10. Week 10

Proposition 10.1 (Cauchy-Schwarz Inequality)**.** *Let X and Y be RVs defined on the same probability space. Then,*

$$(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2),$$

provided the expectations exist. The equality occurs if and only if $\mathbb{P}(Y = cX) = 1$ or $\mathbb{P}(X = cY) = 1$ for some $c \in \mathbb{R}$.

Proof. First we consider the case when $\mathbb{E}X^2 = 0$. Then $\mathbb{P}(X = 0) = 1$ and consequently $\mathbb{P}(XY = 0) = 1$ and $\mathbb{E}(XY) = 0$. The equality holds.

Now, assume that $\mathbb{E}X^2 > 0$. Now, for all $c \in \mathbb{R}$, we have $\mathbb{E}(Y - cX)^2 = c^2\mathbb{E}X^2 - 2c\mathbb{E}(XY) + \mathbb{E}Y^2 \geq 0$. Hence, the discriminant $(2\mathbb{E}(XY))^2 - 4\mathbb{E}X^2\,\mathbb{E}Y^2$ must be non-positive, which proves the statement.

If the equality holds for some $\mathbb{E}(Y - cX)^2 = 0$ for some c, then we have $\mathbb{P}(Y = cX) = 1$. If $\mathbb{P}(Y = cX) = 1$ for some c, then $\mathbb{E}(Y - cX)^2 = 0$. Interchanging the roles of X and Y, we can discuss the case involving $\mathbb{E}(X - cY)^2$ and $\mathbb{P}(X = cY)$. □

Corollary 10.2. *Let X and Y be RVs defined on the same probability space. Then,*

$$(Cov(X, Y))^2 \leq Var(X)\,Var(Y),$$

provided the covariance and the variances exist.

Proof. Take $U = X - \mathbb{E}X$ and $V = Y - \mathbb{E}Y$. Applying the Cauchy-Schwarz inequality to U and V, the result follows. □

Remark 10.3. (a) Recall from the discussion in Definition 6.33 that

$$Var(X) = \mathbb{E}\,(X - \mathbb{E}X)^2 = \inf_{c \in \mathbb{R}} \mathbb{E}(X - c)^2,$$

for any RV X with finite second moment. This result can be interpretated in the following way: $\mathbb{E}X$ is the best constant to approximate an RV X with an 'error' $\sqrt{Var(X)}$.

(b) Now, given two non-degenerate RVs X and Y, we can ask the following question: what are the best constants a, b when we try to approximate Y by $aX + b$? To be precise, we are looking at the problem of minimizing $\mathbb{E}(Y - aX - b)^2$ over all $a, b \in \mathbb{R}$.

(c) Continue with the problem mentioned above. Fix $a \in \mathbb{R}$. Then, the best b which minimizes $\mathbb{E}(Y - aX - b)^2$ is $b = \mathbb{E}(Y - aX)$ with an 'error' $\sqrt{Var(Y - aX)}$. We may therefore consider

$$\inf_{a,b\in\mathbb{R}} \mathbb{E}(Y-aX-b)^2 = \inf_{a\in\mathbb{R}} \mathbb{E}\left[Y - aX - \mathbb{E}(Y - aX)\right]^2 = \inf_{a\in\mathbb{R}} \left[Var(Y) - 2aCov(X,Y) + a^2 Var(X)\right].$$

By a simple computation, we conclude that $a = \frac{Cov(X,Y)}{Var(X)}, b = \mathbb{E}(Y - aX)$ attains the infimum in

$$\inf_{a,b\in\mathbb{R}} \mathbb{E}(Y - aX - b)^2 = Var(Y)\left[1 - \left(\frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}\right)^2\right] \leq Var(Y) = \inf_{a\in\mathbb{R}} \mathbb{E}(Y - a)^2.$$

The inequality above follows from the Cauchy-Schwarz inequality (Proposition 10.1).

Definition 10.4 (Correlation between RVs)**.** Let X and Y be RVs defined on the same probability space. If $0 < Var(X) < \infty, 0 < Var(Y) < \infty$, then we call

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$$

as the Correlation between X and Y. We say X and Y are uncorrelated if $\rho(X,Y) = 0$ or equivalently $Cov(X,Y) = 0$.

Note 10.5. By Corollary 10.2, $|\rho(X,Y)| \leq 1$ for any two RVs X and Y defined on the same probability space.

Remark 10.6 (Correlation and Independence)*.* If X and Y are independent RVs defined on the same probability space, then by Remark 9.32(i), $Cov(X,Y) = 0$ and hence X and Y are uncorrelated. However, the converse is not true. We illustrate this problem with examples.

(a) Let $X = (X_1, X_2)$ be a bivariate discrete random vector, i.e. a 2-dimensional discrete random vector with joint p.m.f. given by

$$f_X(x_1, x_2) = \begin{cases} \frac{1}{2}, \text{ if } (x_1, x_2) = (0, 0), \\ \frac{1}{4}, \text{ if } (x_1, x_2) = (1, 1) \text{ or } (1, -1), \\ 0, \text{ otherwise.} \end{cases}$$

The marginal p.m.fs are

$$f_{X_1}(x_1) = \begin{cases} \frac{1}{2}, \text{ if } x_1 \in \{0, 1\} \\ 0, \text{ otherwise} \end{cases} \quad , \quad f_{X_2}(x_2) = \begin{cases} \frac{1}{2}, \text{ if } x_2 = 0 \\ \frac{1}{4}, \text{ if } x_2 \in \{1, -1\} \\ 0, \text{ otherwise} \end{cases}$$

We have $f_{X_1,X_2}(0, 0) = \frac{1}{2} \neq \frac{1}{4} = f_{X_1}(0) f_{X_2}(0)$ and hence X_1 and X_2 are not independent. But, $\mathbb{E}X_1 = \frac{1}{2}, \mathbb{E}X_2 = 0, \mathbb{E}(X_1X_2) = 0, Var(X_1) > 0$ and $Var(X_2) > 0$. Therefore $Cov(X_1, X_2) = 0$ and hence X_1 and X_2 are uncorrelated.

(b) Let $X = (X_1, X_2)$ be a bivariate continuous random vector, i.e. a 2-dimensional continuous random vector with joint p.d.f. given by

$$f_X(x_1, x_2) = \begin{cases} 1, & \text{if } 0 < |x_2| \leq x_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\mathbb{E}(X_1X_2) = \int_0^1 \int_{-x_1}^{x_1} x_1 x_2 \, \mathrm{d}x_2 \, \mathrm{d}x_1 = 0,$$

and

$$\mathbb{E}(X_1) = \int_0^1 \int_{-x_1}^{x_1} x_1 \, \mathrm{d}x_2 \, \mathrm{d}x_1 = \frac{2}{3}, \quad \mathbb{E}(X_2) = \int_0^1 \int_{-x_1}^{x_1} x_2 \, \mathrm{d}x_2 \, \mathrm{d}x_1 = 0.$$

Hence, $\mathbb{E}(X_1X_2) = (\mathbb{E}X_1)(\mathbb{E}X_2)$, which implies $Cov(X_1, X_2) = 0$. A similar computation shows $Var(X_1)$ and $Var(X_2)$ exists and are non-zero. Hence, X_1 and X_2 are uncorrelated.

Now, by computing the marginal p.d.f.s f_{X_1} and f_{X_2}, it is immediate that the equality

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

does not hold for all $x = (x_1, x_2) \in \mathbb{R}^2$. Here, X_1 and X_2 are not independent. The verification with the marginal p.d.f.s is left as an exercise in practice problem set 9.

We now discuss the concept of equality of distribution for random vectors. As we shall see, the ideas remain the same as in the case of RVs.

Definition 10.7 (Identically distributed random vectors)**.** Let X and Y be two p-dimensional random vectors, possibly defined on different probability spaces. We say that they have the same law/distribution, or equivalently, X and Y are identically distributed or equivalently, X and Y are equal in law/distribution, denoted by $X \stackrel{d}{=} Y$, if $F_X(x) = F_Y(x), \forall x \in \mathbb{R}^p$.

Remark 10.8. As discussed in the case of RVs in Remark 7.1, we can check whether two random vectors are identically distributed or not via other quantities that describe their law/distribution.

(a) Let X and Y be p-dimensional discrete random vectors with joint p.m.f.s f_X and f_Y, respectively. Then X and Y are identically distributed if and only if $f_X(x) = f_Y(x), \forall x \in \mathbb{R}^p$.

(b) Let X and Y be p-dimensional continuous random vectors with joint p.d.f.s f_X and f_Y, respectively. Then X and Y are identically distributed if and only if $f_X(x) = f_Y(x), \forall x \in \mathbb{R}^p$.

(c) Let X and Y be p-dimensional random vectors such that their joint MGFs M_X and M_Y exist and agree on $(-a_1, a_1) \times (-a_2, a_2) \times \cdots (-a_p, a_p)$ for some $a_1, a_2, \cdots, a_p > 0$, then X and Y are identically distributed.

(d) Let X and Y be identically distributed p-dimensional random vectors. Then for any function $h : \mathbb{R}^p \to \mathbb{R}^q$, we have $h(X)$ and $h(Y)$ are identically distributed q-dimensional random vectors.

Notation 10.9 (i.i.d RVs)**.** We say that RVs $X_1, \cdots, X_p$ defined on the same probability space are independent and identically distributed, then we usually use the short hand notation i.i.d. and say that $X_1, \cdots, X_p$ are i.i.d..

Note 10.10. The concept of independence of random vectors can be discussed in the same way as done for independence of random variables.

Definition 10.11. (a) A random sample is a collection of i.i.d. RVs.

(b) A random sample of size n is a collection of n i.i.d. RVs $X_1, X_2, \cdots, X_n$.

(c) Let $X_1, X_2, \cdots, X_n$ be a random sample of size n. If the common DF is F or the common p.m.f./p.d.f. is f, then we call $X_1, X_2, \cdots, X_n$ to be a random sample from a distribution having a DF F or p.m.f./p.d.f. f.

(d) A function of one or more RVs that does not depend on any unknown parameter is called a statistic.

Example 10.12. Suppose that $X_1, \cdots, X_n$ are i.i.d. with the common distribution being $Poisson(\theta)$ or $Exponential(\theta)$ for some unknown $\theta \in (0, \infty)$. Here, θ is a unknown parameter.

(a) $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ is a statistic and is usually referred to as the sample mean.

(b) $X_1 - \theta$ is not a statistic.

(c) $S_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a statistic and is usually referred to as the sample variance. Depending on the situation, we sometimes work with $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

(d) The value of S_n such that S_n^2 is the sample variance, is referred to as the sample standard deviation.

(e) For $r = 1, \cdots, n$, we denote by $X_{(r:n)}$ the r-th smallest of $X_1, \cdots, X_n$. By definition, $X_{(1:n)} \leq \cdots \leq X_{(n:n)}$ and these are called the order statistics of the random sample. If n is understood, then we simply write $X_{(r)}$ to denote the r-th order statistic.

Note 10.13. Let X_1, X_2 be a random sample of size 2. Then $X_{(1)} = \min\{X_1, X_2\} = \frac{1}{2}(X_1 + X_2) - \frac{1}{2}|X_1 - X_2|$ and $X_{(2)} = \max\{X_1, X_2\} = \frac{1}{2}(X_1 + X_2) + \frac{1}{2}|X_1 - X_2|$ are RVs. Using similar arguments, it follows that the order statistics from any random sample of size n are RVs. The joint distribution of the order statistics is therefore of interest.

Note 10.14. Let $X_1, \cdots, X_n$ be a random sample of continuous RVs with the common p.d.f. f. Then,

$$\mathbb{P}(X_{(1)} < X_{(2)}) < \cdots < X_{(n)}) = 1$$

and hence $X_{(r)}, r = 1, \cdots, n$ are defined uniquely with probability one.

Proposition 10.15. *Let $X_1, \cdots, X_n$ be a random sample of continuous RVs with the common DF F and the common p.d.f. f. The joint p.d.f. of $(X_{(1)}, \cdots, X_{(n)})$ is given by*

$$g(y_1, \cdots, y_n) = \begin{cases} n! \prod_{i=1}^{n} f(y_i), & \text{if } y_1 < \cdots < y_n, \\ 0, & \text{otherwise.} \end{cases}$$

Further the marginal p.d.f. of $X_{(r)}$ is given by

$$g_{X_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!}(F(y))^{r-1}(1 - F(y))^{n-r} f(y), \forall y \in \mathbb{R}.$$

Proof. Observe that a sample value $(y_1, \cdots, y_n)$ of $(X_{(1)}, \cdots, X_{(n)})$ is related to a sample $(x_1, \cdots, x_n)$ of $(X_1, \cdots, X_n)$ in the following way

$$(y_1, \cdots, y_n) = (x_{(1)}, \cdots, x_{(n)}),$$

$x_{(r)}$ being the r-th smallest of $x_1, \cdots, x_n$. Note that $y_r = x_{(r)}$.
Now, the actual values $x_1, \cdots, x_n$ may have been arranged in a different order than $x_{(1)}, \cdots, x_{(n)}$. In fact, the values $x_{(1)}, \cdots, x_{(n)}$ arise from one of the $n!$ permutations of the values $x_1, \cdots, x_n$. But, any such transformation/permutation is obtained by the action of a permutation matrix on the vector $(x_1, \cdots, x_n)$. For example, if $x_1 < x_2 < \cdots < x_{n-2} < x_n < x_{n-1}$, then $x_{(1)} = x_1, \cdots, x_{(n-2)} = x_{n-2}, x_{(n-1)} = x_n, x_{(n)} = x_{n-1}$ which interchanges the $n-1$ and n-th values, i.e. x_{n-1} and x_n.
Hence, the Jacobian matrix for this transformation is the same as the corresponding permutation matrix and the Jacobian determinant is ± 1.
Since $X_1, \cdots, X_n$ are i.i.d., the joint p.d.f. of $(X_1, \cdots, X_n)$ is given by

$$f_{X_1, \cdots, X_n}(x_1, \cdots, x_n) = f(x_1) \times \cdots \times f(x_n), \forall (x_1, \cdots, x_n) \in \mathbb{R}^n.$$

Using Theorem 9.24, we have joint p.d.f. of $(X_{(1)}, \cdots, X_{(n)})$ is given by

$$g(y_1, \cdots, y_n) = \begin{cases} n! \prod_{i=1}^{n} f(y_i), \text{ if } y_1 < \cdots < y_n, \\ 0, \text{ otherwise.} \end{cases}$$

The marginal p.d.f. of $X_{(r)}$ can now be computed for $y \in \mathbb{R}$,

$$g_{X_{(r)}}(y)$$
$$= \int_{y_{r-1}=-\infty}^{y} \int_{y_{r-2}=-\infty}^{y_{r-1}} \cdots \int_{y_1=-\infty}^{y_2} \int_{y_{r+1}=y}^{\infty} \int_{y_{r+2}=y_{r+1}}^{\infty} \cdots \int_{y_n=y_{n-1}}^{\infty} n! \prod_{i=1}^{n} f(y_i)\, dy_n dy_{n-1} \cdots dy_{r+1} dy_1 dy_2 \cdots dy_{r-1}$$

The above integral simplifies to the result stated above. □

Example 10.16. Let X_1, X_2, X_3 be a random sample from $Uniform(0, 1)$ distribution. The common p.d.f. here is given by

$$f(x) = \begin{cases} 1, \text{ if } x \in (0, 1) \\ 0, \text{ otherwise.} \end{cases}$$

By the above result, the joint p.d.f. of $(X_{(1)}, X_{(2)}, X_{(3)})$ is given by

$$g(y_1, y_2, y_3) = \begin{cases} 6, \text{ if } 0 < y_1 < y_2 < y_3 < 1, \\ 0, \text{ otherwise.} \end{cases}$$

and the marginal p.d.f. of $X_{(1)}$ is

$$g(y_1) = \begin{cases} 3(1 - y_1)^2, \text{ if } y_1 \in (0, 1) \\ 0, \text{ otherwise.} \end{cases}$$

Remark 10.17. For random samples from discrete distributions, there is no general formula or result which helps in computing the joint distribution of the order statistics. Usually they are done by a case-by-case analysis. Let X_1, X_2, X_3 be a random sample from $Bernoulli(p)$ distribution, for

some $p \in (0,1)$. The common p.m.f. here is given by

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

Note that $X_{(1)}$ is also a $\{0,1\}$-valued RV with $X_{(1)} = \min\{X_1, X_2, X_3\} = 1$ if and only if $X_1 = X_2 = X_3 = 1$. Then using independence,

$$\mathbb{P}(X_{(1)} = 1) = \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 1) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1)\mathbb{P}(X_3 = 1) = p^3$$

and $\mathbb{P}(X_{(1)} = 0) = 1 - \mathbb{P}(X_{(1)} = 1) = 1 - p^3$. Therefore, $X_{(1)} \sim Bernoulli(p^3)$. Similarly, $X_{(3)} \sim Bernoulli(1-(1-p)^3)$. The distribution of $X_{(2)}$ is left as an exercise in problem set 10.

Earlier, we have discussed the concept of conditional distributions and the concept of expectation of a random vector. Combining these two concepts, we are led to the following.

Definition 10.18 (Conditional Expectation, Conditional Variance and Conditional Covariance)**.** Let $X = (X_1, X_2, \cdots, X_{p+q})$ be a $p+q$-dimensional random vector with joint p.m.f./p.d.f. f_X. Let the joint p.m.f./p.d.f. $Y = (X_1, X_2, \cdots, X_p)$ and $Z = (X_{p+1}, X_{p+2}, \cdots, X_{p+q})$ be denoted by f_Y and f_Z, respectively. Let $h : \mathbb{R}^p \to \mathbb{R}$ be a function. Let $z \in \mathbb{R}^q$ be such that $f_Z(z) > 0$.

(a) The conditional expectation of $h(Y)$ given $Z = z$, denoted by $\mathbb{E}(h(Y) \mid Z = z)$, is the expectation of $h(Y)$ under the conditional distribution of Y given $Z = z$.

(b) The conditional variance of $h(Y)$ given $Z = z$, denoted by $Var(h(Y) \mid Z = z)$, is the variance of $h(Y)$ under the conditional distribution of Y given $Z = z$.

(c) Let $1 \leq i \neq j \leq p$. The conditional covariance between X_i and X_j given $Z = z$, denoted by $Cov(X_i, X_j \mid Z = z)$, is the covariance between X_i and X_j under the conditional distribution of (X_i, X_j) given $Z = z$.

Notation 10.19. On $\{z \in \mathbb{R}^q : f_Z(z) > 0\}$, consider the function, $g_1(z) := \mathbb{E}(h(Y) \mid Z = z)$. We denote the RV $g_1(Z)$ by $\mathbb{E}(h(Y) \mid Z)$. Similarly, define the RVs $Var(h(Y) \mid Z)$ and $Cov(X_1, X_2 \mid Z)$

Proposition 10.20. *The following are properties of Conditional Expectation, Conditional Variance and Conditional Covariance. Here, we assume that the relevant expectations exist.*

(a) $\mathbb{E}h(Y) = \mathbb{E}(\mathbb{E}(h(Y) \mid Z)$.

(b) $Var(h(Y)) = Var(\mathbb{E}(h(Y) \mid Z)) + \mathbb{E}Var(h(Y) \mid Z)$.

(c) $Cov(X_1, X_2) = Cov(\mathbb{E}(X_1 \mid Z), \mathbb{E}(X_2 \mid Z)) + \mathbb{E}Cov(X_1, X_2 \mid Z)$.

Proof. We only prove the first statement under a simple assumption. The general case and other statements can be proved using appropriate generalization.

Take $p = q = 1$ and let $X = (Y, Z)$ be a 2-dimensional continuous random vector. Then,

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(h(Y) \mid Z) &= \int_{-\infty}^{\infty} \mathbb{E}(h(Y) \mid Z = z) f_Z(z)\, dz \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} h(y) f_{Y|Z}(y \mid z)\, dy \right] f_Z(z)\, dz \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y) f_{Y,Z}(y, z)\, dydz \\
&= \mathbb{E}h(Y).
\end{aligned}$$

□

Example 10.21. We shall see computations for conditional expectations in a later lecture.

We look at examples of discrete RVs in relation with random experiments.

Remark 10.22 (Binomial RVs via random experiments)*.* Recall that in Remark 7.20, we have seen Bernoulli RVs arising from Bernoulli trials. Now, consider the same random experiment with two outcomes 'Success' and 'Failure' with probability of success $p \in (0, 1)$. Now, consider n independent Bernoulli trials of this experiment with the RV X_i being 1 for 'Success' and 0 for 'Failure' in the i-th trial for $i = 1, 2, \cdots, n$. Then, $X_1, X_2, \cdots, X_n$ is a random sample of size n from the $Bernoulli(p)$ distribution. Now, the total number X of successes in the n trials is given by $X = X_1 + X_2 + \cdots + X_n$ and hence, by Remark 9.23, $X \sim Binomial(n, p)$. A $Binomial(n, p)$ RV can therefore be interpretated as the number of successes in n trials of a random experiment with two outcomes 'Success' and 'Failure' with probability of success $p \in (0, 1)$. Here, we have kept p fixed over all the trials.

Example 10.23. Suppose that a standard six-sided fair die is rolled at random 4 times independently. We now consider the probability that all the rolls result in a number at least 5. In each roll, obtaining at least 5 has the probability $\frac{2}{6} = \frac{1}{3}$ - we treat this as the probability of success in one trial. Repeating the trial three times independently gives us the number of success as $X \sim Binomial(4, \frac{1}{3})$. The probability that all the rolls result in successes is given by $\mathbb{P}(X = 4)$ – which can now be computed from the Binomial distribution. If we now consider the probability that at least two rolls result in a number at least 5, then that probability is given by $\mathbb{P}(X \geq 2)$.

Example 10.24 (Negative Binomial RV)**.** Consider a random experiment with two outcomes 'Success' and 'Failure' with probability of success $p \in (0, 1)$. We consider repeating the experiment until we have r successes, with r being a positive integer. Let X denote the number of failures observed till the r-th success. Then X is a discrete RV with the support of X being $S_X = \{0, 1, \cdots\}$. Note that for $k \in S_X$, using independence of the trials we have

$$
\begin{aligned}
&\mathbb{P}(X = k) \\
&= \mathbb{P}(\text{there are } k \text{ failures before the } r\text{-th success}) \\
&= \mathbb{P}(\text{first } k + r - 1 \text{ trials result in } r - 1 \text{ successes and the } k + r\text{-th trial results in a success}) \\
&= \mathbb{P}(\text{first } k + r - 1 \text{ trials result in } r - 1 \text{ successes}) \times \mathbb{P}(\text{the } k + r\text{-th trial results in a success}) \\
&= \binom{k + r - 1}{r - 1} p^{r-1}(1 - p)^k \times p \\
&= \binom{k + r - 1}{k} p^r (1 - p)^k.
\end{aligned}
$$

Therefore the p.m.f. of X is given by

$$
f_X(x) = \begin{cases} \binom{x+r-1}{x} p^r (1 - p)^x, \text{ if } x \in S_X, \\ 0, \text{ otherwise.} \end{cases}
$$

In this case, we say X follows the negative Binomial(r, p) distribution or equivalently, X is a negative Binomial (r, p) RV. Here, r denotes the number of successes at which the trials are

terminated and p being the probability of success. The MGF can now be computed as follows.

$$\begin{aligned} M_X(t) &= \mathbb{E}e^{tX} \\ &= \sum_{k=0}^{\infty} e^{tk} \binom{k+r-1}{k} p^r (1-p)^k \\ &= \sum_{k=0}^{\infty} \binom{k+r-1}{k} p^r \left[(1-p)e^t\right]^k \\ &= p^r [1-(1-p)e^t]^{-r}, \forall t < -\ln(1-p). \end{aligned}$$

Using the MGF, we can compute the mean and variance of X as $\mathbb{E}X = \frac{rq}{p}, Var(X) = \frac{rq}{p^2}$, with $q = 1-p$.

Remark 10.25 (Connection between negative Binomial distribution and the Geometric distribution). A negative Binomial$(1, p)$ RV X has the p.m.f.

$$f_X(x) = \begin{cases} p(1-p)^x, \text{ if } x \in \{0, 1, \cdots\}, \\ 0, \text{ otherwise.} \end{cases}$$

which is exactly the same as the p.m.f. for the Geometric(p) distribution. Since the p.m.f. of a discrete RV determines the distribution, we conclude that a *Geometric*(p) RV can be identified as the number of failures observed till the first success in independent trials of a random experiment with two outcomes 'Success' and 'Failure' with probability of success $p \in (0, 1)$.

Note 10.26 (No memory property for Geometric Distribution)**.** Let $X \sim$ *Geometric*(p) for some $p \in (0, 1)$. For any non-negative integer n, we have

$$\mathbb{P}(X \geq n) = \sum_{k=n}^{\infty} p(1-p)^k = p(1-p)^n \sum_{k=0}^{\infty} (1-p)^k = (1-p)^n.$$

Then, for any non-negative integers m, n, we have

$$\mathbb{P}(X \geq m+n \mid X \geq m) = \frac{\mathbb{P}(X \geq m+n \text{ and } X \geq m)}{\mathbb{P}(X \geq m)} = \frac{\mathbb{P}(X \geq m+n)}{\mathbb{P}(X \geq m)} = (1-p)^n = \mathbb{P}(X \geq n).$$

Here, the probability of obtaining at least n additional failures (till the first success) beyond the first m or more failures remain the same as in the the probability of obtaining at least n failures till the first success. In the situation where we stress test a device under repeated shocks, if we consider the survival or continued operation of the device under shocks as 'Failures' in our trial and if the number of shocks till the device breaks down follows $Geometric(p)$ distribution, then we can interpret that the age of the device (measured in number of shocks observed) has no effect on the remaining lifetime of the device. This property is usually referred to as the 'No memory' property of the Geometric distribution.

Note 10.27. See problem set 9 for a similar property for the Exponential distribution.

Example 10.28. Let us consider the random experiment of rolling a standard six-sided fair die till we observe an outcome of at least 5. As mentioned in Example 10.23, the probability of success is $\frac{1}{3}$. Since the last roll results in a success, the number Y of rolls required is exactly one more than the number X of failures observed. Here $X \sim Geometric(\frac{1}{3})$. Then, the probability that an outcome of 5 or 6 is observed in the 10-th roll for the first time is given by

$$\mathbb{P}(Y = 10) = \mathbb{P}(X = 9) = \frac{1}{3}\left(\frac{2}{3}\right)^9.$$

If we want to look at Z which is the number of failures observed till 5 or 6 is rolled twice, then Z follows negative Binomial$(2, \frac{1}{3})$. Now, the number of rolls required is $Z + 2$. The probability that 10 rolls are required is given by

$$\mathbb{P}(Z + 2 = 10) = \mathbb{P}(Z = 8) = \binom{9}{8}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^8.$$

Note 10.29. Suppose that a box contains N items, out of which M items have been marked/labelled. In our experiment, we consider all labelled items to be identical and the same for all the unlabelled items. If we draw items from the box with replacement, then the probability of drawing a marked/labelled item is $\frac{M}{N}$ does not change between the draws. If we draw n items at random with replacement, then the number X of marked/labelled items follow $Binomial(n, \frac{M}{N})$ distribution. The case where the draws are conducted without replacement is of interest.

Example 10.30 (Hypergeometric RV)**.** In the setup of Note 10.29, consider drawing n items at random without replacement. Here, the probability of drawing a marked/labelled item may change between the draws and the number X of marked/labelled items in the n drawn items need not follow $Binomial(n, \frac{M}{N})$ distribution. Here, the number of labelled items among the items drawn satisfies the relation

$$0 \leq X \leq \min\{n, M\} \leq N$$

and the number of unlabelled items among the items drawn satisfies the relation

$$0 \leq n - X \leq N - M$$

and hence X is a discrete RV with support $S_X = \{\max\{0, n-(N-M)\}, \max\{0, n-(N-M)\}+1, \cdots, \min\{n, M\}\}$. The p.m.f. of X is given by

$$f_X(x) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, & \text{if } x \in S_X, \\ 0, & \text{otherwise.} \end{cases}$$

In this case, we say X follows the Hypergeometric distribution or equivalently, X is a Hypergeometric RV. This distribution has the three parameters N, M and n. Using properties of binomial coefficients, we can compute the factorial moments of X (left as exercise in problem set 10) and using these values we have,

$$\mathbb{E}X = \frac{nM}{N}, \quad Var(X) = \frac{nM}{N^2(N-1)}(N-M)(N-n) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}.$$

Note 10.31. In the setup of a Hypergeometric RV, if we consider $p = \frac{M}{N}$ as the probability of success and n as the number of trials, then $\mathbb{E}X$ matches with that of a $Binomial(n, \frac{M}{N})$ RV and $Var(X)$ is close to that of a $Binomial(n, \frac{M}{N})$ RV for small sample sizes n.

Example 10.32. Suppose that there are multiple boxes each containing 100 electric bulbs and we draw 5 bulbs from each box for testing. If a box contains 10 defective bulbs, then the number X of defective bulbs in the drawn bulbs follows Hypergeometric distribution with parameters

$N = 100, M = 10, n = 5$. Here,

$$\mathbb{P}(X = 2) = \frac{\binom{10}{2}\binom{100-10}{5-2}}{\binom{100}{5}}.$$

Note 10.33. We continue with the setting of Note 10.29, where a box contains N items, out of which M items have been marked/labelled or are defective. In our experiment, we consider all labelled items to be identical and the same for all the unlabelled items. If we draw items from the box with replacement until the r-th defective item is drawn, then the number of draws required can be described in terms of negative Binomial$(r, \frac{M}{N})$ distribution, where the last draw yields the r-th defective item (see Example 10.28). The case where the draws are conducted without replacement is of interest.

Example 10.34 (Negative Hypergeometric RV)**.** In the setting of Note 10.33, consider drawing the items without replacement till the r-th defective item is obtained. We then have $1 \leq r \leq M$. Let X be the number of draws required. Then X is a discrete RV with support $S_X = \{r, r+1, \cdots, N\}$. For $k \in S_X$, using independence of the draws we have

$$\begin{aligned}
&\mathbb{P}(X = k) \\
&= \mathbb{P}(\text{first } k-1 \text{ trials result in } r-1 \text{ defective items and the } k\text{-th trial results in a defective item}) \\
&= \mathbb{P}(\text{first } k-1 \text{ trials result in } r-1 \text{ defective items}) \times \mathbb{P}(\text{the } k\text{-th trial results in a defective item}) \\
&= \frac{\binom{M}{r-1}\binom{N-M}{k-r}}{\binom{N}{k-1}} \times \frac{M-(r-1)}{N-(k-1)}.
\end{aligned}$$

Therefore the p.m.f. of X is given by

$$f_X(x) = \begin{cases} \frac{M-(r-1)}{N-(x-1)} \frac{\binom{M}{r-1}\binom{N-M}{x-r}}{\binom{N}{x-1}}, & \text{if } x \in \{r, r+1, \cdots, N\}, \\ 0, & \text{otherwise.} \end{cases}$$

In this case, we say X follows the negative Hypergeometric distribution or equivalently, X is a negative Hypergeometric RV.