $u - x$ plane, will resemble hills and valleys along the x axis but look flat (straight) along the F direction. Thus the height of this undulating sheet is prescribed by $u(x)$. If we were to consider another surface that cuts through this sheet, that emanates out of the $u - x$ plane, under the *guidance* of the curve traced by $F(x)$, then a *fence*-type surface will emerge whose height is given by $u(x)$. This is shown in Figure 3.8. Clearly the area of the projection of this *fence* on the $u - x$ plane gives the familiar area under the curve $u(x)$ that can be computed by the integral $\int_{\mathfrak{D}} u(x)dx$. The projection of this *fence* on the $u - F$ plane, on the other hand, is denoted by the shaded shadow region whose area is given by the *Riemann–Stieltjes integral* $\int_{\mathfrak{D}} u(x)dF(x)$. For the special case $F(x) = x$ in $\mathfrak{D}$, the Riemann–Stieltjes integral (integration with respect to the distribution $F(x)$) becomes identical to the more familiar Riemann integral $\int_{\mathfrak{D}} u(x)dx$. This aforementioned explanation is based on the discussion reported in the article published in the *The American Mathematical Monthly* by Gregory L. Bullock[5].

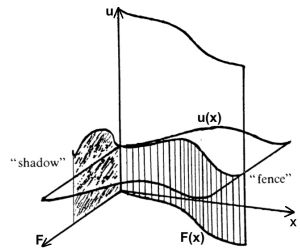

Figure 3.8: Geometrical meaning of an integral with respect to a distribution function. This figure is borrowed from the work of Gregory Bullock referenced below.

[5] Gregory L. Bullock. "A Geometric Interpretation of the Riemann-Stieltjes Integral". In: *The American Mathematical Monthly* 95.5 (1988), pp. 448–455

3.4 *Discrete vs continuous probability distributions*

In the previous chapter, under the section on random variables, we have encountered two different types, namely, discrete and continuous random variables. In case of the former type, the random variables take on distinct values. A simple example is the outcome of tossing a fair coin - it is a *head* or a *tail* with a designated random value of 1 or 0, each with a probability equal to $\frac{1}{2}$. In the continuous case, the outcomes are such that the random variable may take on a continuum of values over a range prescribed by the sample space Ω. A classic example is the test score obtained by a student who is registered in a course, this score may be any real number between 0 and 100. It may be considered a random number because without any specific information about this student's performance in the test or his/her talent/skill-level in the subject or the nature of the test itself, it may be hard to make a definitive prediction of the student's score. Under the circumstances, it may be prudent to deduce the score by sampling without bias from the bell-shaped Gaussian distribution (**cf.** Figure 2.8).

In this section, we will formally define the probability distribution profile for discrete and continuous random variables. Consequently, these definitions will be helpful to calculate statistical moments (means, variances, etc.) of the respective random variable and thereby make forecasts of important events. These calculations become very useful when we do not have access to any sample data but instead have some understanding of the underlying stochastic phenomenon which allows us to identify, with reasonable accuracy, the relevant probability/stochastic model and utilise its probability distribution profile.

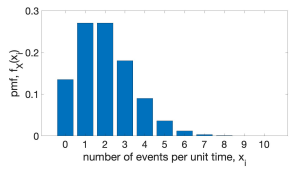

Figure 3.9: Probability mass function $f_X(x_i)$ of a certain discrete random variable. It may be verified that $\sum_{x_i \in \Omega} f_X(x_i) = 1$.

3.4.1 *Definition: Probability mass function*

For a discrete random variable, each possible observable $x_i \in \Omega$ has a certain probability of occurrence $p_i := P(X = x_i)$ which we can think of as a *probability mass*. Obviously, $\sum_{x_i \in \Omega} P(X = x_i) = 1$ due to the second axiom of probability (axiom of unitarity) and this serves as a conservation law. We will use the notation $f_X(x_i) \equiv p_i$ to denote the probability mass function (p.m.f.) of a discrete random variable. The profile of $f_X(x_i)$ presents a visual depiction of how the probability masses are interspersed over the sample space Ω (**cf.** Figure 3.9).

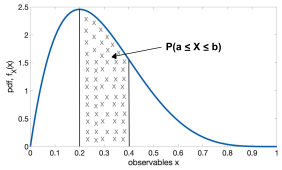

Figure 3.10: Probability density function $f_X(x)$ of a certain continuous random variable.

3.4.2 *Definition: Probability density function*

In the case of a continuous random variable X, the probability mass is spread continuously over the range of the observables. Therefore, it is appropriate to use the notion of a *density function* $f_X(x)$, instead of probability mass. The unitarity axiom of probability enforces the following normalization of the probability density function (p.d.f.): $\int_{x \in \Omega} f_X(x)dx = 1$. It follows that $P(a \leq X \leq b) = \int_a^b f_X(x)dx$ represents the area under the curve f between a and b (**cf.** Figure 3.10).

3.4.3 *Definition: Cumulative distribution function*

The cumulative distribution function (c.d.f.) $F_X : \mathbb{R} \to [0,1]$ is defined as $F_X(x) \equiv F(x) := P(X \leq x)$, $x \in \mathbb{R}$. It follows that $P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$. The c.d.f. F_X has the following properties:

(i) $\lim_{y \downarrow -\infty} F(y) = 0$,

(ii) $\lim_{y \uparrow \infty} F(y) = 1$,

(iii) $\lim_{y \downarrow x} F(y) = F(x), \quad \forall x \in \mathbb{R}$ (i.e. F_X is right-continuous).

The first two properties imply that F is always a non-decreasing function. It must be noted that the definition of the c.d.f. as stated above is valid for both discrete and continuous random variables. Let us consider the case of a continuous random variable. Clearly $F(x) = \int_{-\infty}^{x} f(\zeta)d\zeta$ for a continuous real-valued function f (p.d.f.) whence F is uniformly continuous and differentiable, thereby $\frac{dF(x)}{dx} = f(x) \implies dF(x) = f(x)dx$ which is a direct consequence of the *fundamental theorem of calculus*.

There are two main interpretation of the distribution function $F_X(x)$ that is noteworthy to mention here.

(I) $F_X(x)$ prescribes the *distribution of probability mass* on the real line. Concomitantly, $F(b) - F(a)$ is the mass concentrated in the interval $(b - a)$. For the discrete case,

locations of concentrated point mass on the real line (x_i) are points of discontinuity of F_X with jumps proportional to $p_i \equiv F_X(x_i + 0) - F_X(x_i - 0)$.[6] There are a finite or a countable number of such jumps and F_X is continuous everywhere else.

(II) $F_X(x)$ encompasses the accumulation of probability masses (or density) up to x. Therefore, it is *additive*, non-negative, and has a unit maximum value. Thus, the c.d.f. F qualifies as a *measure* (*F-measure*). In section 3.3 above, we have commented on this aspect of interpreting the linear functional $F(u)$ defined by $F(u) = \int_{\mathfrak{D}} u(x)dF(x)$ as an integral of a measurable function $u(x)$ over $\mathfrak{D}$ with respect to the F-measure $F(x)$.[7]

[6] A distribution with only concentrated point masses is a discrete distribution and one without is a continuous distribution.

[7] The F-measure corresponds to the *Jordan-Peano* measure (Jordan content) that extends the notion of *size* to more complicated geometry.

3.4.4 *Statistical moments and their significance*

Earlier in section 2.8, we have encountered the notion of expected value and variance of a random variable. Thence we had defined the expectation and variance as a weighted aggregate of the observables and mean squared deviation respectively. The weight factors were taken to be the probability masses. In this section, we will extend the definitions in terms of the distribution functions. In what follows here, we will consider X as the random variable (discrete or continuous) and the observables $x \in \Omega$. The statistical moments, defined below, determine the shape of the distribution function and hence characterize data. Renowned Russian mathematician Pafnuty Lvovich Chebyshev was the first to systematically define and use statistical moments of random variables during the mid-nineteenth century.

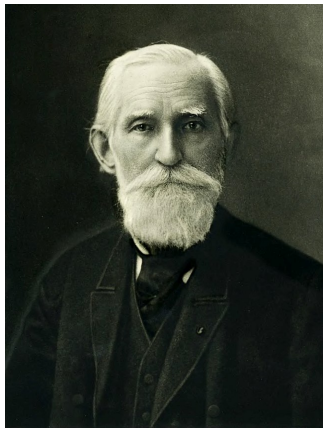

Figure 3.11: Pafnuty Lvovich Chebyshev (1821–1894) was a prominent Russian mathematician and professor of algebra, number theory, and probability at St. Petersburg University (*courtesy: Wikimedia Commons*).

i) **Mean** (μ or $E(X)$) is the *first statistical moment*.

$$E(X) = \sum_{x\in\Omega} xP(X = x) \quad \text{(discrete case)}, \tag{3.1}$$

$$E(X) = \int_{x\in\Omega} xf(x)dx = \int_{x\in\Omega} xdF(x) \quad \text{(continuous case)}. \tag{3.2}$$

ii) **Variance** (σ^2 or $Var(X)$) is the *second statistical moment*.

$$Var(X) = E((X-\mu)^2) = \sum_{x\in\Omega} (x-\mu)^2 P(X = x) \quad \text{(discrete case)}, \tag{3.3}$$

$$Var(X) = \int_{x\in\Omega} (x-\mu)^2 f(x)dx \quad \text{(continuous case)}. \tag{3.4}$$

Equivalently, $\boxed{Var(X) = E(X^2) - \big(E(X)\big)^2}$, a result that follows by algebraically unravelling the expression $E((X-\mu)^2)$.

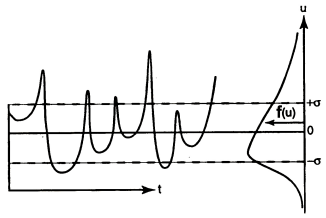

Figure 3.12: Time series data $u(t)$ with positive skewness ($\mu_3 > 0$).

iii) **Skewness** (μ_3) is the *third standardised moment*.

$$\mu_3 = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right) = \frac{E((X-\mu)^3)}{(Var(X))^{3/2}} \tag{3.5}$$

μ_3 measures the degree of asymmetry of the probability density function. A p.d.f. that is symmetric about the mean has zero skewness. All higher order odd moments of such a symmetric p.d.f. will also be identically zero. Data $u(t)$ with positive skewness is characterized by a p.d.f. with a longer tail for $X - \mu > 0$ than for $X - \mu < 0$ (here X represents the random variable that is sampled over time). Hence a positive skewness

means that deviation $X - \mu$ is more likely to take on large positive values than large negative values. For instance, a time series data with long periods of small negative values and a few instances of large positive values, with zero temporal mean, has positive skewness (**cf.** Figure 3.12).

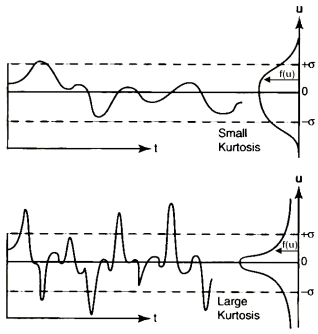

Figure 3.13: Time series data $u(t)$ with small (top) and large (bottom) kurtosis μ_4. Large values of μ_4 correspond to data with intermittent extreme events.

iv) **Kurtosis** (μ_4) is the *fourth standardised moment*.

$$\mu_4 = E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) = \frac{E((X-\mu)^4)}{(Var(X))^2} \tag{3.6}$$

A p.d.f. with longer tails will have a larger kurtosis than a p.d.f. with narrower tails. A time series data $u(t)$ with most measurements clustered around the mean has low kurtosis. A time series dominated by intermittent extreme events has high kurtosis.

3.4.5 *Discrete and continuous probability distribution models: construction and applications*

Models of the real world processes must account for the element of uncertainty that are inherently omnipresent. Therefore, there is a strong case for designing and using probabilistic models that suitably capture random phenomena. In this section, we will study many important probabilistic models of both discrete and continuous processes.

Discrete probability models

i) ***Bernoulli distribution:*** This is a binary probability model with only two possible outcomes. Some examples of this model are the outcomes of tossing a fair coin, success or failure of a projectile in hitting its target, etc. Let us consider that the random variable X can take one of two possible values 1 or 0 with probability p and $1-p$. The probability mass function is defined below.

$$X \sim Bernoulli(p).$$
$$f_X(x) = \begin{cases} p, & \text{when } x = 1, \\ 1-p, & \text{when } x = 0. \end{cases} \tag{3.7}$$

The expected value and variance of a Bernoulli random variable are calculated as follows:

$$E(X) = ((1 \times p) + (0 \times (1-p)) = p. \tag{3.8}$$
$$\begin{aligned} Var(X) &= \sum_{x=\{0,1\}} (x - E(X))^2 f_X(x) = (0-p)^2(1-p) + (1-p)^2 p \\ &= p(1-p). \end{aligned} \tag{3.9}$$

Example: fixing wooden planks to a wall by a battery of nail guns

A battery of ten nail guns are fired simultaneously to affix wooden planks to a wall. The probability that a nail gun successfully drives a nail into a wooden

plank is 0.95 (hit rate). The plank attaches to the wall if at least seven of the ten nails are driven through the plank into the wall successfully. Do you think this battery of guns can successfully affix planks to a wall on an average?

Let $X_i \sim Bernoulli(p = 0.95)$ where $i = 1, 2, 3, ..., 10$. $X_i = 1$ represents an event that a nail is successfully driven through the plank into the wall. $X_i = 0$ represents a compromised nail. Consider the random variable $Y = \sum_{i=1}^{10} X_i$ that captures the total number of successful hits by the battery.

$$E(Y) = E(\sum_{i=1}^{10} X_i) = \sum_{i=1}^{10} E(X_i) = 10p = 9.5. \tag{3.10}$$

So the average hit rate of the battery is 9.5 that is greater than 7. So, on an average, the battery of guns can be successfully used to affix planks to a wall.

Figure 3.14: A battery of nail guns is used to fix a wooden plank to a wall. The failure rate of each gun $(1 - p) = 0.05$ is small enough to ensure that the battery successfully fixes planks to the wall on an average.

ii) ***Binomial distribution:*** The above example demonstrates that a sequence of Bernoulli trials can be used to a model a Binomial random process. Here, we are interested in accounting for a random number (X) of successes (each with probability p) in n independent Bernoulli trials.

$$X \sim Bin(n, p).$$

$$f_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & \text{for } k = 0, 1, 2, 3, ..., n, \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

Taking a cue from the calculation shown in equation (3.10), we can estimate the expected value and variance of a Binomial random variable as follows.

$$E(X) = np. \tag{3.12}$$

$$Var(X) = np(1-p). \tag{3.13}$$

Figure 3.15: Boca-cola vs Moca-cola contest: two colas in a blind taste campaign.

Example: free cola tasting and smart marketing campaign

Boca-cola is a newly emerging cola brand in the market that faces stiff competition from a very popular and old cola brand namely Moca-cola. Both colas look the same. The Boca-cola company made a smart move of inviting one hundred sworn customers of Moca-cola to a free cola tasting campaign during the innings interval of the world cricket championship final held in the city of Borgo Verde. Two anonymous samples of cola (one from each brand) was offered to each of the hundred tasters and they were asked to vote their preference. The Boca-cola company made profitable use of the Binomial distribution which has most of its mass concentrated within three standard deviations of the mean. The following calculations demonstrate how they made a compelling case for a surge in Boca-cola sales even in the face of stiff competition from their bigger rival Moca-cola.

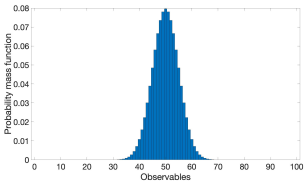

Figure 3.16: The p.m.f. of $Y \sim Bin(100, 0.5)$.

The campaign was successful on the premise that a typical Moca-cola drinker will not be able to tell the difference between two colas during a blind test. Hence, they are equally likely to prefer one cola over the other (i.e., probability that a random participant in the blind cola tasting test prefers Boca-cola (or Moca-cola) is one-half, $p = 0.5$). Consider a Bernoulli random variable X which takes a value 1 (when Boca-cola is preferred) or 0 (when Moca-cola is preferred). Let $Y = \sum_{i=1}^{100} X_i$ denote the number of instances when a sworn Moca-cola drinker preferred the Boca-cola during the blind test. Clearly, $Y \sim Bin(n = 100, p = 0.5)$. The following lines of Matlab code is used to construct the relevant probability mass function of Y (**cf.** Figure 3.16).

```
n = 100; p = 0.5;
x = 0:n;
y = binopdf(x,n,0.5);
figure, bar(x,y,1);
xlabel('Observables');
ylabel('Probability mass function');
set(gca,'FontSize',60);
```

The mean and variance of Y is calculated as $E(Y) = np = 100 \times 0.5 = 50$ and $\sigma^2 = Var(Y) = np(1-p) = 25 \implies \sigma = 5$. It is clear from the profile of the distribution of Y in Figure 3.16 that almost all the probability mass is concentrated between $Y = 35$ and $Y = 65$, i.e. within three standard deviation of the mean. Therefore, $P(Y \geq 35) \approx 1$ which signifies that more than 35% of sworn Moca-cola drinkers prefer the newly launched Boca-cola.[8] A 35% switch by sworn clients of a competing brand was indeed a compelling advertisement campaign!

[8] We can check using the Matlab command `»sum(y(35:end))` that $P(Y \geq 35) = 0.9996 \approx 1$.

iii) ***Geometric distribution:*** Consider a sequence of Bernoulli trials with probability of success equal to p. Let X be the number of failures before the first success. The probability mass function for such a random variable is defined as follows.

Geometric distribution of type-0

$$X \sim geom_0(p)$$

$$P(X = x) = \begin{cases} (1-p)^x p, & \text{for } x = 0, 1, 2, 3, ..., \\ 0, & \text{otherwise.} \end{cases} \tag{3.14}$$

$$E(X) = \frac{1-p}{p} \tag{3.15}$$

$$Var(X) = \frac{1-p}{p^2} \tag{3.16}$$

If Y is the random variable that counts the number of Bernoulli trials until first success,

the p.m.f. of Y is slightly different from the case above as mentioned below.

Geometric distribution of type-1

$$X \sim geom_1(p)$$

$$P(Y = y) = \begin{cases} (1-p)^{y-1}p, & \text{for } y = 1, 2, 3, ..., \\ 0, & \text{otherwise.} \end{cases} \tag{3.17}$$

$$E(X) = \frac{1}{p} \tag{3.18}$$

$$Var(X) = \frac{1-p}{p^2} \tag{3.19}$$

Example: Derivation of $E(X) = \frac{1}{p}$ for $X \sim geom_1(p)$

We will begin with the definition of expectation.

$$\begin{aligned} E(X) &= \sum_{x=1,2,...} xP(X = x) \\ &= \sum_{x=1,2,...} x(1-p)^{x-1}p \\ &= p\sum_{x} x(1-p)^{x-1} =: S \end{aligned} \tag{3.20}$$

Consider $S_1 = \sum_{x=1,2,...} (1-p)^x = \frac{1}{1-(1-p)}$ which is a convergent infinite geometric series. $\frac{dS_1}{dp} = \sum_{x=1,2,...} x(1-p)^{x-1}(-1) = -\frac{1}{p^2}$. Therefore, using equation 3.20, $S = \frac{1}{p^2}$ and consequently $E(X) = \frac{1}{p}$.
Likewise, using $Var(X) = E(X^2) - (E(X))^2$, writing $E(X^2) = E(X(X-1)) + E(X)$, and following similar steps as shown above here, we can deduce the expression for $Var(X) = \frac{1-p}{p^2}$.

The distribution function (c.d.f.) of the geometric distribution of type-1 is $F_X(k) := P(X \leq k) = 1 - P(X > k)$. $P(X > k)$ can be calculated as follows.

$$\begin{aligned} P(X > k) &= P(X = k+1) + P(X = k+2) + \cdots \\ &= (1-p)^k p + (1-p)^{k+1}p + \cdots \\ &= (1-p)^k p(1 + (1-p) + (1-p)^2 + + \cdots) \\ &= (1-p)^k p \frac{1}{1-(1-p)} \\ &= (1-p)^k \end{aligned} \tag{3.21}$$

Therefore $F_X(k) = 1 - (1-p)^k$.

Figure 3.17:
Cunning fox, while passing by,
Thought to taste some grapes on high,
So he leaps, and leaps again,
For they say - there is no gain without pain!
The story of the fox and the grapes, adapted from *Aesopica*, is an old anthem to learn new math tricks with the geometric distribution.

Example: Memoryless property of geometric distribution

Any random variable X has a memoryless property if, for any $n, m \geq 0$, we have $P(X > n+m | X > m) = P(X > n)$. We will demonstrate that this is certainly true of the geometric random variable $X \sim geom_1(p)$.

$$\begin{aligned} P(X > n+m | X > m) &\overset{\text{definition of conditional probability}}{=} \frac{P(\{X > n+m\} \cap \{X > m\})}{P(X > m)} \\ &= \frac{P(X > n+m)}{P(X > m)} \\ &= \frac{(1-p)^{n+m}}{(1-p)^m} \\ &= (1-p)^n \\ &\overset{\text{using equation 3.21}}{=} P(X > n) \end{aligned} \tag{3.22}$$

Practically what this means is the following. *Suppose we are about to start flipping a fair coin for which the probability of observing a head is p. Then the probability distribution of X := "number of flips until the first head is observed" is $geom_1(p)$. Now suppose that the first m flips are tails. Then the probability distribution of the new random variable Y := "number of additional flips until the first head is observed" is still $geom_1(p)$. It is as if the knowledge of the outcomes of the first m flips has been lost from memory!* The geometric distribution is the only known discrete probability distribution with this property.

Figure 3.18: Bummer spent the whole day in class wondering whether or not he flushed the toilet before he left his apartment! Does our friend Bummer have a cognitive state that follows the geometrical distribution? (*Courtesy: Michael Tran, Daily Bruin*).

iv) ***Poisson distribution:*** The number of occurrences of an event (e.g., arrival of busses in a bus stand, calls to a telephone operator, etc.) in a fixed interval of time can be random. In order to derive an expression for the distribution of such a random variable, we make two fundamental assumptions about the random instantiations (henceforth referred to as "arrivals").

(a) *Homogeneity:* The arrival rate λ is constant with respect to time. The expected number of arrivals in a given interval of time Δt is $\lambda \Delta t$. This is also known as *weak stationarity* as will be discussed in one of the latter chapters in this book.

(b) *Independence:* The number of arrivals in any two disjoint intervals of time are independent of each other.

Let N_t be the number of arrivals in an interval $[\tau, \tau + t]$ for any $\tau > 0$. We seek to know the distribution of N_t. Homogeneity implies that $E(N_t) = \lambda t$. Next, we proceed with constructing n canonical intervals of time t/n in such a way that as $n \to \infty$, M_j is a Bernoulli random variable representing the number of arrivals (0 or 1) in the interval $I_{j,n} := [(j-1)\frac{t}{n}, j\frac{t}{n}]$ for any $j \in \mathbb{I}$. By definition, $E(M_j) = 0(1-p_j) + 1(p_j) = p_j =$

$E(N_{\frac{t}{n}}) = \lambda\frac{t}{n}$ where p_j is the probability that $M_j = 1$ and $(1 - p_j)$ is the probability that $M_j = 0$ in the interval $I_{j,n}$.[9] Consequently,

$$N_t = \sum_{j=1}^{n} M_j \sim Bin(n, p),$$

where $p \equiv p_j = \lambda\frac{t}{n}$.[10] Therefore,

$$P(N_t = k) = \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1 - \frac{\lambda t}{n}\right)^{n-k}, \tag{3.23}$$

for $k = 0, 1, 2, ..., n$. Explicitly, we have not yet considered the assumption $n \to \infty$ in the mathematical expressions above. We hope that after taking this limit, the distribution function will stabilize. We will consider the limit of each of these terms separately and collate the results thereafter.

$$\lim_{n\to\infty}\binom{n}{k}\frac{1}{n^k} = \lim_{n\to\infty}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}\frac{1}{k!} = \frac{1}{k!}, \tag{3.24}$$

$$\lim_{n\to\infty}\left(1 - \frac{\lambda t}{n}\right)^n = e^{-\lambda t}, \tag{3.25}$$

result from elementary calculus

and certainly

$$\lim_{n\to\infty}\left(1 - \frac{\lambda t}{n}\right)^{-k} = 1. \tag{3.26}$$

Now combining the results from equations 3.24, 3.25, and 3.26, we have

$$\lim_{n\to\infty} P(N_t = k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}. \tag{3.27}$$

By carefully inspecting the term on the right hand side of equation 3.27, we notice that the following result holds.

$$e^{-\lambda t}\sum_{k=0}^{\infty}\frac{(\lambda t)^k}{k!} = e^{-\lambda t}e^{\lambda t} = 1. \tag{3.28}$$

Results 3.27 and 3.28 entail that we have indeed chanced upon a legitimate probability distribution (the *Poisson distribution*) that complies with the unitary axiom of probability over the sample space $\Omega = \{0, 1, 2, \cdots\}$. In the expression on the right hand side of equation 3.27, we have only one parameter λt. This chain of thought motivates the definition of the Poisson distribution with parameter $\mu > 0$[11] to model the counting process of random number of arrivals in fixed intervals of time.

$$X \sim Poisson(\mu)$$

$$P(X = k) = \frac{\mu^k}{k!}e^{-\mu} \quad \text{for } k = 0, 1, 2, ... \tag{3.29}$$

$$E(X) = \mu \tag{3.30}$$

$$Var(X) = \mu \tag{3.31}$$

[9] This is a direct consequence of the homogeneity axiom: $E(N_{\Delta t}) = \lambda(\Delta t)$.

[10] Here we have used the fact that the number of arrivals (successes) N_t in n Bernoulli trials is $Bin(n., p)$ where p is the probability of one arrival (success) in each canonical interval $\frac{t}{n}$.

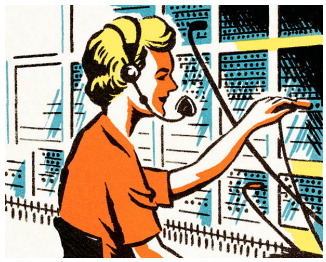

Figure 3.19: The number of calls received per hour by a telephone operator follows a Poisson distribution.

[11] You may think of μ as λt.

The mean can be deduced from the fact that in the preceding paragraph $N_t \sim Bin(n, \frac{\lambda t}{n})$ and hence $E(N_t) = n\frac{\lambda t}{n} = \lambda t,\ \forall n$. Further, $\lim_{n\to\infty} Var(N_t) = \lim_{n\to\infty} n\frac{\lambda t}{n}\left(1 - \frac{\lambda t}{n}\right) = \lambda t$. Therefore, for $X \sim Poisson(\mu)$, we expect that $E(X) = \mu$ and $Var(X) = \mu$. In fact,

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k\frac{\mu^k}{k!}e^{-\mu} \\ &= e^{-\mu}\sum_{k=1}^{\infty}\frac{\mu^k}{(k-1)!} \\ &= \mu e^{-\mu}\sum_{k=1}^{\infty}\frac{\mu^{k-1}}{(k-1)!} \\ &= \mu e^{-\mu}\sum_{j=0}^{\infty}\frac{\mu^j}{j!} \\ &= \mu e^{-\mu}e^{\mu} \\ &= \mu. \end{aligned} \tag{3.32}$$

The calculation for $Var(X)$ follows a similar approach and is left to the reader as a self-exercise.

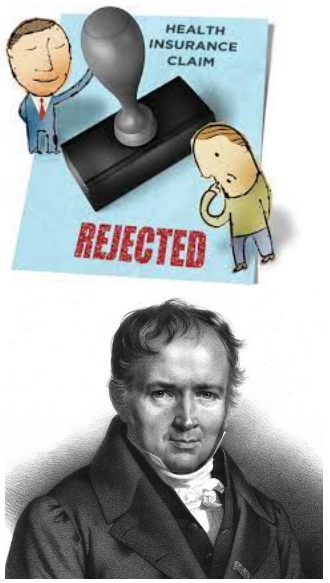

Figure 3.20: The saga of your insurance claims may have been scripted by the French Mathematician - Siméon Denis Poisson (*courtesy: Wikimedia Commons*).

Example: Risk of loss incurred by Carepal from insurance payouts

A watered-down version of the chapter project is considered in this example. A workers' insurance company named *Carepal* has introduced a new insurance policy for factory workers to cover certain types of injuries sustained at work. Since a worker may be inflicted by a diverse type of injuries in the factory, and that the insurance policy may not be applicable for all types of injuries as per the coverage plan, only a certain number out of the total claims get re-reimbursed by Carepal. Further, the insurance scheme allows only a standard payment of ₹ 1,00,000 for all claims approved by Carepal. The annual premium for the policy is ₹ 15 for each insured person. Based on claims data available with the company, it was found that on an average, a total of about 100 claims per year get approved for similar schemes. There are about $10,00,000$ policy holders of this scheme. What is the risk (calculated in terms of a probability) that this factory workers' scheme will yield an annual loss for Carepal?[12]

The insurance scheme is designed in such a manner that the probability of successful approval of any given claim is very small. Of course not all claims made by the workers will get approved by Carepal. Thus, from the perspective of a policy period, there are several probability events being played out at once whence only those claims that will be eventually approved by Carepal may be regarded as a *successful instantiation* of a claim (event). Let X denote the total number of claims that will be approved by Carepal during one policy period of a single year. Consequently, X may be regarded as a Poisson random variable with rate $\mu = 100$ such that $E(X) = 100$ and $Var(X) = 100$. Carepal will incur an annual loss if the aggregate of all claims payouts turns out to be greater than the total revenue gen-

erated by selling this insurance scheme to 10,00,000 customers. Let us use Matlab to compute the critical number of approved claims that will determine whether Carepal incurs loss from this insurance scheme.

```
total_customers = 1000000;
premium = 15;
std_pay_per_claim = 100000;
total_payout_for_loss = 15000000;
mu = 100;
income = premium*total_customers;
Number_payouts = total_payout_for_loss/std_pay_per_claim;
```

Here the value of `Number_payouts` turns out to be 150 which is the critical number of approved claims above which the company will incur loss. Our next objective is to estimate the probability that more than 150 numbers of claims will be approved by Carepal in the given policy year.[13]

$$P(X > \texttt{Number\_payouts}) = 1 - \sum_{k=0}^{\texttt{Number\_payouts}} e^{-\mu}\frac{\mu^k}{k!}$$

is calculated as follows.

```
k = [0:1:Number_payouts];
prob_mass_of_payouts = exp(-mu)*(mu.^k)./(factorial(k));
Risk_of_loss = 1 - sum(prob_mass_of_payouts)
```

It turns out that `Risk_of_loss = 1.2331e-06`, i.e. Carepal's probability of incurring a loss from this scheme is incredibly minuscule and the said insurance scheme is risk free by and large.

[12] Here we are interested in knowing the risk of incurring an annual loss stemming from this particular insurance scheme alone.
[13] Note that Carepal cannot simply enforce a hard stop on its claims approval process in order to arrest a likely annual loss in its business because such an intervention would severely dent its reputation and credibility as a trusted insurance company among factory workers. Further, it cannot raise its premium arbitrarily because such a measure would almost certainly play into the hands of its competitors.

Poisson heuristic: Recall that the construction of the Poisson distribution, as elucidated above, demands that the probability of an event happening (arrival) is small. Specifically, it is proportional to $\frac{\lambda t}{n}$ where $n \to \infty$ which makes $\frac{\lambda t}{n} \to 0$. Further, recall that the number of arrivals in any fixed time interval were assumed to be independent. It turns out that even if we relax the criterion of independence and consider *weakly dependent Bernoulli trials*, the Poisson distribution (more specifically the *Poisson heuristic*) may still provide a reasonable estimate. We will demonstrate this point with the help of an example below.

Figure 3.21: If *love is blind* then there is perhaps a case for *blind dates* and some Poisson heuristics.

Example: your luck with the blind dating app RATATOON

RATATOON is a new mobile application for blind dates. It is hosted on several cloud based servers. Every month, up to 20 randomly picked clients (10 men and 10 women) from the RATATOON provincial database can register on one of their cloud servers. The intelligence integrated within the software application ensures that hugely different profiles (for instance men and women more than 20 years apart in age, etc.) are not usually hosted on the same server in order to avoid absurd match-ups. Each server allows its customers to post a preferred date every month when they may be up for a blind date. The app locks in a potential blind date if the preferred dates of any two clients (one man and one woman) are within one day of each other and it is left up to the pair if they are willing to make this minor adjustment in their chosen dates. After a date is locked in, the individual pairs may decide whether to proceed with the date or not but chances are that they will give it a shot. What is the probability that two or more clients are matched up for a date by a RATATOON server every month? Further, what is the probability that at least 5% of the possible pair-ups (combinations) are matched-up for blind dates by RATATOON?

Had RATATOON paired up potential dates only if the preferred dates of each man and woman in the pair matched identically, then there is a precise and direct way of computing the required probability. However, by allowing up to one day difference in their preferred dates, the probability in question becomes very difficult to calculate exactly. In this case, the Poisson heuristic provides a reasonably good approximation of the asked probability.
We begin by identifying the total number of all possible pairs of men and women from a pool of 10 men and 10 women. A combination is represented as (M_j, F_j) where M stands for a man and F stands for a woman, and $j = 1, 2, ..., 10$ tags each of the ten men and women registered in the same server. This computation may be undertaken by asking in order to fill up each of two slots that constitute a pair, the first slot may be filled up in 10 different ways (corresponding to the 10 different men) and the second slot may be filled up in 10 different ways for similar reasons. Thus there are a total of $n = 10 \times 10 = 100$ trial combinations out of which some combinations may lead to a match-up if the preferred dates of the individuals are within a day of each other (the latter are the successful trials). It is not too difficult to find that each of the n trials have the same success probability $p = \frac{3}{30}$.[14] Each of the day of a month may be mapped to an empty slot of a 30 slot array. For the i^{th} combination (M_i, F_i), let us suppose that F_i picks the slot k. Now, if M_i picks any of the 30 slots but $(k-1)$, k or $(k+1)$, then it will be a *failed* match-up. This failure can happen with a probability $p_c = \frac{27}{30}$ and this estimate will be true of any failed match-up. Thus the probability of success for any combination (M, F) is $p = 1 - p_c = 1 - \frac{27}{30} = \frac{3}{30} = 0.1$.[15] Let X denote a random variable that denotes the number of successful trials (match-ups). Then the prob-

ability that two or more clients are matched-up for a date by RATATOON is the same as $P(X \geq 1)$.
A careful reflection of the situation reveals that the aforementioned Bernoulli trials are *weakly dependent* (as opposed to being independent). This is because that it is extremely unlikely that preferred dates coming from a diverse population will be clustered together. Consider F_1 prefers the 14^{th} of the month and M_1 picks the 13^{th}. (M_1, F_1) turns out to be a match-up because each of their preferences belong to the set $\Big([12,14] \cap [13,15]\Big)$. Additionally, (M_1, F_i) for $i = 3, 5$ are also match-ups due to similar date preferences (e.g., $[12,14]$). But with every such match-up for M_1, it becomes less probable that (M_1, F_j), where $j \neq 1, 3, 5$, will also be a match-up because it is unlikely that all women from the pool of 10 would have preferred dates between 12^{th} and 14^{th} of the month (unless of course we are talking about special months like February when many single women (and men) may prefer dates around a special day on the 14^{th} of February which is valentine's day). Thus, in some sense, the outcome of two different trial combinations may bear some element of dependency. This dependence of trial outcomes may be considered *weak* because of a relatively large number (30) of available preferred dates.[16]
Therefore, $X \approx Poisson(\mu)$ where $\mu = np = 100 \times 0.1 = 10$. Consequently,

$$P(X \geq 1) \approx 1 - P(X = 0) = 1 - e^{-\mu} = 0.9999546. \tag{3.33}$$

Further,

$$P(X \geq 5) \approx 1 - \sum_{k=0}^{k=4} P(X = k) = 0.97074731. \tag{3.34}$$

[14] For simplicity, we have considered only 30 day months.
[15] For preferred dates on the start and the end of the month, we have considered that the possible match-up dates can be 30^{th}, $1^{st}, 2^{nd}$ and $29^{th}, 30^{th}, 1^{st}$.
[16] The estimate would be better if the available number of preferred dates was larger, for example, instead of considering match-ups every month, one could consider match-ups every three months.

In order to establish the veracity of the calculations and results of the above example, we will design a simple computer experiment whereby ten male and ten female bots will randomly pick their preferred dates for a blind date on a given month. Consequently, match-ups will be generated by the computer by considering those pairs that have preferred to go on a blind date within a day of each other. This experiment is repeated one hundred thousand times in order to estimate the probability that at least a certain percentage (q) of pair-ups are matched-up for a blind date. This experimental result is then compared with the theoretical estimations prescribed by the above example.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%% START of CODE %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%% start of parameters %%%%%%%%%%%%%
nm = 10; % num of males
nf = 10; % num of females
n=nm*nf; % total number of possible pair-ups (combinations)
total_period = 30; % duration of monthly cycle of blind match-ups
match_interval = 3; % this ensures that a match-up happens when the preferred
```

```
%dates of a male and female pair is within a day of each other
match_percent = 5; % minimum target match-ups in percentage
req_matchups = ceil(0.01*match_percent*n);% 0.0x*n means atleast x% must be
%matchups
kmax = req_matchups - 1; % the running index for the Poisson heuristic
nmax = 100000; % number of times the computer experiment is conducted
%%%%%%%%%%%%%% end of parameters %%%%%%%%%%%%%
match_cnt = 0;
for i=1:nmax
    % RATATOON
    males = ceil(total_period*rand(1,nm)); % list of date preferences by
                                           % by males
    females = ceil(total_period*rand(1,nf));% list of date preferences by
                                           % by females
    [men,women] = meshgrid(males, females);
    pairs = [men(:) women(:)]; % all pair-ups are stored in matrix form
    diff_pairs = mod(abs(pairs(:,1) - pairs(:,2)),total_period); % difference
                                             % between dates preferred by
                                             % the male and the female in a
                                             % pair
    Ans = [diff_pairs pairs]; % concatenating diff_pairs and pairs in one matrix
    matchups = find(Ans(:,1)<=1); % matchups happen if diff_pairs is within
                                  % one unit
    num_matchups = length(matchups);
    prob_sim_matchups = num_matchups/(n); % this is not really necessary but
                                          % just for fun to compare with
                                          % p
    if num_matchups >=  req_matchups % this condition calculates the required
        match_cnt = match_cnt + 1;   % probability over multiple repetition
    else                             % of the experiment
        continue;
    end
end
prob_sim_final = match_cnt/nmax;    % experimental estimate of the required
                                    % probability


% poisson heuristic approximation
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
p=(1 - ((total_period - match_interval)/total_period)); % p = 1 - pc
mu = n*p;
k=[0:kmax];
prob = 1-sum(exp(-mu)*(mu.^k)./(factorial(k)));  % calculating the required
                                                 % probability by using the
                                                 % Poisson heuristic
prob_final = vpa(prob,8);  % expressing the answer up to 8 decimal places
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% END of CODE %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

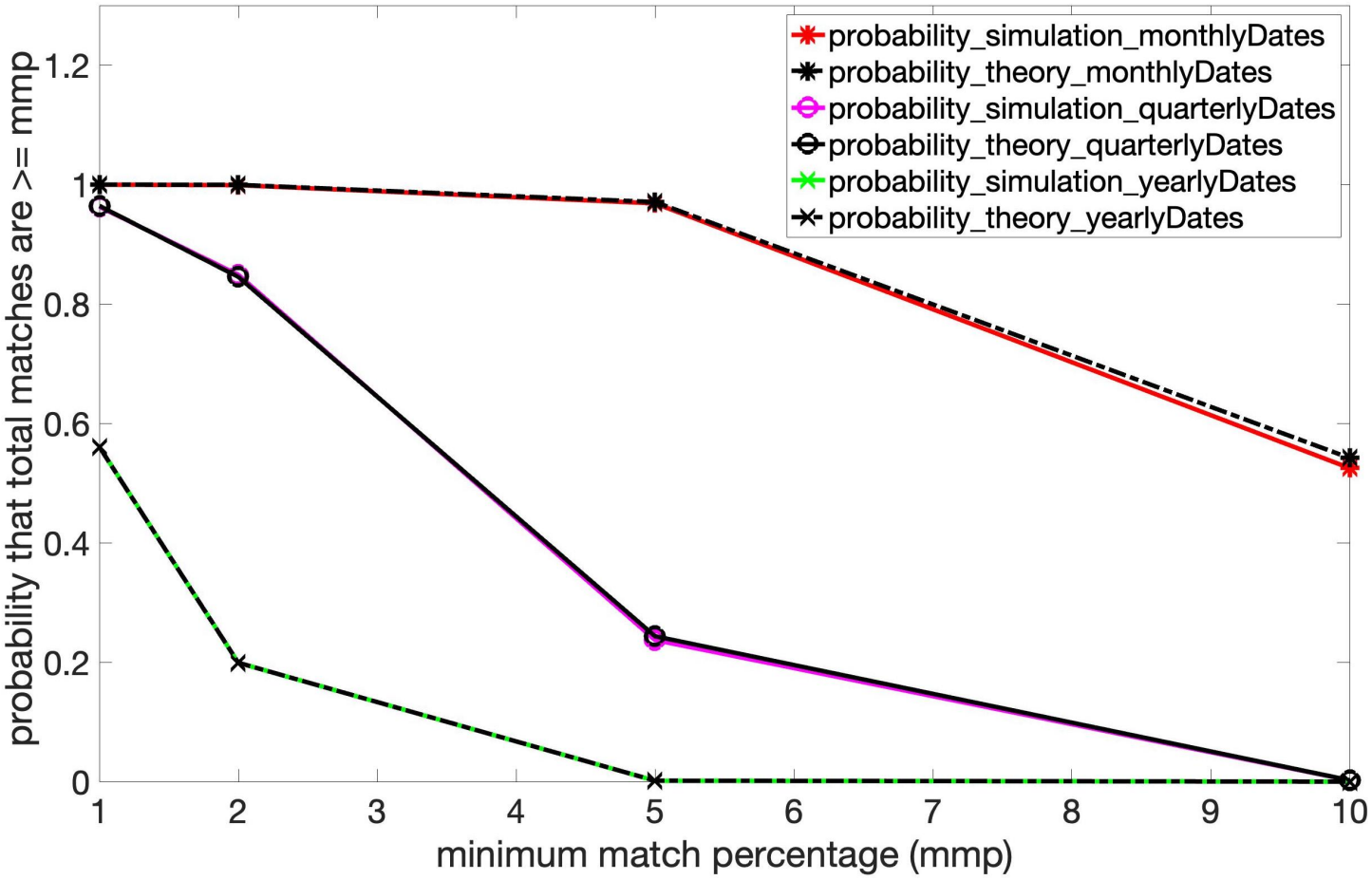

Figure 3.22: Comparison between simulated experiment of dating match-ups and theoretical predictions of the same clearly shows that the Poisson heuristic is a very good approximation when n is large, p is small, and $\mu = np$ is moderate in magnitude.

The results of the above computer experiment are elucidated graphically in Figure 3.22. Firstly, the theoretical predictions of the Poisson heuristic is strikingly similar to the actual simulated probability values. This establishes the veracity of the Poisson heuristic when n is large, p is small, and $\mu = np$ is of moderate magnitude. Secondly, both simulated and the Poisson model demonstrate that the probability, $P(X \geq \text{minimum match percentage})$, decreases with increasing values of the minimum match percentage (mmp). Thirdly, both simulated and the Poisson model show that the corresponding probabilities fall significantly with increasing duration of the experiments for the same mmp. This entails that in order to keep clients hooked to the RATATOON application, it may be economically prudent (from a profitable business perspective) to keep the duration (period) of the subscriptions shorter (monthly as opposed to yearly). In the experiments and analyses considered here, we have chosen a fairly large sample size of 100 clients per server to ensure statistical validity of reported findings.

There are several interesting examples illustrating the utility of the Poisson heuristic. Some of the more familiar ones in the literature are that of the *birthday problem* and the

matching problem. Interested readers are referred to the texts mentioned in the chapter bibliography. These problems have many important applications in cryptography and information security. We will return to the Poisson distribution once again during our discussion of the chapter project. Poisson processes will also be discussed in chapter sections pertaining to Markov chains and queuing models.

Figure 3.23: Rolling a die yields an outcome that has equal probability of happening as that of any other possible outcomes (*courtesy: Wikimedia Commons*).

v) ***Uniform distribution (discrete):*** Let $X \sim Unif([1,m])$ be a random variable on m successive integers starting with 1. Each outcome has an associated identical probability (uniform) prescribed as follows.

$$\begin{aligned} X &\sim Unif([1,m]) \\ p_i = P(X=i) &= \frac{1}{m},\ \forall i \in [1,m] \\ E(X) &= \sum_{x=1}^{m} x\frac{1}{m} = \frac{(m+1)}{2} \\ Var(X) &= E(X^2) - \big(E(X)\big)^2 = \frac{m^2-1}{12} \end{aligned}$$

A simple example is the outcome of rolling a six-sided die. Each of the $m = 6$ sides has a probability of occurrence equal to $\frac{1}{m} = \frac{1}{6}$. We will again return to the uniform distribution when we discuss continuous random variables.

vi) ***Negative-Binomial and Pascal distributions:*** In many practical situations we may be interested in knowing the chance of the r^{th} success of Bernoulli trials in $\nu = r + k$ trials, where $k = 0, 1, 2, \ldots$. This situation is equivalent to occurrence of exactly k failures prior to r successes. So if X is the random variable that denotes $\nu = k + r$ trials for the r^{th} success to occur, we may also re-phrase this as counting $\nu - 1 = r + k - 1$ trials with exactly k failures and a success in the very next trial. The number of possible ways this may happen is obviously $\binom{r+k-1}{k} = \binom{\nu-1}{k}$. This beckons the definition of the probability mass function as follows.

$$\begin{aligned} X &\sim Pa(k;r,p) \text{ or } X \sim NB(k;r,p) \\ P(X=k) &= \binom{r+k-1}{k}(1-p)^k p^r \equiv \binom{-r}{k}(-(1-p))^k p^r \ \text{ for } k = 0,1,2,3,\ldots \end{aligned}$$

Indeed, it may be verified that $\sum_{k=0}^{\infty} P(X=k) = 1$ is true as demanded by the unitarity axiom of probability. This entails that an infinite sequence of Bernoulli trials is bound to yield r successes. Thus we may infer that the Pascal or negative binomial distribution is a model for *the waiting time to the r^{th} success.*

Example: Did Blaise Pascal play badminton?

A game of badminton involves two players Li Pen and Brendon Hart with skill levels $p = 0.6$ and $h = 0.4$ respectively. Here skill levels may be interpreted as

probabilities of winning a rally by the respective players. In order to win a game of badminton, 21 individual victories are required by either player. Further, consider that a game can last at most 41 rallies (i.e. a winning scoreline of $21 - 20$ is permissible by the rules of the game). Answer the following questions.

i. What is the probability that Li Pen will win the game of badminton against Brendon Hart in 26 rallies?

ii. What is the probability that Li Pen will the game of badminton against Brendon Hart?

iii. Whats is the probability that a game of badminton between Pen and Hart will end in 26 rallies?

We proceed with the calculations inspired by our formulation of the Pascal distribution above.

i. A game lasts at least $2\nu + 1 = 21$ rallies and at most $4\nu + 1 = 41$ rallies. Thus $\nu = 10$. So if Pen wins at rally number $4\nu + 1 - r = 26$, we have $r = 15$. Let us define $P_r = P_{15}$ as the probability that Pen wins the game in $4\nu + 1 - r = 26$ rallies. Then, Pen must have won $2\nu = 20$ out of $4\nu - r = 25$ rallies and lost 5 out of 25 rallies. Thus, $P(\text{Pen wins in 26 rallies}) = P_{15} = \binom{25}{20}(0.6)^{21}(0.4)^5 = 0.0119$.

ii. $P(\text{Pen wins}) = P(\text{Pen wins in 1 rally}) + P(\text{Pen wins in 2 rallies}) + ... + P(\text{Pen wins in 41 rallies}) = P_{20} + P_{19} + ... + P_1 + P_0 = 0.9035$.

iii. $P(\text{game ends in 26 rallies}) = P_{15} + H_{15} = 0.0120$ where $H_{15} = \binom{25}{20}(0.4)^{21}(0.6)^5$.

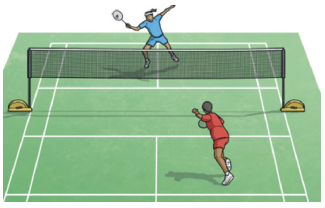

Figure 3.24: A game of badminton between Li Pen and Brendon Hart.

Continuous probability models

i) ***Exponential distribution:*** X is a positive continuous random variable with *rate* parameter $\mu > 0$. X is exponentially distributed with the p.d.f. prescribed below.

$$\begin{aligned} X &\sim exp(\mu) \\ f_X(x) &= \begin{cases} \mu e^{-\mu x} \text{ for } x > 0; \\ 0, \text{ otherwise.} \end{cases} \\ E(X) &= \frac{1}{\mu} \\ Var(X) &= \frac{1}{\mu^2}. \end{aligned}$$

If we consider the case of Poisson distributed arrivals in a fixed interval of time, then the inter-arrival times are exponentially distributed random variables. The c.d.f. is $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(\zeta)d\zeta$ where f is the p.d.f. mentioned above. It follows

that

$$F_X(x) = \begin{cases} 1 - e^{-\mu x} \text{ for } x \geq 0; \\ 0 \text{ for } x < 0. \end{cases} \tag{3.35}$$

Example: Memorylessness of the exponentially distributed random variable

A continuous random variable X has a memoryless property if

$$P\big(X > t + s \big| X > s\big) = P(X > t) \text{ for all } t > 0,$$

regardless of the value of $s > 0$. This is easy to show because if we consider the definition of the conditional probability, then

$$P\big(X > t + s \big| X > s\big) = \frac{P(X > t + s)}{P(X > s)} = \frac{-\mu(t+s)}{e^{-\mu s}} = e^{-\mu t} = P(X > t).$$

The exponential distribution is the only known continuous distribution with the memoryless property.

Perhaps in light of the above example and the one discussed earlier using the geometric distribution about the memoryless property, it is not difficult to intuit that there must be a connection between the exponential distribution and the geometric distribution. This is further unravelled by the following discussion. Exponentially distributed random variables are used to model time until the occurrence of a rare event.

Example: Equivalence of geometric and exponential distributions

Recall from equation 3.21 that for $X \sim geom_1(p)$, we have $P(X > k) = (1-p)^k$. This means $F_X(k) = 1 - (1-p)^k$. Further, let $T \sim exp(\mu)$ whence $F_T(t) = 1 - e^{-\mu t}$. If we suppose $\mu = -\log(1-p)$ and $t = \lfloor n\tau \rfloor$[17] for all $\tau > 0$ and $n = 0, 1, 2, 3, ...$, then clearly $F_X(\lfloor n\tau \rfloor) = 1 - e^{\big(\log(1-p)\big)\lfloor n\tau \rfloor} = 1 - (1-p)^{\lfloor n\tau \rfloor} = F_T(\lfloor n\tau \rfloor)$, i.e. the probability distributions of X (geometrically distributed random variable) and T (exponentially distributed random variable) are the same.
The connection may also be established by considering the case when the exponentially distributed random variable $T \in (n-1, n]$ for $n = 1, 2, 3, ...$ and $p = 1 - e^{-\mu}$ (equivalently, $\mu = -\log(1-p)$).

$$\begin{aligned} P(n-1 < T \leq n) &= F_T(n) - F_T(n-1) \\ &= (1 - e^{-\mu n}) - (1 - e^{-\mu(n-1)}) \\ &= p(1-p)^{n-1}. \end{aligned} \tag{3.36}$$

The r.h.s. of equation 3.36 is identical to the p.m.f. of the geometrical distribution of type-1 (i.e. $P(X = n)$).

[17] Here $\lfloor\ \rfloor$ is the greatest integer function akin to the *floor operation*. For instance, $\lfloor 2.63 \rfloor = 2$.

There is a more rigorous relationship between the geometric distribution and the exponential distribution that we will simply state here without providing a proof. *Let* $Z_n = \frac{Y_n}{n}$ *where* Y_n *is a geometric random variable with parameter* $p_n = \frac{\lambda}{n}$ *and* $n > \lambda > 0$. *Then* Z_n *converges in distribution*[18] *to an exponential random variable with parameter* λ.

[18] Let $X_1, X_2, ...$ be a sequence of real-valued random variables that converges in distribution to X, $X_n \xrightarrow{\mathfrak{D}} X$, if $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ at which $F(x)$ is continuous. $F_i(x)$ is the c.d.f. of X_i, $\forall i = 1, 2, ...$ and $F(x)$ is the c.d.f. of X.

Example: Chance of total power failure in a single-engine jet aircraft

Most single-engine jet aircrafts have an auxiliary power unit (APU) as backup power supply in case of an engine failure. Typically the APU is activated when the main engine fails and while the pilot initiates an *engine re-start* procedure. If the APU also fails in flight in addition to the main engine, then we have a total system (power) failure. The lifetime of an operating power unit is exponentially distributed with expected value $\frac{1}{\mu}$. Let X denote the time until the first total system failure. It can be shown that $E(X) = \frac{2-e^{-\mu\tau}}{\mu(1-e^{-\mu\tau})} = 500$ flying hours, where $\tau > 0$ is the fixed time to re-start the main engine. What is the probability that the aircraft will encounter a total system failure after 100 flying hours?

$P(X > 100) \approx e^{-\frac{100}{E(X)}} = 0.8187$, i.e. there is a 82% chance of a total system failure after 100 flying hours.

Figure 3.25: This single engine jet has an auxiliary power unit as a backup during engine failure. How likely will this built-in redundancy prove to be helpful after one hundred flying hours?

Below, we have provided a summary of few other continuous probability distributions.

ii) ***Gamma distribution:*** Here the parameters $\alpha > 0$ is a scale parameter and $\beta > 0$ is a rate parameter. $\Gamma(\alpha)$ is the well known gamma function.

$$\begin{aligned} X &\sim Gamma(\alpha, \beta) \\ f_X(x) &= \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ for } x > 0; \\ 0, \text{ otherwise.} \end{cases} \\ E(X) &= \frac{\alpha}{\beta} \\ Var(X) &= \frac{\alpha}{\beta^2}. \end{aligned}$$

The Gamma distribution is used to model aggregate insurance claims and the amount of rainfall accumulated in a reservoir. It is used for modelling attenuation of signal strength in wireless communications. It finds applications in oncology for modelling age distribution of cancer incidence. It is also used in Bayesian statistical models.

iii) ***Beta distribution:*** Here both the parameters $\alpha > 0$ and $\beta > 0$ are shape parameters. $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function that is defined here in terms of the gamma function.

$$\begin{aligned} X &\sim Beta(\alpha, \beta) \\ f_X(x) &= \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \text{ for } x \in [0,1]; \\ 0, \text{ otherwise.} \end{cases} \\ E(X) &= \frac{\alpha}{\alpha+\beta} \\ Var(X) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

The Beta distribution is used in the theory of order statistics, in subjective logic in the form of posteriori probability estimates of binary events, and in wavelet analysis. Beta distribution is also used in project management models.

iv) ***Pareto distribution:*** Here the parameters $x_m > 0$ and $\alpha > 0$ are scale and shape parameters respectively. $\Gamma(\alpha)$ is the well known gamma function.

$$\begin{aligned} X &\sim Pareto(x_m, \alpha) \\ f_X(x) &= \begin{cases} \frac{\alpha x^\alpha}{x^{\alpha+1}} \text{ for } x \in [x_m, \infty); \\ 0, \text{ otherwise.} \end{cases} \\ E(X) &= \begin{cases} \infty, \text{ for } \alpha \leq 1 \\ \frac{\alpha x_m}{\alpha-1}, \text{ for } \alpha > 1 \end{cases} \\ Var(X) &= \begin{cases} \infty, \text{ for } \alpha \leq 2 \\ \frac{\alpha x_m^2}{(\alpha-1)^2(\alpha-2)}, \text{ for } \alpha > 2 \end{cases} \end{aligned}$$

The Pareto distribution was originally used to model the distribution and allocation of wealth among individuals in a society where the greatest fortune is owned by a small fraction of the population.

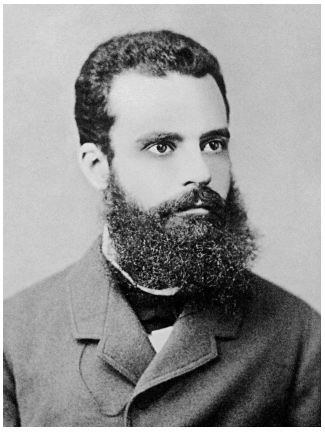

Figure 3.26: The Pareto distribution is named after Italian sociologist Vilfredo Pareto (*courtesy: Wikimedia Commons*).

v) ***Uniform distribution (continuous):*** Here the parameters $-\infty < a < b < \infty$ define the extent of the uniform distribution.

$$\begin{aligned} X &\sim Unif(a, b) \\ f_X(x) &= \begin{cases} \frac{1}{b-a} \text{ for } x \in [a,b]; \\ 0, \text{ otherwise.} \end{cases} \\ E(X) &= \frac{a+b}{2} \\ Var(X) &= \frac{(b-a)^2}{12}. \end{aligned}$$

vi) ***Normal distribution (a.k.a. Gaussian distribution):***

We will take a more detailed look at this familiar bell-shaped Normal distribution in the chapter on *Statistical Experiments*. Here we shall summarize some very essential

features. In what follows, μ can be negative or positive but finite, $\sigma^2 > 0$ by definition.

$$\begin{aligned} X &\sim N(\mu, \sigma^2) \\ f_X(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}} \text{ for } x \in \mathbb{R}; \\ E(X) &= \mu \\ Var(X) &= \sigma^2. \end{aligned}$$

Normal distribution is one of the most widely used probability models in statistics partly because of their relevance in connection to the central limit theorem which we will discuss later.

3.4.6 *Definition: Compound probability distribution*

Consider the sum $Y = X_1 + X_2 + \cdots + X_N$ where (i) N is a random number, (ii) X_i, $i = 1, 2, 3, ..., N$ are independent and identically distributed random variables with c.d.f. F_X,[19] and (iii) each X_i are independent of N.[20] By the law of total probability, the *compounded* distribution of Y is prescribed as follows:

[19] X is a random variable with mean μ_X and variance σ_X^2.

[20] N is a random variable with mean μ_N and variance σ_N^2.

$$\begin{aligned} f_Y(y) &= P(Y = y) \\ &= \sum_{n=0}^{\infty} P(X_1 + X_2 + \cdots + X_N = y | N = n) P(N = n) \\ &= \sum_{n=0}^{\infty} f_Y^{(n)} P(N = n), \end{aligned} \tag{3.37}$$

where $f_Y^{(n)}$ is the $n-$fold convolution of f_Y.[21] Next, we will compute the first two moments of a random variable with compound distribution.

[21] We have used the notation $f_X(x)$ and $p(x)$, and $f_X^{(n)}(x)$ and $p^{(n)}(x)$ interchangeably in the case of discrete random variables. For the continuous case, we have used $f_X(x)$ and $f_X^{(n)}(x)$ exclusively.

$$E(Y) = E_N(E_Y(Y|N))$$

Law of iterated expectation, eq. 2.15

$$\begin{aligned} &= \sum_{n=0}^{\infty} E(Y|N = n) P(N = n) \\ &= \sum_{n=0}^{\infty} nE(X) P(N = n) \\ &= \mu_X \sum_{n=0}^{\infty} nP(N = n) \\ &= \mu_X \mu_N. \end{aligned} \tag{3.38}$$

$$Var(Y) = E_N(Var(Y|N)) + Var_N(E(Y|N))$$

Law of total variance, eq. 2.19

$$\begin{aligned} &= E_N(NVar(X)) + Var(NE(X)) \\ &= Var(X)E(N) + (E(X))^2 Var(N) \\ &= \mu_N \sigma_X^2 + \mu_X^2 \sigma_N^2. \end{aligned} \tag{3.39}$$

The example presented in this section is very similar to the chapter project and should serve as a building block to solve it.

Example: aggregate claims of an insurance policy

Let the number of claims, N, generated by a portfolio of insurance policies over a fixed duration has Poisson distribution with rate parameter $\lambda = 3$ claims per policy period. Individual claim amounts X_i (for all values of $i = 1, 2..., N$) can be 1 or 2 million euros with probabilities $q = 0.6$ and $p = 0.4$, respectively. The aggregate claim $Y = X_1 + X_2 + \cdots + X_N$ is a compound Poisson distributed random variable. Find $P(Y = k)$ for $k = 0, 1, 2, 3, 4$. Also find the expected aggregate claim $E(Y)$.

We begin by noting that Y may take a maximum value equal to 4 as the insurance portfolio is capped at 4 million euros. This means that the number of claims may be $N = 0, 1, 2, 3, 4$. X_i is a Bernoulli random variable with probability of success (claim = 2 million euros), $p = 0.4$ and probability of fail (claim = 1 million euros), $q = 0.6$. Then, for a certain realization $N = n$, $(n = 1, 2, 3, 4)$, the sum $X_1 + \cdots + X_n$ is a binomial distributed random variable $\sum_{i=1}^{n} X_i \sim bin(n, p)$. The $n-$ fold p.m.f. are listed below.

$$
\begin{aligned}
p^{(1)}(1) &= 0.6 \\
p^{(1)}(2) &= 0.4 \\
p^{(2)}(2) &= \binom{2}{0}(0.4)^0(0.6)^2 = 0.36 \\
p^{(2)}(3) &= \binom{2}{1}(0.4)^1(0.6)^1 = 0.48 \\
p^{(2)}(4) &= \binom{2}{2}(0.4)^2(0.6)^0 = 0.16 \\
p^{(3)}(3) &= \binom{3}{0}(0.4)^0(0.6)^3 = 0.216 \\
p^{(3)}(4) &= \binom{3}{1}(0.4)^1(0.6)^2 = 0.432 \\
p^{(3)}(5) &= \binom{3}{2}(0.4)^2(0.6)^1 = 0.288 \\
p^{(3)}(6) &= \binom{3}{3}(0.4)^3(0.6)^0 = 0.064 \\
p^{(4)}(4) &= \binom{4}{0}(0.4)^0(0.6)^4 = 0.1296 \\
p^{(4)}(5) &= \binom{4}{1}(0.4)^1(0.6)^3 = 0.3456 \\
p^{(4)}(6) &= \binom{4}{2}(0.4)^2(0.6)^2 = 0.3456 \\
p^{(4)}(7) &= \binom{4}{3}(0.4)^3(0.6)^1 = 0.1536 \\
p^{(4)}(8) &= \binom{4}{4}(0.4)^4(0.6)^0 = 0.0256
\end{aligned}
$$

In order to find the p.m.f., we simply scale the $n-$ fold p.m.f. by the Poisson distributed probability weights with $\lambda = 3$.

(i)

$$
\begin{aligned}
P(Y=0) &= e^{-\lambda} = 0.0498 \\
P(Y=1) &= 0.6\lambda e^{-\lambda} = 0.0896 \\
P(Y=2) &= 0.4\lambda e^{-\lambda} + 0.36\frac{\lambda^2 e^{-\lambda}}{2} = 0.1404 \\
P(Y=3) &= 0.48\frac{\lambda^2 e^{-\lambda}}{2} + 0.216\frac{\lambda^3 e^{-\lambda}}{6} = 0.1559 \\
P(Y=4) &= 0.16\frac{\lambda^2 e^{-\lambda}}{2} + 0.432\frac{\lambda^3 e^{-\lambda}}{6} + 0.1296\frac{\lambda^4 e^{-\lambda}}{24} = 0.1544
\end{aligned}
$$

(ii) Without the capping of four million euros on the aggregate claim per policy period, the insurance company realizes that the expected aggregate claim is $E(Y) = \mu_X \mu_N = \big(2(p) + 1(q)\big)\mu_N = 1.4 \times 3 = 4.2$ million euros. This estimation may have served as a guide to cap the aggregate claim to approximately 4 million euros.

3.4.7 *Definitions: Joint and marginal probability distributions*

Consider a discrete random variable X with distribution F_X and another independent discrete random variable Y with distribution F_Y. Further, let X and Y be defined on the same sample space. The collection of points (x_i, y_j), $i, j = 1, 2, 3, ...$ that prescribes the joint event $\{X = x_i, Y = y_j\}$ forms an *event space* with probabilities, known as *joint probability mass function*, written as $P(X = x_i, Y = y_j) = p(x_i, y_j) \equiv f_{XY}(x_i, y_j)$.

The *marginal probability mass functions*, $p(x_i) \equiv f_X(x_i)$ and $p(y_j) \equiv f_Y(y_j)$ can be computed by integrating out the complementary dimension.[22] Therefore,

$$
\begin{aligned}
p(x_i) &\equiv f_X(x_i) = \sum_{y_j} P(X = x_i, Y = y_j) \equiv f_{XY}(x_i, y_j), \\
&\text{and} \\
p(y_j) &\equiv f_Y(y_j) = \sum_{x_i} P(X = x_i, Y = y_j) \equiv f_{XY}(x_i, y_j).
\end{aligned} \tag{3.40}
$$

[22] X and Y may be regarded as *complementary* to each other in the joint event space.

Obviously, the unitary axiom of probability ensures that $\sum_{y_j}\sum_{x_i} f_{XY}(x_i, y_j) = 1$. It is important to note here that if $f_{XY}(x, y) = f_X(x) f_Y(y)$, then the random variables X and Y are *independent* of each other (**cf.** section 2.4.5 for the definition of *independent* events).[23]

[23] Further, it can be shown that if X and Y are independent random variables, then $E(XY) = E(X)E(Y)$.

Example: Typographical errors in a manuscript

Consider a bouquet of manuscripts each of which are three pages long. Based on several editorial proof checking exercises over many years, it has been observed that typically authors commit three typographical errors per manuscript of this type. In this example, we will investigate the distribution of error-free pages in such manuscripts.

As an illustrative example, we have chosen a more tractable case here, of three errors distributed across three printed pages. Let us denote a typographical error by the

symbol † and a page as a container in a 3-tuple representation of the aforementioned manuscript family (i.e., (-/††/†) represents a case when two of the three typographical errors are found in page 2 and one in page 3 while page 1 is observed to be free of any errors). The event space is tabulated below.

(†††/-/-)	(-/†††/-)	(-/-/†††)	(††/†/-)	(††/-/†)
(†/††/-)	(-/††/†)	(†/-/††)	(-/†/††)	(†/†/†)

Let N denote the number of pages that have at least one error. Let X_i denote the number of errors in the i^{th} page. Let us begin by considering the joint distribution of X_1 and N. For each event case, we enlist below the realizations of X_1 and N.

i. (†††/-/-): $X_1 = 3,\ N = 1$,

ii. (-/†††/-): $X_1 = 0,\ N = 1$,

iii. (-/-/†††): $X_1 = 0,\ N = 1$,

iv. (††/†/-): $X_1 = 2,\ N = 2$,

v. (††/-/†): $X_1 = 2,\ N = 2$,

vi. (†/††/-): $X_1 = 1,\ N = 2$,

vii. (-/††/†): $X_1 = 0,\ N = 2$,

viii. (†/-/††): $X_1 = 1,\ N = 2$,

ix. (-/†/††): $X_1 = 0,\ N = 2$, and

x. (†/†/†): $X_1 = 1,\ N = 3$.

Figure 3.27: A day in the life of a copy editor who is busy finding typographical errors in a manuscript. The probability distribution of typographical errors in manuscripts is a useful marker for publishing houses to allocate resources in the copy editing process.

The enlisting of the event space above enables us to write the joint probability mass function $f_{X_1 N}(x, n)$ by computing the relative frequency of occurrence. The following table captures both the joint probability mass function as well as the marginal probability mass functions (a.k.a. the *marginals*). The marginals are calculated from the joint p.m.f. by summing along the rows and the columns respectively.

				X_1		$f_N(n)$
		0	1	2	3	
	1	$\frac{2}{10}$	0	0	$\frac{1}{10}$	$\frac{3}{10}$
N	2	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	0	$\frac{6}{10}$
	3	0	$\frac{1}{10}$	0	0	$\frac{1}{10}$
$f_{X_1}(x)$		$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	

Likewise, we can calculate the joint p.m.f.s for (X_2, N) and (X_3, N) followed by their respective marginals.

				X_2		$f_N(n)$
		0	1	2	3	
	1	$\frac{2}{10}$	0	0	$\frac{1}{10}$	$\frac{3}{10}$
N	2	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	0	$\frac{6}{10}$
	3	0	$\frac{1}{10}$	0	0	$\frac{1}{10}$
$f_{X_2}(x)$		$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	

				X_3		$f_N(n)$
		0	1	2	3	
	1	$\frac{2}{10}$	0	0	$\frac{1}{10}$	$\frac{3}{10}$
N	2	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	0	$\frac{6}{10}$
	3	0	$\frac{1}{10}$	0	0	$\frac{1}{10}$
$f_{X_3}(x)$		$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	

Clearly, all the joint distribution functions for (X_1, N), (X_2, N), and (X_3, N) are identical because it is natural to expect that the occurrences of the typographical errors are page-agnostic. Further, in each case, the marginals f_{X_1}, f_{X_2}, and f_{X_3} are also identical. Additionally, we can easily check by inspection that $f_{X_1N}(0,1) \neq f_{X_1}(0)f_N(1)$. In fact, $f_{X_iN}(x_i, n_j) \neq f_{X_i}(x_i)f_N(n_j)$ for all $i = 1,2,3$ and $N = 1,2,3$. This means that the random variables X_i and N are <u>not</u> independent. Lastly, the marginal p.m.f. $f_N(n)$ implicitly describes the probability distribution of error-free pages in manuscripts under consideration here.

The notion of joint and marginal distribution holds true for continuous probability distributions as well. Two continuous random variables X and Y have a *joint p.d.f.* $f_{XY}(x,y)$ if for any subset A of $\mathbb{R}^2$, we have

$$P\Big((X,Y) \in A\Big) = \int\int_A f_{XY}(x,y)dxdy, \tag{3.41}$$

where $f_{XY}(x,y)$ is a non-negative function (because it represents a probability measure) and normalizes to unit magnitude upon integration, i.e. $\int\int_{\mathfrak{S}} f_{XY}(x,y)dxdy = 1$, where $\mathfrak{S}$ is an appropriate sample set. The *marginal p.d.f.s* are also defined akin to the discrete counterpart as follows:

$$f_X(x) = \int_y f_{XY}(x,y)dy, \quad f_Y(y) = \int_x f_{XY}(x,y)dx. \tag{3.42}$$

Further, the condition for independence of the random variables X and Y is $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all $(x,y) \in \mathfrak{S}$.

Example: revisiting the problem of total system failure of an aircraft's power plant

Let us reconsider the problem we investigated earlier about estimating the risk of a total system failure after 100 flying hours as explained in the example of an exponential distribution (also **cf.** the example alongside Figure 3.25).[24] The lifetime of each of the two engines (the main power unit (MPU) and the auxiliary power unit (APU)) is represented by two different random variables X and Z with joint p.d.f.

$f_{XZ}(x,z) = \mu^2 e^{-\mu x} e^{-\mu z}$, for $x, z > 0$ and $\mu = \frac{1}{E(X)} = \frac{1}{E(Z)} = \frac{1}{500}$. Let $Y = X + Z$ be the time until the total system failure. What is the p.d.f. of the time until the first engine failure and the total system failure? Further, what is the risk (in terms of a probability measure) of a total system failure after 100 flying hours have elapsed?

The p.d.f. of the time until the MPU failure and the time until the APU failure are the respective marginal p.d.f.s $f_X(x)$and $f_Z(z)$.

$$f_X(x) = \int_0^\infty f_{XZ}(x,z)dz = \mu e^{-\mu x}, \text{ for } x > 0. \tag{3.43}$$

$$\text{Likewise, } f_Z(z) = \mu e^{-\mu z}, \text{ for } z > 0. \tag{3.44}$$

The risk of a total system failure after 100 flying hours is estimated by calculating the right tail distribution

$$\begin{aligned} P(Y > 100) &= 1 - P(Y \leq 100) \\ &= 1 - P(X + Z \leq 100) \\ &= 1 - \int_0^{100} P\Big(X \leq 100 - Z \Big| Z = z\Big) f_Z(z)dz \end{aligned}$$

conditioning & law of total probability

$$\begin{aligned} &= 1 - \int_0^{100} P\Big(X \leq 100 - z\Big) f_Z(z)dz \\ &= 1 - \int_0^{100} \Big(\int_0^{100-z} f_X(x)dx\Big) f_Z(z)dz \\ &= 0.9825. \end{aligned} \tag{3.45}$$

There is a 98.25% risk of a total system failure after 100 flying hours.

[24] In this case, we will not consider an "engine re-start" option. The situation of failure of both power plants during flight would result in total system failure without an opportunity to salvage the crisis.

The concept of joint and marginal distributions (for both discrete and continuous cases) can be extended in an analogous manner for more than two random variables.

Example: application of joint and marginal probability distributions

Let X and Y are independent random variables with distribution $geom_1(p)$. Answer the following questions:

1. What is the probability distribution of $min(X, Y)$?
2. Compute $P(Y \geq X)$.
3. What is the probability distribution of $X + Y$?
4. Compute $P(Y = y | X + Y = z)$ for $z \geq 2$ and $y = 1, 2, ..., z - 1$.

The calculations to solve the above questions are shown here below.

1. Let $Z = min(X, Y)$. If $min(X, Y) > z$, then $X > z$ and $Y > z$. Therefore

$$\begin{aligned} P(Z > z) = P(min(X, Y) > z) &= P(X > z, Y > z) \\ &= P(X > z)P(Y > z) \\ &\quad \nearrow \\ &\quad X \text{ and } Y \text{ are independent} \\ &= (1-p)^z(1-p)^z \\ &\quad \nearrow \\ &\quad \textbf{cf.} \text{ eq. 3.21} \\ &= \left((1-p)^2\right)^z. \end{aligned} \tag{3.46}$$

   This implies that $Z = min(X, Y) \sim geom_1\Big(1 - (1-p)^2\Big)$.

2. Consider the countable events $\{X = 1\} \cap \{Y \geq X\}$, $\{X = 2\} \cap \{Y \geq X\}$, $\{X = 3\} \cap \{Y \geq X\}$, ... that partition the sample space of geometrically distributed random variables X and Y where $Y \geq X$.

$$\begin{aligned} P(Y \geq X) &= P\Big(\bigcup_{x=1}^{\infty}\big(\{X = x\} \cap \{Y \geq X\}\big)\Big) \\ &= \sum_{x=1}^{\infty} P\Big(\{X = x\} \cap \{Y \geq x\}\Big) \\ &\quad \nearrow \\ &\quad \text{probabilities are additive for disjoint events} \\ &= P(X = x)P(Y \geq x) \\ &\quad \nearrow \\ &\quad \text{independence of } X \text{ and } Y \\ &= \sum_{x=1}^{\infty} p(1-p)^{x-1}(1-p)^{x-1} \\ &= \frac{1}{2-p}. \end{aligned} \tag{3.47}$$

3. The distribution for $X + Y$ is prescribed by a convolution sum.

$$\begin{aligned} P(X + Y = z) &= \sum_{x=1}^{z-1} P(X = x, X + Y = z) \\ &= \sum_{x=1}^{z-1} P(X = x, Y = z - x) \\ &= \sum_{x=1}^{z-1} P(X = x)P(Y = z - x) \\ &= (z-1)p^2(1-p)^{z-2}. \end{aligned}$$

4. We use the definition of conditional probability in the following calculation.

$$\begin{aligned} P(Y = y \big| X + Y = z) &= \frac{P(Y = y, X + Y = z)}{P(X + Y = z)} \\ &= \frac{P(X = z - y)P(Y = y)}{P(X + Y = z)} \\ &= \frac{1}{z-1}. \end{aligned} \tag{3.48}$$

3.5 Chapter project: Predicting insurance claim aggregates during a policy period

3.5.1 Interlude: Computing the moments of the compound Poisson distribution and estimating aggregate insurance claims by clients by theoretical analysis

Consider $Y_j = X_1 + X_2 + \cdots + X_{N_j}$ is the aggregate of a random number of claims N_j per quarter (policy period) where $N_j \sim Poisson(\lambda_j)$, $j = 1, 2, 3, 4$ (corresponding to each of four quarters) and $X_i \sim Bernoulli([1, 2], p_2)$ are individual claims with probability $p_1 = \frac{2}{3}$ and $p_2 = \frac{1}{3}$ corresponding to claims denominations of $\$\ 100,000$ and $\$\ 200,000$ respectively. Further, $\lambda_1 = 2$, $\lambda_2 = 3$, $\lambda_3 = 1$, $\lambda_4 = 3$. $Z = \sum_{j=1}^{4} Y_j$ is the yearly total of all claims made to the firm. Answer the following questions.

1. **Identify the distribution of Y_j.**

2. **Compute $E(Z)$ and $Var(Z)$.**

3. **Compute $P(Y_2 > 5)$ and compute $P(Y_3 > 5)$ analytically (without a computer simulation). Subsequently, comment on the discrepancy between the two results (if any).**

Note: A word of caution is in order here. While X_i is indeed a binary random variable, in order for your calculations to work out in accordance with a Bernoulli random process, you may have to carry forth a simple transformation of the observables $[1, 2] \to [0, 1]$. This will be especially necessary while computing the statistical moments such as $E(Z)$!

Figure 3.28: What are the risks in the insurance business?

3.6 Transformation of a random variable

This section is coming soon!

3.7 Moment generating functions and their applications

The moment generating function (m.g.f.) of a random variable X is denoted by $m_X(t)$ and is defined as follows.

$$m_X(t) := E\left(e^{tX}\right) = \begin{cases} \sum_x e^{tx} P(X = x), & \text{if } X \text{ is a discrete r.v.} \\ \int_x e^{tx} f_X(x)dx, & \text{if } X \text{ is a continuous r.v.} \end{cases} \tag{3.49}$$

The following four results underscore the utility of m.g.f. to find probability distributions of sum of random variables and also to find the higher moments of random variables.

1. If $X_1, X_2, ..., X_n$ are n independent random variables, and $Y := X_1 + X_2 + \cdots + X_n$, then

$$m_Y(t) = m_{X_1}(t)m_{X_2}(t)\cdots m_{X_n}(t) = \prod_{i=1}^{n} m_{X_i}(t). \tag{3.50}$$

2. If X and Y are two random variables with finite m.g.f. $m_X(t)$ and $m_Y(t)$ for all t, then X and Y have the same probability distributions.

3. The *joint m.g.f.* of X and Y is

$$m_{X,Y}(s,t) = E\left(e^{sX+tY}\right). \tag{3.51}$$

The following always holds.

$$\text{If } X \text{ and } Y \text{ are independent} \iff m_{X,Y}(s,t) = m_X(s)m_Y(t),\ \forall s,t. \tag{3.52}$$

4. ***Computing higher order moments from m.g.f.***: For all $n \geq 0$, $E\left(X^n\right) = m_X^{(n)}(0)$.[25]

[25] Here $m_X^{(n)}(0)$ is the n^{th} derivative of the m.g.f. of X evaluated at $t = 0$.

Let us consider the following example in order to fully appreciate the utilitarian nature of the above results in the context of: (i) identifying the distribution of sums of random variables, and (ii) computing all the moments of the random variables.

Figure 3.29: Blue cars enter the highway lane at rate λ_1 and red cars enter the lane at λ_2. The lanes merge after sometime whence the rate of flow of cars in the narrow lane is $\lambda_1 + \lambda_2$.

Example: Sum of two Poisson random variable is another Poisson random variable

Consider that X and Y are independent Poisson random variables with rate parameters λ_1 and λ_2 respectively. What is the probability distribution of $X + Y$?

First, let us compute the m.g.f. of X.

$$\begin{aligned} m_X(t) = E\left(e^{tX}\right) &= \sum_{k=0}^{\infty} \frac{e^{tk}e^{-\lambda_1}\lambda_1^k}{k!} \\ &= e^{-\lambda_1}\sum_{k=0}^{\infty} \frac{(\lambda_1 e^t)^k}{k!} \\ &= e^{-\lambda_1}e^{\lambda_1 e^t} \end{aligned}$$

(Taylor series of exponential function)

$$= e^{\lambda_1(e^t-1)}. \tag{3.53}$$

Similarly, $m_Y(t) = e^{\lambda_2(e^t-1)}$.
Now, by using the result of equation 3.50, we obtain $m_{X+Y}(t) = m_X(t)m_Y(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)}$ which is the m.g.f. of a Poisson random variable with rate parameter $\lambda_1 + \lambda_2$. Therefore, $X + Y \sim Poisson(\lambda_1 + \lambda_2)$.

Example: computing moments of random variables

Consider a r.v. X with p.d.f. $f_X(x) = \frac{e^x}{(1+e^x)^2}$ for $-\infty < x < \infty$. Use the m.g.f. to find $E(X)$ and $Var(X)$.

The m.g.f. $m_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{e^x}{(1+e^x)^2} dx = \int_0^1 \left(\frac{1-u}{u}\right)^t du$ for $-1 < t < 1$. Here we have used the transformation $u = \frac{1}{1+e^x}$. Now, recall that $\frac{d}{dt} a^t = e^{at} \log a$ where $a^t = e^{t \log a}$. Using this result, we have $m'_X(t) = \int_0^1 \log(\frac{1-u}{u})(\frac{1-u}{u})^t du$. Thus $m'_X(0) = \int_0^1 \Big(\log(1-u) - \log u\Big) du = 0$. Likewise, $m''_X(0) = \frac{\pi^2}{3}$. This entails that $E(X) = 0$ and $Var(X) = E(X^2) - (E(X))^2 = \frac{\pi^2}{3}$.

3.7.1 *Cumulants and their applications*

This section is coming soon!

3.8 *Asymptotic results*

We will simply state and present to the readers how to apply the important results of this section. Proofs of some these theorems are omitted here in this introductory text. Interested readers may refer to some excellent books mentioned in the chapter bibliography to study the proofs.

First a few clarifications are in order on the issue of *integrability*. If the r.v. X is integrable, i.e. $X \in \mathfrak{L}^1$, then $E(|X|) < \infty$. If the r.v. X is square-integrable, i.e. $X \in \mathfrak{L}^2$, then $E(X^2) < \infty$.

3.8.1 *Markov's inequality*

For a non-negative random variable $X \in \mathfrak{L}^1$, and a constant $c > 0$, we have a following upper bound for the probability tail.

$$P(X \geq c) \leq \frac{E(X)}{c}. \tag{3.54}$$

3.8.2 *Chebyshev's inequality*

If $E(X^2) < \infty$, $k \in \mathbb{R}^+$ and $E(X) = \mu$, $Var(X) = \sigma^2$, then the following are true.

$$P(|X| \geq k) \leq \frac{E(X^2)}{k^2}, \tag{3.55}$$

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}. \tag{3.56}$$

3.8.3 *Law of large numbers*

There are two important theorems on the law of large numbers which forms the basis of many widely used statistical algorithms like the Monte Carlo methods.

- ***Weak Law of Large Numbers (WLLN):*** Let $X_1, X_2, ..., X_n$ be independent and identically distributed (i.i.d.) random variables in $\mathfrak{L}^2$ with mean μ and variance σ^2. $S_n = X_1 + X_2 + \cdots + X_n$. Then for every fixed $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \to 0, \quad \text{as } n \to \infty. \tag{3.57}$$

The WLLN is a direct consequence of the Chebyshev's inequality. This may be easily checked thusly. $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n}$ is also a random variable whose variance is as follows. $Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2}Var(X_1 + \cdots + X_n) = \frac{n\sigma^2}{n^2}$ because all the covariance terms of the form $Cov(X_i, X_j)$, $i \neq j$ are zero. This is due to the independence of the $\{X_i\}$s. Now as $n \to \infty$, $Var\left(\frac{S_n}{n}\right) \to 0$ and the WLLN follows.

- ***Strong Law of Large Numbers (SLLN):*** Let $X_1, X_2, ..., X_n$ are i.i.d. random variables in $\mathfrak{L}^1$ with mean μ. $S_n = X_1 + X_2 + \cdots + X_n$. Then,

$$P\left(\frac{S_n}{n} \to \mu\right) = 1. \tag{3.58}$$

Here it is assumed that the set ω s.t. $\frac{S_n(\omega)}{n} \to \mu$ is an event.

We now illustrate an application of the law of large numbers which will be our first introduction to Monte-Carlo family of simulations.

Figure 3.30: A street magician performing tricks to earn money for a bottle of beer. What are his odds to make enough money?

Example: will the magician collect enough for his evening beer?

A street artist performs a magic show in the main street of Mingletown for three hours. He hopes to earn enough for a bottle of beer which costs INR 350. Throughout the three hours, people give him coins at random. So we will assume that the coins arrive in his bag according to a Poisson distribution. The amount of money each person gives is random with the following distribution.

$$\begin{aligned} P(INR\ 5) &= 0.4 \\ P(INR\ 10) &= 0.4 \\ P(INR\ 20) &= 0.2 \end{aligned}$$

On average, 5 people per hour give money to the street artist. This implies that the Poisson process has intensity $\lambda = 5$. What is the probability the artist accumulates enough money to get his beer? In other words, what is $\hat{I} = P(X_3 \geq 350)$ where X_i is the money accumulated after $i = 1, 2, 3$ hours?

We can estimate this easily using the Monte Carlo simulation which makes use of the

strong law of large numbers. The estimate $\hat{I} = \frac{1}{N}\sum_{j=1}^{N} Z_j$ where Z_j is a Bernoulli random variable with output 0 or 1 depending on whether the i^{th} iteration (out of many thousand iterations) of the Monte Carlo method resulted in the artist bagging enough money for the beer. The law of large number implies that $\hat{I} \to E(Z_j)$ as $n \to \infty$ almost surely. Succinctly, $\hat{I} = P(X_3 \geq 350) = E\left(\mathbb{I}_{X_3 \geq 350}\right)$ where we have used the indicator random variable and a useful result from sec. 2.8.1. The computational estimate of $E\left(\mathbb{I}_{X_3 \geq 350}\right)$ is basically an ensemble averaging process accomplished over many thousand Monte Carlo iterations as is illustrated in the code below.

```
% Monte Carlo Simulation of Compound Poisson process
t = 3; lambda = 5; N = 10^6;
beer = zeros(N,1); beer_price = 350;

for i = 1:N
    n = poissrnd(lambda * t);
    if n~=0
        coins = zeros(n,1);
        for j = 1:n
            U = rand;
            coins(j) = (U <= 2/5)*5 + ...
            (U > 2/5 && U <= 4/5)*10 + (U > 4/5)*20;
        end
    end
beer(i) = (sum(coins) >= beer_price);
end
l_hat = mean(beer)  % l_hat = P(X3 >= beer_price)
relErr_hat = std(beer) / (l_hat * sqrt(N))
% relative error of l_hat by crude Monte Carlo simulation
```

The estimate `l_hat = 8.1000e-05` entails that there is very little chance that the artist will accumulate enough money to buy a bottle of beer. His chance of making enough money for beer could be enhanced if the chance of receiving the largest denomination of money, i.e. INR 20, is increased at the expense of the lower denominations and/or if the frequency at which the people offered money (i.e. the value of λ) increased and/or if he performed his show for a longer duration.

3.8.4 *Central limit theorem (CLT)*

Consider a random sample of size n denoted by $X_1, X_2, ..., X_n$ which is drawn from a population of mean μ (i.e. $E(X_i) = \mu,\ i = 1, 2, ..., n$) and variance σ^2 (i.e. $Var(X_i) = \sigma^2$). Then the asymptotic distribution of $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ as $n \to \infty$ is the standard normal distribution $N(0, 1)$.

We will revisit the central limit theorem in a subsequent chapter on *Statistical Experiments*. Here we will simply use this to arrive at the following to asymptotic results.

3.8.5 *Normal approximation to the Binomial distribution*

Let $X \sim Bin(n,p)$. Then $\frac{X-np}{\sqrt{np(1-p)}} \to N(0,1)$ as $n \to \infty$.

Figure 3.31: How well did I learn from Prof. Bean?

Example: Approval rating of a teacher

A survey is conducted among 100 students in a class to verify if a teacher is performing well. Since the teacher under review is an average student, it turns out that students are equally likely to approve or disapprove of his teaching style. What is the probability that exactly (i) 50 students, and (ii) 75 students approve the teacher's style?

Let us first use the Binomial distribution to compute this probability. Here $n = 100$ and $p = \frac{1}{2}$. Let Y be the number of students who approve the teacher's style. $Y \sim Bin(100, \frac{1}{2})$. Therefore,

$$\begin{aligned} P(Y=50) &= P(Y \leq 50) - P(Y \leq 49) \\ &= 0.0796. \end{aligned} \tag{3.59}$$

The above calculation is accomplished by using the following Matlab command

```
>> binocdf(50,100,0.5) - binocdf(49,100,0.5)
```

Likewise,

$$\begin{aligned} P(Y=75) &= P(Y \leq 75) - P(Y \leq 74) \\ &= 1.9131 \times 10^{-7}. \end{aligned} \tag{3.60}$$

Now, let us attempt to estimate the same quantity by using the Normal approximation $Y \approx N(\mu, \sigma^2)$ where $\mu = np = 100 \times 0.5 = 50$ and $\sigma^2 = np(1-p) = 50 \times 0.5 = 25$. Applying what is known as the *continuity correction*, $P(Y = 50) = P(49.5 < Y < 50.5)$, we proceed as follows:

$$\begin{aligned} P(49.5 < Y < 50.5) &= P\left(\frac{49.5-50}{5} < \frac{Y-\mu}{\sigma} < \frac{50.5-50}{5}\right) \\ &= P(-0.1 < Z < 0.1), \text{ where } Z \sim N(0,1) \text{ as a result of the CLT,} \\ &= 0.0797. \end{aligned} \tag{3.61}$$

The above calculation of the probability for the standard normal distribution is accomplished using Matlab commands given below.

```
mu = 0;
sigma = 1; % for Z ~ N(0,1)
pd = makedist('Normal','mu',mu,'sigma',sigma);
probability = cdf(pd,0.1) - cdf(pd,-0.1)
```

Likewise,

$$\begin{aligned} P(Y = 75) &= P(74.5 < Y < 75.5) \\ &= P(\frac{24.5}{5} < Z < \frac{25.5}{5}) \\ &= 3.09 \times 10^{-7}. \end{aligned} \tag{3.62}$$

A close inspection of the above calculation reveals that the normal approximation is a reasonable estimate of the binomial distributed random process. The approximation is exceptionally good near the mean of the distribution where most of the probability mass is concentrated.

3.8.6 *Normal approximation to the Poisson distribution*

Let $X \sim Poisson(\lambda)$. Then $\frac{X-\lambda}{\sqrt{\lambda}} \to N(0,1)$ as $\lambda \to \infty$.

The annual number of earthquakes registering at least 3.5 on the Richter Scale and having an epicenter within 50 kms of downtown Colombo follows a Poisson distribution with mean 6.5. What is the probability that at least 10 such earthquakes will strike next year?[26]

Let $X \sim Poisson(\lambda = 6.5)$ be the random number of earthquakes that strikes Colombo on a given year. $P(X \geq 10) = 1 - P(X \leq 9) = 0.1226$. The Matlab calculations for the above are done as follows.

```
lambda = 6.5;
pd = makedist('Poisson','lambda',lambda);
prob = 1 - cdf(pd,9);
```

Now, even though λ is not very large here, the Poisson estimate can be reasonably approximated by the Normal distribution as shown below.

$$\begin{aligned} P(X \geq 10) &\overset{\text{continuity correction}}{=} P(X > 9.5) \\ &= P\left(\frac{X-\lambda}{\sqrt{\lambda}} > \frac{9.5-6.5}{\sqrt{6.5}}\right) \\ &= P(Z > 1.1767), \text{ where } Z \sim N(0,1), \\ &= 0.119. \end{aligned} \tag{3.63}$$

The above is computed in Matlab as follows.

```
mu=0;
sigma = 1;
pd = makedist('Normal','mu',mu,'sigma',sigma);
probability = 1 - cdf(pd,(9.5-6.5)/(sqrt(6.5)));
```

[26] This example is adapted from the textbook by Richard J. Larsen and Morris L. Marx, *An Introduction to Mathematical Statistics and its Applications*, sixth edition, Pearson, 2018.

3.8.7 *Poisson approximation to the Binomial distribution*

Consider $S_n \sim Bin(n, p_n)$ s.t. $p_n \to 0$, $np_n \to \lambda$ as $n \to \infty$, then $S_n \sim Poisson(\lambda)$ in the asymptotic limit.

Example: application of Poisson approximation to a binomial distributed random process

Suppose we roll two dice 24 times and X be the number of times a pair of aces appear. Compute the following: $P(X = 0)$, $P(X = 2)$.

Here $n = 24$ and $p = \frac{1}{36}$ which means $np = \frac{2}{3} = \lambda$. Estimation of $P(X = k)$ for $k = 0, 2$ is shown below using the Binomial distribution and the corresponding Poisson approximation is shown alongside.

	Binomial distribution $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	Poisson approximation $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$
$P(X = 0)$	$\left(1 - \frac{1}{36}\right)^{24} = 0.5086$	$e^{-2/3} = 0.5134$
$P(X = 2)$	$\binom{24}{2}(\frac{1}{36})^2\left(1 - \frac{1}{36}\right)^{22} = 0.1146$	$e^{-2/3}(\frac{2}{3})^2 \frac{1}{2!} = 0.1141$

From the probability estimates tabulated above, we may infer that the Poisson approximation is very close to the predictions of the binomial distribution model.

3.8.8 *DeMoivre-Laplace limit theorem and its application*

This section is coming soon!

3.9 *Chapter project: Predicting insurance claim aggregates during a policy period*

3.9.1 *Prologue: Predicting risk of monetary loss associated with the insurance scheme for the company using a Monte Carlo simulation*

In this section, use the *crude* Monte Carlo simulation (and thereby the law of large numbers) to predict the following.

1. **Estimate $P(Y_2 > 5)$ and $P(Y_3 > 5)$ using the crude Monte Carlo simulation. Compare your simulation results here with the analytical results you obtained in section 3.5.1. Comment on your comparisons.**

2. **Let the total annual income on the sale of insurance premiums be $\$\, 1,000,000$. What is the risk of yearly loss for the company in terms of $P(Z > 1,000,000)$? You may provide your analysis of the risk by using an appropriate Monte Carlo simulation.**

3.10 Selected bibliography

1. *An Introduction to Probability Theory and Its Applications: Vol. 1* by *William Feller*, John Wiley & Sons, Inc. (third edition), 1968.
2. *An Introduction to Probability Theory and Its Applications: Vol. 2* by *William Feller*, John Wiley & Sons, Inc. (second edition), 1971.
3. *Theory of Probability* by *Bruno de Finetti*, John Wiley & Sons, Inc. (special edition), 2017.
4. *Probability: A Lively Introduction* by *Henk Tijms*, Cambridge University Press (twelfth edition), 2019.
5. *Elementary Probability for Applications* by *Rick Durrett*, Cambridge University Press (first edition), 2009.
6. *Weighing the Odds* by *David Williams*, Cambridge University Press (first edition), 2010.
7. *Fundamentals of Mathematical Statistics* by *S. C. Gupta* and *V. K. Gupta*, Sultan Chand and Sons (eleventh edition), 2017.
8. *Introduction to Probability Models* by *Sheldon M. Ross*, Academic Press (Elsevier) (twelfth edition), 2019.
9. *Elements of Distribution Theory* by *Thomas A. Severini*, Cambridge University Press (first edition), 2005.
10. *Theory of Distributions - a non-technical introduction* by *Ian Richards* and *Heekyung Youn*, Cambridge University Press (first edition), 2007.

3.11 Exercise problems

1. (***Expectation of positive integer valued random variables***) Let X be a random variable with values on the positive integer set. Prove that $E(X) = \sum_{k=1}^{\infty} P(X \geq k) \leq \infty$.
2. Consider a random variable X with p.d.f. $f_X(x) = 6x(1-x)$ for $x \in (0,1)$ and zero otherwise. Find $E(X)$ and $Var(X)$.
3. (***Needle and*** π) A famous attempt of estimating the value of π was made by repeatedly dropping a needle of length $L \leq 1$ on the floor made of long continuous tiles of unit width. This was done by estimating the probability that the dropped needle would touch a tile edge and then comparing (equating) this answer with computer simulated experiments of multiple needles dropped on such a tiled floor. What is the probability that a dropped needle would touch a tile edge?
4. (***Stuck in traffic!***) Cars start successively at the origin and travel at different but constant speeds along an infinite narrow road on which no passing is possible. When a car reaches a slower car it is compelled to follow it at the same speed. In this way platoons will be formed whose ultimate size depends on the speeds of the cars but not on the times between successive departures. Consider the speeds of the cars to be independent random variables with a common continuous distribution. Choose a car at random. Prove the following.

(i) The probability that the chosen car does not trail any other car tends to half.

(ii) The probability that it leads a platoon of total size n (with exactly $n-1$ cars following it) tends to $\frac{1}{(n+1)(n+2)}$.

(iii) The probability that the given car is the last in a platoon of size n tends to the same limit as in (ii).

5. (***Why do two bass guitars sound twice as loud as one?***) Consider the sound wave emanating from the guitar is given by a random unit vector. Waves coming from two independent guitars would superimpose to give a resultant vector.

(i) Use the *law of the cosines* to write the square of the length of the resulting vector as $2+2\cos\theta$, where θ is the angle between the two random vectors.

(ii) Predict an appropriate distribution for θ.

(iii) Calculate the expected value of *loudness*[27] of the resulting vector.

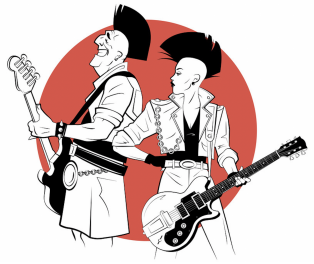

Figure 3.32: Two bass guitarists in full song!

[27] Loudness is proportional to square of the amplitude of vibration.

6. (***How long will I bemoan my bad luck at the toll plaza?***) While on a road trip through the Grand Himalayan Highway, I come across a toll plaza with multiple parallel counters. I am in a dilemma - *which of the several counter options do I exercise so that I can get off the queue as fast as I can?* Often in such situations, the car right behind (car-x) becomes a marker of my decision - *assuming the car right behind me chose a different counter, my assessment of my luck is determined by who among the two of us stays ahead in the queue?* To make the problem simpler to solve, we will consider the following simplifications:

(a) all cars in the toll plaza are of equal size,

(b) all queues are stochastically independent,

(c) time interval between successive moves in any queue is an independent continuous random variable with a common probability distribution,

(d) successive moves constitute Bernoulli trials where "success" means I move ahead by a unit distance and "failure" means car-x moves ahead on a certain trial,

(e) the probability of success is $p = 0.5$.

Answer the following questions.

(i) What are my odds of getting ahead of car-x in the queue while I am at the toll plaza?

(ii) If I do end up being ahead of car-x at some point, what is my expected waiting time before I am ahead?[28]

(iii) In what manner would the answers to the above two questions differ if $p \neq 0.5$?

Figure 3.33: The Grand Himalayan toll plaza (*courtesy: The Times of India).*

[28] The longer I have to wait to get ahead of car-x, the longer I will end up bemoaning my bad luck at the toll plaza!

7. (***Twin-engine failure***) The $\mathfrak{M}$-09 is a twin engine jet. Let X be the random time to failure of engine-1 and Y be the random time to failure of engine-2. X and Y are independent random variables with distribution $exp(\mu_1)$ and $exp(\mu_2)$ with mean times to failure $\frac{1}{\mu_1} = \frac{1}{\mu_2} = 100$ flying hours. What is the probability that there is a dual engine flame out in more than 75 flying hours?

Figure 3.34: What are the odds of a dual engine flame out?

8. Let X and Y be independent random variables each with an exponential distribution $exp(\lambda)$. What is the p.d.f. of the following random variables?

(i) $Z = X + Y$.

(ii) $W = Y - X^2$.

9. (***Application of moment generating function***) Consider a random variable X with p.d.f. $f_X(x) = e^{-2x} + \frac{1}{2}e^{-x},\ x > 0$. Find the m.g.f. of X and use it to find the $Var(X)$.

10. (***Generating functions for branching process***) Branching processes have wide applications. For example, in the area of biology, a cell may die or split into two in each generation leading to the survival or extinction of the species. We may be interested in knowing the chances of survival of the species after n generations. The extinction probability $u_n := P(X_n = 0)$ where X_n is the size of the n^{th} generation. The extinction probability u_n can be computed iteratively by using the *probability generating function* (p.g.f.): $P(z) := E(z^X) = \sum_{j=0}^{\infty} p_j z^j$, where $0 \leq z \leq 1$ and p_j is the probability $P(X = j)$, that in the lifetime of an individual cell, it produces $j = 0, 1, 2, ...$ new offspring.

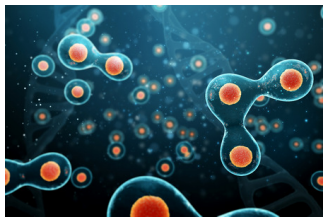

Figure 3.35: Cell division.

Thence

$$u_n = P(u_{n-1}) \text{ for } n = 2, 3, 4, ... \tag{3.64}$$

with $u_0 = 0$ and $u_1 = p_0$. It follows that in the asymptotic limit $n \to \infty$, u_∞ satisfies the fixed point iteration $u = P(u)$ and is its smallest positive root. Answer the following questions about a branching process.

(a) Deduce the equation 3.64 by using the definition of the probability generating function.[29]

(b) A carcinoma begins with a single cell. In each generation, a carcinogenic cell dies with a probability $\frac{1}{3}$ or doubles with probability $\frac{2}{3}$. What is the probability that the cancer will die out in the third generation? What is the probability that the cancer will proliferate forever? What are the above probabilities if the cancer initially began with two cells?

[29] You may have to deduce an intermediary step $u_n = \sum_{k=0}^{\infty} u_{n-1}^k p_k$, for $n = 2, 3, ...$ and use the law of total probability.