1. Mean and Variance

Let X be a random variable with distribution F_X. We shall assume that it has pmf (or, pdf) denoted by f_X.

Definition 1. The *expected value* (also called *mean*) of X is defined as the quantity $\mathbf{E}[X] = \sum_t t f(t)$ if f is a pmf, and $\mathbf{E}[X] = \int_{-\infty}^{+\infty} t f(t) dt$ if f is a pdf (provided the sum, or the integral converges absolutely).

In other words,

$$\mathbf{E}[X] = \begin{cases} \sum_x x f(x), & \text{if } \sum_x |x| f(x) < \infty \text{ for discrete } X, \\ \int_{-\infty}^{+\infty} x f(x), & \text{if } \int_{-\infty}^{+\infty} |x| f(x) < \infty \text{ for continuous } X. \end{cases}$$

Note that it is possible to define expected value for distributions without pmf or pdf, but we shall not do it here.

Exercise 2. Find the expectation of random variables with the following pdf/pmf:

(a) $f(x) = \frac{1}{\pi(1+x^2)}$ when $x \in \mathbb{R}$,

(b) $f(x) = \frac{1}{|x|(1+|x|)}$ when $x \in S = \{(-1)^n n | n \in \mathbb{N}\}$.

Properties of expectation: Let X, Y be random variables both having pmf f, g (or, pdf f, g).

(1) Then, $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$ for any $a, b \in \mathbb{R}$. In particular, for a constant random variable (i.e., $X = a$ with probability 1 for some a, $\mathbf{E}[X] = a$). This is called *linearity* of expectation.

(2) If $X \geq Y$ (meaning, $X(\omega) \geq Y(\omega)$ for all ω), then $\mathbf{E}[X] \geq \mathbf{E}[Y]$.

(3) If $\varphi : \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}[\varphi(X)] = \begin{cases} \sum\limits_t \varphi(t) f(t) & \text{if } f \text{ is a pmf} \\ \int_{-\infty}^{+\infty} \varphi(t) f(t) dt & \text{if } f \text{ is a pdf,} \end{cases}$$

provided they exist (i.e., the sum, or the integral converges absolutely).

For random variables on a discrete probability space (they have a pmf), we can easily prove all these properties. For random variables with pdf, a proper proof require a bit of work.

Note: Expectation is a very important quantity. Using it, we can define several other quantities of interest.

Discussion: For simplicity, let us take random variables to have densities in this discussion. You may adapt the remarks to the case of pmf easily. The density has all the information we need about a random variable. However, it is a function, which means that we have to know $f(t)$ for every t. In real life, we often have random variables whose pdf is unknown, or impossible to determine. It would be better to summarize the main features of the distribution (i.e., the density) in a few numbers. That is what the quantities defined below try to do.

Mean: Mean is another term for expected value.

Moments: The quantity $\mu'_k = \mathbf{E}[X^k]$ (if it exists) is called the k^{th} *moment* of X for $k \in \{1, 2, \ldots\}$.

Central Moments: The quantity $\mu_k = \mathbf{E}[(X - \mu)^k]$ (if it exists) is called the k^{th} *central moment* of X for $k \in \{1, 2, \ldots\}$. Here, and henceforth $\mu = \mu'_1$ denotes the mean.

Variance: Let $\mu = \mathbf{E}[X]$ and define $\sigma^2 := \mathbf{E}\left[(X - \mu)^2\right]$. This is called the *variance* of X, also denoted by $\text{Var}(X)$. It can be written in other forms. For example,

$$\begin{aligned}
\sigma^2 &= \mathbf{E}[X^2 + \mu^2 - 2\mu X] \qquad \text{(by expanding the square)} \\
&= \mathbf{E}[X^2] + \mu^2 - 2\mu\mathbf{E}[X] \qquad \text{(by property (1) above)} \\
&= \mathbf{E}[X^2] - \mu^2.
\end{aligned}$$

That is $\text{Var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$.

Standard deviation: The standard deviation of X is defined as $\text{s.d.}(X) := \sqrt{\text{Var}(X)}$.

Computing expectations from the pmf: Let X be a random variable on (Ω, p) with pmf f. Then, we claim that

$$\mathbf{E}[X] = \sum_{t \in \mathbb{R}} t f(t).$$

Indeed, let $\text{Range}(X) = \{x_1, x_2, \ldots\}$. Let $A_k = \{\omega : X(\omega) = x_k\}$. By the definition of pmf, we have $\mathbf{P}(A_k) = f(x_k)$. Further, A_k are pairwise disjoint and exhaustive. Hence

$$\mathbf{E}[X] = \sum_{\omega \in \Omega} X(\omega) p_\omega = \sum_k \sum_{\omega \in A_k} X(\omega) p_\omega = \sum_k x_k \mathbf{P}(A_k) = \sum_k x_k f(x_k).$$

Similarly, $\mathbf{E}[X^2] = \sum_k x_k^2 f(x_k)$.

Remark 3. More generally, if $h : \mathbb{R} \to \mathbb{R}$ is any function, then the random variable $h(X)$ has expectation $\mathbf{E}[h(X)] = \sum_k h(x_k) f(x_k)$. Although this sounds trivial, there is a very useful point here. To calculate $\mathbf{E}[X^2]$ we do not have to compute the pmf of X^2 first, which can be done but would be more complicated. Instead, in the above formulas, $\mathbf{E}[h(X)]$ has been computed directly in terms of the pmf of X.

Exercise 4. Find $\mathbf{E}[X]$, $\mathbf{E}[X^2]$ and $\text{Var}(X)$ in each case.

(1) $X \sim \text{Bin}(n, p)$.

(2) $X \sim \text{Geo}(p)$.

(3) $X \sim \text{Pois}(\lambda)$.

(4) $X \sim \text{Hypergeo}(b, w, m)$.

Example 5. Let $X \sim N(\mu, \sigma^2)$. Recall that its density is $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. We can compute

$$\mathbf{E}[X] = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \mu.$$

On the other hand

$$\begin{aligned}
\text{Var}(X) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}(x-\mu)^2e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \;&=\; \sigma^2\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}u^2e^{-\frac{u^2}{2}}\,du \qquad (\text{substitute } x = \mu + \sigma u)\\
= \sigma^2\frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty}u^2e^{-\frac{u^2}{2}}\,du \;&=\; \sigma^2\frac{2\sqrt{2}}{\sqrt{2\pi}}\int_{0}^{+\infty}\sqrt{t}e^{-t}\,dt \qquad (\text{substitute } t = u^2/2)\\
= \sigma^2\frac{2\sqrt{2}}{\sqrt{2\pi}}\Gamma(3/2) \;&=\; \sigma^2.
\end{aligned}$$

To get the last line, observe that $\Gamma(3/2) = \frac{1}{2}\Gamma(1/2)$ and $\Gamma(1/2) = \sqrt{\pi}$. Thus, we now have a meaning for the parameters μ and σ^2 - they are the mean and variance of the $N(\mu, \sigma^2)$ distribution. Again, note that the mean is the same for all $N(0, \sigma^2)$ distributions but the variances are different, capturing the spread of the distribution.

Exercise 6. Let $X \sim N(0, 1)$. Show that $\mathbf{E}[X^n] = 0$ if n is odd. If n is even, then $\mathbf{E}[X^n] = (n-1)(n-3)\cdots(3)(1)$ (product of all odd numbers upto and including $n-1$). What happens if $X \sim N(0, \sigma^2)$?

Exercise 7. If X is a non-negative integer-valued random variable with finite expectation, then

$$\mathbf{E}(X) = \sum_{k=1}^{\infty} \mathbf{P}(X \geq k)$$

Exercise 8. Find $\mathbf{E}[X]$, $\mathbf{E}[X^2]$ and $\mathrm{Var}(X)$ in each case.

(1) $X \sim \mathrm{Exp}(\lambda)$.

(2) $X \sim \mathrm{Gamma}(\nu, \lambda)$.

(3) $X \sim \mathrm{Unif}([0, 1])$.

(4) $X \sim \mathrm{Beta}(p, q)$.

Exercise 9. Show that $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$ (hence, $\mathrm{s.d.}(cX) = |c|\mathrm{s.d.}(X)$) for $c \in \mathbb{R}$.

2. Descriptive Measures of Probability Distributions

Mean: Recall that the mean of a random variable (probability distribution) X is given by $\mu = \mathbf{E}(X)$.

Median: A real number m satisfying

$$F_X(m-) \leq \frac{1}{2} \leq F_X(m), \text{ i.e., } \mathbf{P}(\{X < m\}) \leq \frac{1}{2} \leq \mathbf{P}(\{X \leq m\})$$

is called the median (of the probability distribution) of X.

Note: Let us assume that the CDF F_X of X is *strictly increasing and continuous*. Then $F_X^{-1}(t)$ is well defined for every $t \in (0,1)$. For each $t \in (0,1)$, the number $Q_t = F_X^{-1}(t)$ is called the t-quantile. For example, the $1/2$-quantile, also called *median* is the number x such that $F_X(x) = \frac{1}{2}$ (unique when the CDF is strictly increasing and continuous). Similarly, one defines $1/4$-quantile and $3/4$-quantile and these are sometimes called quartiles.[1]

Mode: The mode m_0 of a probability distribution is the value that occurs with highest pmf/pdf, and is defined by $f_X(m_0) = \sup\{f_X(x) : x \in S_X\}$.

Measures of central tendency: Mean and median try to summarize the distribution of X by a single number. Of course one number cannot capture the whole distribution, so there are many densities and mass functions that have the same mean or median. Which is better - mean or median? This question has no unambiguous answer. Mean has excellent mathematical properties (mainly linearity) which the median lacks (med$(X+Y)$ bears no general relationship to med$(X)+$ med(Y)). In contrast, mean is sensitive to outliers, while the median is far less so. For example, if the average income in a village of 50 people is Rs.1000 per month, the immigration of multi-millionaire to the village will change the mean drastically, but the median remains about the same. This is good, if by giving one number we are hoping to express the state of a typical individual in the population.

[1] Another familiar quantity is the percentile, frequently used in reporting performance in competitive exams. For each x, the x-percentile is nothing but $F(x)$. For exam scores, it tells the proportion of exam-takers who scored less than or equal to x.

Coefficient of variation: The coefficient of variation of X is defined as c.v.$(X) = \frac{\text{s.d.}(X)}{|\mathbf{E}[X]|}$.

Mean absolute deviation (m.a.d.): The mean absolute deviation of X is defined as the $\mathbf{E}[|X - \text{med}(X)|]$.

Quartile deviation: Let q_1 and q_3 be real numbers such that

$$F_X(q_1-) \leq \frac{1}{4} \leq F_X(q_1) \quad \text{and} \quad F_X(q_3-) \leq \frac{3}{4} \leq F_X(q_3)$$

i.e., $\mathbf{P}(\{X < q_1\}) \leq \frac{1}{4} \leq \mathbf{P}(\{X \leq q_1\})$ and $\mathbf{P}(\{X < q_3\}) \leq \frac{3}{4} \leq \mathbf{P}(\{X \leq q_3\})$. The quantities q_1 and q_3 are called, respectively, the lower and upper quartiles of the probability distribution of random variable X.
The quartile deviation (or, inter-quartile range) is defined as $q_3 - q_1$.

Measures of dispersion: Suppose the average height of people in a city is 160 cm. This could be because everyone is 160 cm exactly, or because half the people are 100 cm. While the other half are 220 cm., or alternately the heights could be uniformly spread over 150-170 cm., etc. How widely the distribution is spread is measured by standard deviation and mean absolute deviation. Since we want deviation from mean, $\mathbf{E}[X - \mathbf{E}[X]]$ looks natural, but this is zero because of cancellation of positive and negative deviations. To prevent cancellation, we may put absolute values (getting to the m.a.d, but that is usually taken around the median) or we may square the deviations before taking expectation (giving the variance, and then the standard deviation). Variance and standard deviation have much better mathematical properties, and hence are usually preferred.

The standard deviation has the same units as the quantity. For example, if mean height is 160cm measured in centimeters with a standard deviation of 10cm, and the mean weight is 55kg with a standard deviation of 5kg, then we cannot say which of the two is less variable. To make such a comparison we need a dimension free quantity (a pure number). Coefficient of variation is such a quantity, as it measure the standard deviation per mean. For the height and weight data just described, the coefficients of variation are 1/16 and 1/11, respectively. Hence, we may say that height is less variable than weight in this example.

Standardization: Let $\mu = \mathbf{E}[X]$ and $\sigma = \text{s.d.}(X)$, i.e., the mean and the standard deviation of X, respectively. Define $Z = \frac{(X-\mu)}{\sigma}$ to be the standardized variable (independent of units).

Skewness: A measure of skewness of the probability distribution of X is defined by

$$\beta_1 = \mathbf{E}\left(Z^3\right) = \frac{\mathbf{E}\left((X-\mu)^3\right)}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}.$$

Kurtosis: A measure of kurtosis of the probability distribution of X is defined by

$$\gamma_1 = \mathbf{E}\left(Z^4\right) = \frac{\mathbf{E}\left((X-\mu)^4\right)}{\sigma^4} = \frac{\mu_4}{\mu_2^2}.$$

The quantity γ_1 is simply called the kurtosis of the probability distribution of X. It is easy to show that for any values of $\mu \in \mathbb{R}$ and $\sigma > 0$, the kurtosis of $N\left(\mu, \sigma^2\right)$ distribution is $\gamma_1 = 3$. The quantity $\gamma_2 = \gamma_1 - 3$ is called the *excess kurtosis* of the distribution of X.

Exercise 10. Show that for any values of $\mu \in \mathbb{R}$ and $\sigma > 0$, the kurtosis of $N\left(\mu, \sigma^2\right)$ distribution is $\gamma_1 = 3$. This implies that $\gamma_2 = 0$.

Moment generating function: We are familiar with the Laplace transform of a given real-valued function defined on $\mathbb{R}$. We also know that under certain conditions, the Laplace transform of a function determines the function almost uniquely. In probability theory, the Laplace transform of a pdf/pmf of a random variable X plays an important role and is referred to as moment generating function (of probability distribution) of the random variable X.

Define $M_X : A \to \mathbb{R}$ by

$$M_X(t) = \mathbf{E}[e^{tX}], \quad t \in A.$$

We call M_X the moment generating function (mgf) of the random variable X. We say that the mgf of a random variable X exists if there exists a positive real number a such that $(-a, a) \subseteq A$ (i.e., if $M_X(t) = \mathbf{E}[e^{tX}]$ is finite in an interval containing 0).

Note that $M_X(0) = 1$, and therefore, $A = \{t \in \mathbb{R} : \mathbf{E}[e^{tX}] \text{ is finite}\} \neq \emptyset$. Moreover, we have $M_X(t) > 0 \;\forall\, t \in A$. The name moment generating function to the transform M_X is derived from the fact that M_X can be used to generate moments of random variable X. Let X be a random variable with mgf M_X, which is finite on an interval $(-a, a)$, for some $a > 0$ (i.e., mgf of X exists). Then, we have the following:

(i) $\mu'_r = \mathbf{E}[X^r]$ is finite for each $r \in \{1, 2, \dots\}$,

(ii) $\mu'_r = \mathbf{E}[X^r] = M_X^{(r)}(0)$, where $M_X^{(r)}(0) = \left[\frac{d^r}{dt^r} M_X(t)\right]_{t=0}$ the r-th derivative of $M_X(t)$ at the point 0 for each $r \in \{1, 2, \dots\}$, and

(iii) $M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu'_r$ with $t \in (-a, a)$.

Proposition 11. *Under the notation and assumptions of the theorem define $\psi_X : (-a, a) \to \mathbb{R}$ by $\psi_X(t) = \ln M_X(t), t \in (-a, a)$. Then*

$$\mu'_1 = \mathbf{E}[X] = \psi_X^{(1)}(0) \quad \textit{and} \quad \mu_2 = \mathit{Var}(X) = \psi_X^{(2)}(0),$$

where $\psi_X^{(r)}$ denotes the r-th $(r \in \{1, 2\})$ derivative of ψ_X.

Proof. We have, for $t \in (-a, a)$

$$\psi_X^{(1)}(t) = \frac{M_X^{(1)}(t)}{M_X(t)} \quad \text{and} \quad \psi_X^{(2)}(t) = \frac{M_X(t) M_X^{(2)}(t) - \left(M_X^{(1)}(t)\right)^2}{(M_X(t))^2}.$$

■

Example 12. Let X be a random variable with pmf

$$f_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}, & \text{if } x \in \{0,1,2,\ldots\} \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$.
(i) Find the mgf $M_X(t)$ for $t \in A = \left\{s \in \mathbb{R} : \mathbf{E}[e^{sX}] < \infty\right\}$ of X. Show that X possesses moments of all orders. Find the mean and variance of X;
(ii) Find $\psi_X(t) = \ln\left(M_X(t)\right)$ for $t \in A$. Hence, find the mean and variance of X;
(iii) What are the first four terms in the power series expansion of M_X around the point 0?

(i) We have

$$M_X(t) = \mathbf{E}\left[e^{tX}\right] = \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda\left(e^t-1\right)}\ \forall\, t \in \mathbb{R},$$

since $A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\} = \mathbb{R}$. For every $r \in \{1,2,\ldots\}$, $\mu_r' = \mathbf{E}\left(X^r\right)$ is finite. Clearly,

$$M_X^{(1)}(t) = \lambda e^t e^{\lambda\left(e^t-1\right)} \quad \text{and} \quad M_X^{(2)}(t) = \lambda e^t e^{\lambda\left(e^t-1\right)}\left(1+\lambda e^t\right)\ \forall\, t \in \mathbb{R}.$$

Therefore,

$$\begin{aligned} \mathbf{E}(X) &= M_X^{(1)}(0) = \lambda, \\ \mathbf{E}\left(X^2\right) &= M_X^{(2)}(0) = \lambda(1+\lambda) \text{ and} \\ \mathrm{Var}(X) &= \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = \lambda. \end{aligned}$$

(ii) We have, for $t \in \mathbb{R}$

$$\psi_X(t) = \ln\left(M_X(t)\right) = \lambda\left(e^t-1\right) \Rightarrow \quad \psi_X^{(1)}(t) = \psi_X^{(2)}(t) = \lambda e^t.$$

Therefore,

$$\mathbf{E}(X) = \psi_X^{(1)}(0) = \lambda \quad \text{and} \quad \mathrm{Var}(X) = \psi_X^{(2)}(0) = \lambda.$$

(iii) We have

$$\begin{aligned} & M_X^{(3)}(t) = \lambda e^t e^{\lambda\left(e^t-1\right)}\left(\lambda^2 e^{2t} + 3\lambda e^t + 1\right)\ \forall\, t \in \mathbb{R} \\ \Rightarrow\quad & \mu_3' = \mathbf{E}\left(X^3\right) = M_X^{(3)}(0) = \lambda\left(\lambda^2 + 3\lambda + 1\right). \end{aligned}$$

Since $A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\} = \mathbb{R}$, we have

$$\begin{aligned} M_X(t) &= 1 + \mu_1' t + \mu_2'\frac{t^2}{2!} + \mu_3'\frac{t^3}{3!} + \cdots \\ &= 1 + \lambda t + \lambda(\lambda+1)\frac{t^2}{2!} + \lambda\left(\lambda^2+3\lambda+1\right)\frac{t^3}{3!} + \cdots, t \in \mathbb{R}. \end{aligned}$$

Example 13. Let X be a random variable with pdf

$$f_X(x) = \begin{cases} e^{-x}, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(i) Find the mgf $M_X(t)$ for $t \in A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\}$ of X. Show that X possesses moments of all orders. Find the mean and variance of X.
(ii) Find $\psi_X(t) = \ln\left(M_X(t)\right)$ for $t \in A$. Hence, find the mean and variance of X.
(iii) Expand M_X as a power series around the point 0 and hence find $\mathbf{E}\left(X^r\right)$ with $r \in \{1, 2, \ldots\}$.

(i) We have

$$M_X(t) = E\left(e^{tX}\right) = \int_0^\infty e^{tx} e^{-x}\,\mathrm{d}x = \int_0^\infty e^{-(1-t)x}\,\mathrm{d}x < \infty, \text{ if } t < 1.$$

Clearly, $A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\} = (-\infty, 1) \supset (-1, 1)$ and $M_X(t) = (1-t)^{-1}$ for $t < 1$. For every $r \in \{1, 2, \ldots\}$, μ_r' is finite. Clearly,

$$M_X^{(1)}(t) = (1-t)^{-2} \quad \text{and} \quad M_X^{(2)}(t) = 2(1-t)^{-3} \text{ for } t < 1.$$

So, we get

$$\begin{aligned} \mathbf{E}(X) &= M_X^{(1)}(0) = 1, \\ \mathbf{E}\left(X^2\right) &= M_X^{(2)}(0) = 2 \text{ and} \\ \operatorname{Var}(X) &= \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = 1. \end{aligned}$$

(ii) For $t < 1$, we have

$$\psi_X(t) = \ln\left(M_X(t)\right) = -\ln(1-t) \Rightarrow \quad \psi_X^{(1)}(t) = \frac{1}{1-t} \text{ and } \psi_X^{(2)}(t) = \frac{1}{(1-t)^2}.$$

So, we get

$$\mathbf{E}(X) = \psi_X^{(1)}(0) = 1 \quad \text{and} \quad \operatorname{Var}(X) = \psi_X^{(2)}(0) = 1.$$

(iii) We have

$$M_X(t) = (1-t)^{-1} = \sum_{r=0}^{\infty} t^r, \quad \text{for } t \in (-1, 1),$$

since $A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\} = (-\infty, 1) \supset (-1, 1)$. So, we conclude that $\mu_r' =$ coefficient of $\frac{t^r}{r!}$ in the power series expansion of M_X around 0.

Exercise 14. Let X be a random variable with pdf

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, -\infty < x < \infty$$

Show that the mgf of X does not exist.

Identically Distributed Random Variables: Two random variables X and Y are said to have the same distribution (written as $X \stackrel{D}{=} Y$) if they have the same distribution function, i.e., if

$$F_X(x) = F_Y(x) \; \forall \; x \in \mathbb{R}.$$

(i) Let X and Y be random variables of discrete type with pmf f_X and f_Y, respectively. Then $X \stackrel{D}{=} Y$ if and only if $f_X(x) = f_Y(x) \; \forall \; x \in \mathbb{R}$.

(ii) Let X and Y be random variables of continuous type with pdf f_X and f_Y, respectively. Then $X \stackrel{D}{=} Y$ if and only if $f_X(x) = f_Y(x) \; \forall \; x \in \mathbb{R}$.

(iii) Let X and Y be random variables having mgfs M_X and M_Y, respectively. Suppose that there exists a positive real number b such that $M_X(t) = M_Y(t) \; \forall \; t \in (-b, b)$. Then, $X \stackrel{D}{=} Y$.

Remark 15. The idea of two random variables being identical in distribution is different from two random variables being identical (equal) pointwise. Which notion is stronger?

Symmetric Distribution: A random variable X is said to have a symmetric distribution about a point $\mu \in \mathbb{R}$ if $X - \mu \stackrel{D}{=} \mu - X$.

Exercise 16. Check that the $N(\mu, \sigma^2)$ distribution is symmetric about the point μ. Find the point of symmetry of Bin$(n, 1/2)$ distrbution, when n is a positive, even integer.

4. MARKOV'S AND CHEBYSHEV'S INEQUALITIES

Let X be a non-negative integer valued random variable with pmf $f(k)$ for $k = 0, 1, 2, \ldots$. Fix any number m, say $m = 10$. Then

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} 10f(k) = 10\mathbf{P}\{X \geq 10\}.$$

More generally, $m\mathbf{P}\{X \geq m\} \leq \mathbf{E}[X]$.

Markov's inequality: Let X be a non-negative random variable with finite expectation. Then, for any $t > 0$, we have $\mathbf{P}\{X \geq t\} \leq \frac{1}{t}\mathbf{E}[X]$.

Proof. Fix $t > 0$ and let $Y = X\mathbf{1}_{X<t}$ and $Z = X\mathbf{1}_{X\geq t}$ so that $X = Y + Z$. Both Y and Z are non-negative random variable and hence $\mathbf{E}[X] = \mathbf{E}[Y] + \mathbf{E}[Z] \geq \mathbf{E}[Z]$. On the other hand, $Z \geq t\mathbf{1}_{X\geq t}$ (Why?). Therefore $\mathbf{E}[Z] \geq t\mathbf{E}[\mathbf{1}_{X\geq t}] = t\mathbf{P}\{X \geq t\}$. Putting these together we get $\mathbf{E}[X] \geq t\mathbf{P}\{X \geq t\}$ as we desired to show. ■

Markov's inequality is simple, but surprisingly useful. Firstly, one can apply it to functions of our random variable and get many inequalities. Here are some.

Variants of Markov's inequality:

(1) If X is a non-negative random variable with finite p^{th} moment, then $\mathbf{P}\{X \geq t\} \leq t^{-p}\mathbf{E}[X^p]$ for any $t > 0$.

(2) If X is a random variable with finite second moment and $\mu = \mathbf{E}[X]$, then $\mathbf{P}[|X - \mu| \geq t] \leq \frac{1}{t^2}\text{Var}(X)$. [*Chebyshev's inequality*]

(3) If X is a random variable with finite exponential moments, then $\mathbf{P}\{X \geq t\} \leq e^{-\lambda t}\mathbf{E}[e^{\lambda X}]$ for any $\lambda > 0$. [*Chernoff's inequality*]

Thus, if we only know that X has finite mean, the tail probability $\mathbf{P}\{X \geq t\}$ must decay at least as fast as $1/t$. But, if we knew that the second moment was finite we could assert that the decay must be at least as fast as $1/t^2$, which is better. If $\mathbf{E}[e^{\lambda X}] < \infty$, then we get much faster decay of the tail, like $e^{-\lambda t}$.

Chebyshev's inequality captures again the intuitive notion that variance measures the spread of the distribution about the mean. The smaller the variance, lesser the spread. An alternate way to write Chebyshev's inequality is

$$\mathbf{P}(|X - \mu| > r\sigma) \leq \frac{1}{r^2},$$

where $\sigma = \text{s.d.}(X)$. This measures the deviations in multiples of the standard deviation. This is a very general inequality. In specific cases we can get better bounds than $1/r^2$ (just like Markov inequality can be improved using higher moments, when they exist).

Jensen's Inequality: Let $I \subseteq \mathbb{R}$ be an interval and let $\varphi : I \to \mathbb{R}$ be a twice differentiable function such that its second order derivative φ'' is continuous on I and $\varphi''(x) \geq 0 \, \forall \, x \in \mathbb{R}$ (i.e., φ is convex). Let X be a random variable with support $S_X \subseteq I$, and finite expectation. Then,

$$\mathbf{E}[\varphi(X)] \geq \varphi(\mathbf{E}[X]).$$

If $\varphi'(x) > 0 \, \forall \, x \in I$, then the inequality above is strict unless X is a degenerate random variable.

Exercise 17. A random variable X is said to be degenerate at a point $c \in \mathbb{R}$ if $\mathbf{P}(X = c) = 1$. Find $\text{Var}[X]$.

AM-GM-HM Inequality: Let X be a random variable with support $S_X \subseteq (0, \infty)$. Then, $\mathbf{E}[X]$ is called the arithmetic mean (AM) of X, $e^{\mathbf{E}[\ln X]}$ is called the geometric mean (GM) of X, and $\frac{1}{\mathbf{E}[1/X]}$ is called harmonic mean (HM) of X (provided these quantities exist). Then,

$$\mathbf{E}[X] \geq e^{\mathbf{E}[\ln X]} \geq \frac{1}{\mathbf{E}[1/X]}.$$

Exercise 18. Prove this inequality.

5. Simulation - I

As we have emphasized, probability is applicable to many situations in the real world. As such one may conduct experiments to verify the extent to which theorems are actually valid. For this we need to be able to draw numbers at random from any given distribution.

For example, take the case of Bernoulli($1/2$) distribution. One experiment that can give this is that of physically tossing a coin. This is not entirely satisfactory for several reasons. Firstly, are real coins fair? Secondly, what if we change slightly and want to generate from Ber(0.45)? In this section, we describe how to draw random numbers from various distributions on a computer. We do not fully answer this question. Instead what we shall show is
If one can generate random numbers from Unif($[0,1]$) distribution, then one can draw random numbers from any other distribution. More precisely, suppose U is a random variable with Unif($[0,1]$) distribution. We want to simulate random numbers from a given distribution F. Then, we shall find a function $\psi : [0,1] \to \mathbb{R}$ so that the random variable $X := \psi(U)$ has the given distribution F.

Important: The question of how to draw random numbers from Unif($[0,1]$) distribution is a very difficult one, and we shall just make a few superficial remarks about that.

Drawing random numbers from a discrete pmf: First start with an example.

Example 19. Suppose we want to draw random numbers from Ber(0.4) distribution. Let $\psi : [0,1] \to \mathbb{R}$ be defined as $\psi(t) = \mathbf{1}_{t \le 0.4}$. Let $X = \psi(U)$, i.e., $X = 1$ if $U \le 0.4$ and $X = 0$ otherwise. Then

$$\mathbf{P}\{X = 1\} = \mathbf{P}\{U \le 0.4\} = 0.4, \qquad \mathbf{P}\{X = 0\} = \mathbf{P}\{U > 0.4\} = 0.6.$$

Thus, X has Ber(0.4) distribution.

It is clear how to generalize this.

General rule: Suppose we are given a pmf f

$$\begin{pmatrix} t_1 & t_2 & t_3 & \dots \\ f(t_1) & f(t_2) & f(t_3) & \dots \end{pmatrix}.$$

Then, define $\psi : [0,1] \to \mathbb{R}$ as

$$\psi(u) = \begin{cases} t_1 & \text{if } u \in [0, f(t_1)] \\ t_2 & \text{if } u \in (f(t_1), f(t_1) + f(t_2)] \\ t_3 & \text{if } u \in (f(t_1) + f(t_2), f(t_1) + f(t_2) + f(t_3)] \\ \vdots & \vdots \end{cases}$$

Then, define $X = \psi(U)$. Clearly X takes the values $t_1, t_2, \ldots$ and

$$\mathbf{P}\{X = t_k\} = \mathbf{P}\left\{\sum_{j=1}^{k-1} f(t_j) < U \leq \sum_{j=1}^{k} f(t_j)\right\} = f(t_k).$$

Thus, X has pmf f.

Exercise 20. Write R codes to draw 100 random numbers from each of the following distributions (and draw the histograms). Compare with the pmf.

(1) Bin(n, p) for $n = 10, 20, 40$ and $p = 0.5, 0.3, 0.9$.

(2) Geo(p) for $p = 0.9, 0.5, 0.3$.

(3) Pois(λ) with $\lambda = 1, 4, 10$.

(4) Hypergeo(N_1, N_2, m) with $N_1 = 100, N_2 = 50, m = 20$, $N_1 = 1000, N_2 = 1000, m = 40$.

Check the file 'Distributions_discrete.R' under 'Resources' on HelloIITK.