Chapter 5. Discrete random variables.

5.1 Random variables.

In some of the examples we have considered the sample space can be represented by a set of integers, *e.g.*, $\Omega = \{1, 2, 3, 4, 5, 6\}$ corresponding to one toss of a die or $\Omega = \{1, 2, \ldots\}$ corresponding to the number of tosses required to obtain a one when tossing a die repeatedly. In other examples we restricted our attention to events described in terms of a numerical value, *e.g.*, the number of heads, $0, 1, \ldots, n$, in n tosses of a coin. We will now consider a more formal treatment of such assignments of numerical values to the outcomes of an experiment.

A function which assigns numerical values (real numbers) to the elements of a sample space is known as a random variable (denoted r.v.). The word variable indicates that the values of the function are numbers. The adjective random is used here as it is used in random experiment to indicate that the value of the random variable or outcome of the experiment is not know with certainty before the experiment is conducted and the value of the random variable or outcome of the experiment is determined.

Example. Consider the experiment of tossing a coin three times with sample space $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
If we let X denote the random variable "the number of heads", then:
1) the elementary outcome HHH is mapped to 3, *i.e.*, $X(HHH) = 3$;
2) each of HHT, HTH, and THH is mapped to 2,
i.e., $X(HHT) = X(HTH) = X(THH) = 2$;
3) each of HTT, THT, and TTH is mapped to 1,
i.e., $X(HTT) = X(THT) = X(TTH) = 1$;
4) the outcome TTT is mapped to 0, *i.e.*, $X(TTT) = 0$.
The sample space (collection of possible values) for this r.v. is $\Omega_X = \{0, 1, 2, 3\}$.

Given a sample space Ω and a random variable X defined on Ω, the r.v. X defines a new sample space Ω_X comprised of all the possible values of X. If Ω_X is discrete (finite or countably infinite), then X is said to be a discrete random variable. In this chapter we will restrict our attention to discrete random variables. If $x \in \Omega_X$, then the event $X = x$ (subset of Ω_X) corresponds to the event $\{\omega \in \Omega : X(\omega) = x\}$ (subset of Ω), *e.g.*, in the coin tossing example the event $X = 1$ ($\{1\}$ when viewed as a subset of Ω_X) corresponds to the event $\{HTT, THT, TTH\}$ relative to the sample space of the underlying experiment. To avoid technicalities involving uncountably infinite Ω, for each $x \in \mathcal{R}$, we will assume that the probability measure Pr on Ω assigns a probability to the event $X \leq x$. For $x \in \mathcal{R}$, we define the probability of the event $X \leq x$ in terms of the original sample space and

probability measure by setting $\Pr(X \leq x) = \Pr(\{\omega \in \Omega : X(\omega) \leq x\})$. The distribution of any r.v. X can be characterized by assigning probabilities $\Pr(X \leq x)$ for each $x \in \mathcal{R}$. If the r.v. X is discrete, then this characterization can be in terms of the probabilities $\Pr(X = x)$ for each $x \in \mathcal{R}$.

5.2 Distributions of discrete random variables.

As noted above, we can characterize the distribution of the discrete r.v. X by specifying the probabilities $\Pr(X = x)$ for each $x \in \mathcal{R}$. Thus we define the probability mass function (denoted p.m.f.) f_X, where

$$f_X(x) = \Pr(X = x),$$

with $f_X(x) \geq 0$ for all $x \in \mathcal{R}$, and $\sum_{\{x \in \Omega_X\}} f_X(x) = 1$. The sample space for X is $\Omega_X = \{x \in \mathcal{R} : f_X(x) > 0\}$. Note that for any event A defined in terms of the r.v. X, *i.e.*, any $A \subset \Omega_X$, the probability of A is the sum of the probabilities of each of its elements, *i.e.*,

$$\Pr(A) = \sum_{x \in A} f_X(x).$$

Note further that we can discuss such probabilities without reference to the original experiment. In fact, it is legitimate to refer to any function f_X with the requisite properties ($f_X(x) \geq 0$ for all $x \in \mathcal{R}$ and $\sum_{\{x \in \Omega_X\}} f_X(x) = 1$) as a p.m.f. on Ω_X without reference to any specific experiment.

Aside: Indicator functions. *Given a set A the indicator function $\mathbf{1}_A$ is defined by*

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}.$$

Our main use of indicator functions will be to provide convenient representations of functions with definitions involving cases. These representations simplify manipulations involving such functions.

If the possible values of the discrete r.v. X are nonnegative integers, then a useful alternate characterization of the distribution of X is provided by its probability generating function (denoted p.g.f.) P_X. Given a nonnegative integer valued r.v. X with p.m.f. f_X, let $p_x = f_X(x)$. The probability generating function of X is

$$P_X(t) = \sum_{x=0}^{\infty} t^x p_x = p_0 + tp_1 + t^2 p_2 + \cdots$$

The dummy variable t is of no significance. This series converges at least for $-1 \leq t \leq 1$.

The usefulness of the probability generating function as a characterization of a distribution will become clear shortly. For now we only note the relationship between derivatives

of $P_X(t)$ and the probabilities p_x. First note that $P_X(0) = p_0$. The derivative of the series $P_X(t)$ is

$$P'_X(t) = \sum_{x=1}^{\infty} x t^{x-1} p_x = p_1 + 2tp_2 + 3t^2 p_3 + \cdots$$

and $P'_X(0) = p_1$. The second derivative of $P_X(t)$ is

$$P''_X(t) = \sum_{x=2}^{\infty} x(x-1) t^{x-2} p_x = 2p_2 + 6tp_3 + 12t^3 p_4 + \cdots$$

and $\frac{P''_X(0)}{2} = p_2$. Continuing with this differentiation process we find that $\frac{P_X^{(k)}(0)}{k!} = p_k$, *i.e.*, for $k = 0, 1, \ldots$, the k^{th} derivative of $P_X(t)$ evaluated at $t = 0$ and divided by $k!$ is equal to $p_k = \Pr(X = k)$.

A parametric family of distributions is a family of distributions of a specified form indexed by a (possibly vector valued) parameter θ, *e.g.*, for a discrete r.v. X the p.m.f. f_X might be written as a function of θ which is completely determined by assigning a value to θ. The collection of parameter values for which the distribution is valid is the parameter space Θ for the family. We will now describe several standard parametric families of discrete distributions.

Several of these parametric families are associated with sequences of Bernoulli trials. A sequence of independent dichotomous trials is said to be a sequence of Bernoulli trials if the probabilities of the two possible outcomes are constant from trial to trial. The two possible outcomes are generically known as success (S) and failure (F) and p denotes the probability of success on a single trial, *i.e.*, $\Pr(S) = p$ and $\Pr(F) = 1 - p = q$. We can think of a Bernoulli trial with success probability p as the outcome of selecting a ball at random from a box containing balls labeled S and F and such that the proportion of balls labeled S is p, with $0 < p < 1$. In this context a sequence of Bernoulli trials with success probability p corresponds to the outcome of a sequence of selections of a ball from the box with the ball being returned to the box after each draw. The binomial, geometric, negative binomial, and Poisson distributions described below can be motivated in terms of sequences of Bernoulli trials.

Binomial distribution.

For a positive integer n, $0 < p < 1$, and $q = 1 - p$, the binomial distribution with parameters n and p has p.m.f.

$$f_X(x) = \binom{n}{x} p^x q^{n-x} \mathbf{1}_{\{0,1,\ldots,n\}}(x),$$

sample space $\Omega_X = \{0, 1, \ldots, n\}$, and p.g.f.

$$P_X(t) = (q + tp)^n.$$

The fact that $\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = 1$ follows from the binomial theorem as demonstrated in the binomial distribution example of Section 3.3. The preceding expression for the p.g.f. is derived similarly as follows

$$P_X(t) = \sum_{x=0}^{n} t^x \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (tp)^x q^{n-x} = (q + tp)^n.$$

If the distribution of the r.v. X is binomial with parameters n and p, then we will say that X is a binomial r.v. and indicate this by writing $X \sim$ binomial (n, p). The binomial $(1, p)$ distribution is also known as the Bernoulli (p) distribution. If we let X denote the number of successes in a sequence of n Bernoulli trials with success probability p, then X is a binomial (n, p) r.v.

Geometric distribution.

For $0 < p < 1$ and $q = 1 - p$, the geometric distribution with parameter p has p.m.f.

$$f_X(x) = pq^x \mathbf{1}_{\{0,1,\ldots\}}(x),$$

sample space $\Omega_X = \{0, 1, \ldots\}$, and p.g.f.

$$P_X(t) = \frac{p}{1 - tq}.$$

The fact that $\sum_{x=0}^{\infty} pq^x = 1$ follows from the fact that the geometric series $\sum_{x=0}^{\infty} q^x = \frac{1}{1-q} = \frac{1}{p}$. The preceding expression for the p.g.f. is derived similarly as follows

$$P_X(t) = \sum_{x=0}^{\infty} t^x pq^x = p \sum_{x=0}^{\infty} (tq)^x = p \left(\frac{1}{1 - tq} \right) = \frac{p}{1 - tq}.$$

If the distribution of the r.v. X is geometric with parameter p, then we will say that X is a geometric r.v. and indicate this by writing $X \sim$ geometric (p).

If we let X denote the number of trials up to but not including the first trial when a success occurs in a potentially infinite sequence of Bernoulli trials with success probability p, then X is a geometric (p) r.v. Note that this means that the first success occurs on trial $X + 1$.

The geometric distribution has an important and interesting lack of memory property. Let a denote a positive integer, b a nonnegative integer, and X a geometric (p) r.v., then

$$\Pr(X = a + b \mid X \geq a) = \Pr(X = b).$$

Before we prove this result note that if we think of X as the waiting time up to but not including the first trial when a success occurs in a sequence of Bernoulli trial with success probability p (the number of failures before the first success), then this result states that the conditional probability that the first success occurs on the $(b+1)^{th}$ trial after the a^{th} trial given that the first a trials resulted in failures is the same as the unconditional probability that the first success occurs on the $(b+1)^{th}$ trial. That is, if we have observed a failures, then conditionally given this information the probability of observing exactly b failures before the first success is exactly the same as it would be if we were starting from scratch. Here is the proof:

$$\Pr(X = a + b | X \geq a) = \frac{pq^{a+b}}{\sum_{x=a}^{\infty} pq^x} = \frac{pq^{a+b}}{q^a \sum_{x=0}^{\infty} pq^x} = pq^b = \Pr(X = b).$$

Aside. *The geometric distribution is the only distribution on the nonnegative integers with this lack of memory property. More formally, if the nonnegative integer valued r.v. X has the property that, for all positive integers k, $Pr(X = k|X \geq k) = p = Pr(X = 0)$, then X is a geometric (p) r.v.*

Proof. Let the nonnegative integer valued r.v. X be given and, for $k = 0, 1, \ldots$, let $p_k = \Pr(X = k)$ and $Q_k = \Pr(X > k) = p_{k+1} + p_{k+2} + \ldots$ Note that $Q_0 = 1 - p_0$. Assume that $\Pr(X = k|X \geq k) = p_0$ for all positive integers k. Then

$$p_0 = \Pr(X = k|X \geq k) = \frac{p_k}{Q_{k-1}} = \frac{Q_{k-1} - Q_k}{Q_{k-1}} = 1 - \frac{Q_k}{Q_{k-1}}.$$

Thus $\frac{Q_k}{Q_{k-1}} = 1 - p_0$ for all positive integers k, *i.e.*, $Q_k = (1 - p_0)Q_{k-1}$, so that $Q_1 = (1-p_0)^2$, $Q_2 = (1-p_0)^3$, $\ldots$ Hence, $\Pr(X > k) = Q_k = (1-p_0)^{k+1}$, $\Pr(X = 0) = p_0$, and for each positive integer k, $\Pr(X = k) = Q_{k-1} - Q_k = (1-p_0)^k(1-(1-p_0)) = p_0(1-p_0)^k$. In other words, X is a geometric (p_0) r.v. □

Negative binomial distribution.

For a positive integer r, $0 < p < 1$, and $q = 1 - p$, the negative binomial distribution with parameters r and p has p.m.f.

$$f_X(x) = \binom{r + x - 1}{x} p^r q^x \mathbf{1}_{\{0,1,\ldots\}}(x),$$

sample space $\Omega_X = \{0, 1, \ldots\}$, and p.g.f.

$$P_X(t) = \left(\frac{p}{1-tq}\right)^r.$$

Verification that the expression above is a valid p.m.f. is provided following Theorem 5.1. Assuming that this is a valid p.m.f. we can derive the p.g.f. as follows

$$P_X(t) = \sum_{x=0}^{\infty} t^x \binom{r+x-1}{x} p^r q^x = \left[\frac{p^r}{(1-tq)^r}\right] \sum_{x=0}^{\infty} \binom{r+x-1}{x} (1-tq)^r (tq)^x = \left[\frac{p}{1-tq}\right]^r.$$

If the distribution of the r.v. X is negative binomial with parameters r and p, then we will say that X is a negative binomial r.v. and indicate this by writing $X \sim$ negative binomial (r, p). Note that the negative binomial $(1, p)$ distribution is the same as the geometric (p) distribution.

Given a positive integer r, if we let X denote the number of trials up to but not including the trial when the r^{th} success occurs in a potentially infinite sequence of Bernoulli trials with success probability p, then X is a negative binomial (r, p) r.v. Note that this means that the r^{th} success occurs on trial $X+1$. Thus, as mentioned above, the geometric distribution is a special case of the negative binomial distribution.

We defined the binomial coefficient $\binom{n}{k}$ for $n, k \in \mathcal{Z}^+$ with $k \leq n$ and with the convention that $\binom{n}{0} = 1$. Note that for any $\alpha \in \mathcal{R}$ and $k \in \mathcal{Z}^+$ the expression

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}$$

is well defined. Adopting this expression as the definition of $\binom{\alpha}{k}$, with the convention that $\binom{\alpha}{0} = 1$ and the restriction that if $\alpha = n$ is an integer, then $\binom{n}{k} = 0$ whenever $k > n$ yields the following extended binomial theorem. We will use this theorem to prove that $\sum_{x=0}^{\infty} \binom{r+x-1}{x} p^r q^x = 1$.

Theorem 5.1 (Extended binomial theorem). *For $\alpha \in \mathcal{R}$ and $-1 < t < 1$*

$$(1+t)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} t^k = 1 + \alpha t + \left(\frac{\alpha(\alpha-1)}{2}\right) t^2 + \left(\frac{\alpha(\alpha-1)(\alpha-2)}{3!}\right) t^2 + \cdots$$

If $\alpha \in \mathcal{Z}^+$, then the terms with $k > \alpha$ vanish, the sum contains a finite number of terms, and this expression is valid for all values of t. If α is not a positive integer, then this is an infinite series.

Proof. Expanding $(1+t)^\alpha$ in a Taylor series about zero yields this expression. □

For $r \in \mathcal{Z}^+$ and $x = 0, 1, \ldots, r$

$$\binom{r+x-1}{x} = \frac{(r+x-1)(r+x-2)\cdots r}{x!} = \frac{r(r+1)\cdots(r+x-1)}{x!} = (-1)^x \binom{-r}{x}$$

since

$$r(r+1)\cdots(r+x-1) = (-1)^x(-r)(-r-1)(-r-2)\cdots(-r-x+1).$$

(It is this expression which leads to the name negative binomial.) Thus

$$\sum_{x=0}^{\infty} \binom{r+x-1}{x} q^x = \sum_{x=0}^{\infty} (-1)^x \binom{-r}{x} q^x = \sum_{x=0}^{\infty} \binom{-r}{x} (-q)^x = (1-q)^{-r} = p^{-r}$$

and

$$\sum_{x=0}^{\infty} \binom{r+x-1}{x} p^r q^x = p^r p^{-r} = 1.$$

Hypergeometric distribution.

For positive integers N_1, N_2, and n, with $n \leq N_1 + N_2$, the hypergeometric distribution with parameters N_1, N_2, and n has p.m.f.

$$f_X(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N_1+N_2}{n}} \mathbf{1}_{\Omega_X}(x),$$

where $\Omega_X = \{\max\{0, n-N_2\}, \ldots, \min\{n, N_1\}\}$ is the sample space. The p.g.f. for this distribution exists but is not useful. The fact that these probabilities sum to one follows from the argument given in the hypergeometric distribution example of Section 3.3. Note that, as discussed in Section 3.3, the hypergeometric distribution serves as a model for the distribution of the number of successes in a random sample of size n selected without replacement from a finite population containing N_1 objects classified as successes and N_2 objects classified as failures.

If the distribution of the r.v. X is hypergeometric with parameters N_1, N_2, and n, then we will say that X is a hypergeometric r.v. and indicate this by writing $X \sim$ hypergeometric (N_1, N_2, n).

Discrete uniform distribution.

For a positive integer N, the discrete uniform distribution on the integers $\{1, 2, \ldots, N\}$ has p.m.f.

$$f_X(x) = \frac{1}{N} \mathbf{1}_{\{1,2,\ldots,N\}}(x),$$

sample space $\Omega_X = \{1, \ldots N\}$, and p.g.f.

$$P_X(t) = \frac{t + t^2 + \cdots + p^N}{N}.$$

If the distribution of the r.v. X is uniform on the set $\{1, 2, \ldots, N\}$, then we will say that X is a discrete uniform r.v. and indicate this by writing $X \sim \text{uniform}(\{1, 2, \ldots, N\})$.

Poisson distribution.

For a positive number λ, the Poisson distribution with parameter λ has p.m.f.

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \mathbf{1}_{\{0,1,\ldots\}}(x),$$

sample space $\Omega_X = \{0, 1, \ldots\}$, and p.g.f.

$$P_X(t) = \exp[\lambda(t-1)].$$

The fact that $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = 1$ follows from the fact that $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$ is a power series expansion of e^{λ}. We can derive the p.g.f. as follows

$$P_X(t) = \sum_{x=0}^{\infty} t^x \frac{\lambda^x}{x!} e^{-\lambda} = \left[\frac{e^{-\lambda}}{e^{-t\lambda}}\right] \sum_{x=0}^{\infty} \frac{(t\lambda)^x}{x!} e^{-t\lambda} = \exp[\lambda(t-1)].$$

If the distribution of the r.v. X is Poisson with parameter λ, then we will say that X is a Poisson r.v. and indicate this by writing $X \sim \text{Poisson}\ (\lambda)$.

We will now provide an interesting connection between the binomial and Poisson distributions. In some situations there is interest in the behavior of a binomial (n, p) r.v. X when n is large, p is small, and $np = \lambda$ for some $\lambda > 0$. The Poisson (λ) distribution can be used to approximate the binomial (n, p) distribution in such a situation. More formally, if we let $n \to \infty$ and $p \to 0$ in such a way that $np = \lambda$, then,

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \to \frac{\lambda^x}{x!} e^{-\lambda}.$$

To see this note that, setting $p = \frac{\lambda}{n}$,

$$\begin{aligned}
\binom{n}{x} p^x (1-p)^{n-x} &= \left(\frac{n(n-1)\cdots(n-x+1)}{x!}\right) \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \left(\frac{n(n-1)\cdots(n-x+1)}{n^x}\right) \left(\frac{\lambda^x}{x!}\right) \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-x}
\end{aligned}$$

Letting $n \to \infty$, and noting that $\left(\frac{n(n-1)\cdots(n-x+1)}{n^x}\right) \to 1$, $\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$, and $\left(1-\frac{\lambda}{n}\right)^{-x} \to 1$ establishes the result.

The Poisson distribution often serves as a useful model for the distribution of the number of occurrences of an event in a fixed interval of time. The motivation of this application given below is not rigorous but it can be made so. Consider events occurring in time such as radioactive disintegrations. Assume that the conditions affecting the times of occurrence of the events are constant over time. Assume also that nonoverlapping time intervals are independent in the sense that information about the number of occurrences in one interval reveals nothing about the number of occurrences in another interval. We will argue that under these assumptions, and some others introduced below, the number X of events which occur in a fixed interval of time, taken to be the interval $(0,1)$ without loss of generality, is a Poisson (λ) r.v., with λ as defined below. Suppose that the interval is divided into a large number n of equal length subintervals and define a success as the occurrence of at least one event in such a subinterval. Since the n subintervals are of equal length the probability of success p_n is the same for each subinterval. Assume further that the probability of two or more events in an interval is negligible in the limit as $n \to \infty$. Let X denote the number of events in the entire interval $(0,1)$, then, for $x = 1, \ldots, n$, as $n \to \infty$ we have the binomial probability

$$\Pr(X = x) = \binom{n}{x} p_n^x (1-p_n)^{n-x}.$$

If we subdivide each of the n intervals into equal length subintervals, then $p_n = 2p_{2n} - p_{2n}^2$ since the event must occur in the left interval, the right interval, or both. Thus $p_n < 2p_{2n}$ and it appears that p_n is an increasing function of n (this can be proved rigorously). If the expected number of events in one of the n subintervals np_n tends to a limit λ as $n \to \infty$, then a slight modification of the argument above for the Poisson approximation to the binomial distribution yields the result that, as $n \to \infty$

$$\Pr(X = x) \to \frac{\lambda^x}{x!} e^{-\lambda}.$$

Thus under these assumptions the number of occurrences of the event in the interval $(0,1)$ is distributed as a Poisson (λ) r.v.

5.3 Joint, marginal, and conditional discrete distributions.

In this section we will introduce some concepts related to relationships among two or more random variables. We will concentrate our attention on the two variable (bivariate) case while providing some indication of how the concepts extend to higher dimensional (multivariate) cases.

We first extend the p.m.f from one dimension (univariate) to the two dimensional (bivariate) case. If two discrete random variables X and Y are defined on the same sample space, then their joint distribution is characterized by their joint probability mass function $f_{X,Y}$ where

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

i.e., the probability that $X = x$ and $Y = y$. The corresponding joint sample space (possible values of the pair (X, Y)) is $\Omega_{X,Y} = \{(x,y) \in \mathcal{R}^2 : f_{X,Y}(x,y) > 0\}$; and, we must have $f_{X,Y}(x,y) \geq 0$ for all $(x,y) \in \mathcal{R}^2$ and $\sum_{\{(x,y)\in\Omega_{X,Y}\}} f_{X,Y}(x,y) = 1$.

Note that the univariate p.m.f.'s f_X and f_Y, which are known as marginal p.m.f.'s in this context, can be obtained by summing the joint p.m.f.. For example, if $\Omega_Y = \{y_1, \ldots, y_n\}$, then, for $x \in \Omega_X$,

$$f_X(x) = \sum_{y\in\Omega_Y} f_{X,Y}(x,y) = f_{X,Y}(x,y_1) + \cdots + f_{X,Y}(x,y_n),$$

since the event $[X = x]$ is the union of the events $[X = x, Y = y_1]$, $\ldots$, $[X = x, Y = y_n]$ and these n events are disjoint.

Next we consider quantification of the effects of knowledge of the value of one variable on the distribution of another. For $x \in \Omega_X$ the conditional p.m.f. of Y given $X = x$, $f_{Y|X=x}(y|X = x)$, is defined by

$$f_{Y|X=x}(y|X = x) = \Pr(Y = y|X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

i.e., the conditional probability that $Y = y$ given that $X = x$. It is important to note that the conditional sample space for Y given $X = x$,
$\Omega_{Y|X=x} = \{y : f_{Y|X=x}(y|X = x) > 0\}$, may vary depending on the value of x; however, for each $x \in \Omega_X$, we must have $\sum_{y\in\Omega_{Y|X=x}} f_{Y|X=x}(y|X = x) = 1$.

The joint p.m.f. $f_{X,Y}$ is conveniently represented in a two way table as demonstrated in the following example. In this context the respective p.m.f.'s f_X and f_Y are known as the marginal p.m.f.'s of X and Y, since these probabilities are the (marginal) row and column sums corresponding to the tabular representation of the joint distribution. The conditional p.m.f. $f_{Y|X=x}$ is obtained by extracting the row corresponding to the fixed value x and

dividing the probabilities in this row by the corresponding marginal probability $f_X(x)$. The conditional p.m.f. $f_{X|Y=y}$ is obtained similarly from the $Y = y$ column and $f_Y(y)$.

Example. Suppose that a pair of fair dice is tossed once. Let X denote the sum of the two observed values and let Y denote the maximum of the two observed values. It is easily verified that the joint p.m.f. $f_{X,Y}(x, y)$ of X and Y corresponds to the probabilities in the body of the table below and the marginal p.m.f.'s of X and Y correspond to the probabilities along the margins.

Joint distribution of $X =$ sum and $Y =$ max.

	$Y = \max$						
$X = \text{sum}$	1	2	3	4	5	6	f_X
2	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{36}$
3	0	$\frac{2}{36}$	0	0	0	0	$\frac{2}{36}$
4	0	$\frac{1}{36}$	$\frac{2}{36}$	0	0	0	$\frac{3}{36}$
5	0	0	$\frac{2}{36}$	$\frac{2}{36}$	0	0	$\frac{4}{36}$
6	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	$\frac{5}{36}$
7	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{6}{36}$
8	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{5}{36}$
9	0	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{4}{36}$
10	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$
11	0	0	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$
12	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
f_Y	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

The conditional p.m.f. of X for a fixed value of Y is formed by selecting the appropriate column of the table (corresponding to the fixed value of Y) and normalizing the probabilities in the column so that they sum to one. The conditional p.m.f. of Y for a fixed value of X is formed analogously operating on rows. Some examples corresponding to fixing the value of the maximum are:

$$f_{X|Y=1}(x|Y = 1) = \mathbf{1}_{\{2\}}(x),$$

i.e., if we know that the maximum is 1, then we know that the sum is 2;

$$f_{X|Y=2}(x|Y = 2) = \frac{2}{3}\mathbf{1}_{\{3\}}(x) + \frac{1}{3}\mathbf{1}_{\{4\}}(x),$$

i.e., if we know that the maximum is 2, then the sum must be 3 or 4; with 3 twice as probable as 4;

$$f_{X|Y=4}(x|Y = 4) = \frac{2}{7}\mathbf{1}_{\{5,6,7\}}(x) + \frac{1}{7}\mathbf{1}_{\{8\}}(x)$$

i.e., if we know that the maximum is 4, then the sum must be 5, 6, 7, or 8; with 5, 6, and 7 equally probable and 8 half as probable as each of the other possible sums. Similarly, fixing the value of the sum yields:

$$f_{Y|X=6}(y|X=6) = \frac{1}{5}\mathbf{1}_{\{3\}}(y) + \frac{2}{5}\mathbf{1}_{\{4,5\}}(y),$$

i.e., if we know that the sum is 6, then the maximum must be 3, 4, or 5; with 4 and 5 being equally probable and 3 half as probable as each of the other possible maximums; and

$$f_{Y|X=7}(y|X=7) = \frac{1}{3}\mathbf{1}_{\{4,5,6\}}(y),$$

i.e., if we know that the sum is 7, then the maximum must be 4, 5, or 6 and these sums are equally probable.

As we can see from the preceding example, the joint p.m.f. of X and Y is not necessarily determined by the marginal p.m.f.'s. If $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathcal{R}$, then X and Y are said to be (stochastically) independent. If the joint p.m.f. does not factor in this way, then X and Y are (stochastically) dependent. The r.v.'s X = sum and Y = maximum in the dice tossing example are dependent. In this same example it is easy to verify that the r.v.'s Z = number on the first die and W = number on the second die are independent.

We will now establish discrete r.v. versions of the multiplication rule (Theorem 4.1) and the law of total probability (Theorem 4.6). First we establish the multiplication rule. If $(x,y) \in \Omega_{X,Y}$, then $f_{X,Y}(x,y) > 0$. Thus if $x \in \Omega_X$, then, since $f_X(x) = \sum_{y\in\Omega_Y} f_{X,Y}(x,y)$, there must be at least one y such that for this (x,y) pair $f_{X,Y}(x,y) > 0$ (such that this pair belongs to $\Omega_{X,Y}$). Similarly, if $f_X(x) = 0$, then for this x we must have $f_{X,Y}(x,y) = 0$ for all values of y. Hence, for all $(x,y) \in \Omega_{X,Y}$

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

with the convention that when $f_X(x) = 0$ and $f_{Y|X=x}(y|X=x)$ is undefined we define the corresponding product to be zero. Now we establish the law of total probability. Let $\Omega_X = \{x_1, x_2, \ldots\}$ and let $y \in \Omega_Y$ be given, then with the convention from above we have

$$\begin{aligned} f_Y(y) &= \sum_{x\in\Omega_X} f_{X,Y}(x,y) = \sum_{x\in\Omega_X} f_X(x)f_{Y|X=x}(y|X=x) \\ &= f_X(x_1)f_{Y|X=x_1}(y|X=x_1) + f_X(x_2)f_{Y|X=x_2}(y|X=x_2) + \cdots \end{aligned}$$

The concepts of joint, marginal, and conditional p.m.f.'s and independence are readily generalized to three or more r.v.'s. For example, given three r.v.'s X_1, X_2 and X_3 defined on the same sample space:

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = \Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3),$$

$$f_{X_1}(x_1) = \sum_{(x_2,x_3)\in\Omega_{X_2,X_3}} f_{X_1,X_2,X_3}(x_1,x_2,x_3),$$

$$f_{X_1,X_2}(x_1,x_2) = \sum_{x_3\in\Omega_{X_3}} f_{X_1,X_2,X_3}(x_1,x_2,x_3),$$

$$f_{X_1,X_2|X_3=x_3}(x_1,x_2|X_3=x_3) = \frac{f_{X_1,X_2,X_3}(x_1,x_2,x_3)}{f_{X_3}(x_3)},$$

and
X_1, X_2, and X_3 are (mutually) independent if, and only if,
$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = f_{X_1}(x_1)f_{X_2}(x_2)f_{X_3}(x_3)$ for all $x_1,x_2,x_3 \in \mathcal{R}$.

We will now present two parametric families of joint distributions, the multinomial and multiple hypergeometric distributions, which generalize the binomial and hypergeometric. For $k \geq 3$, let $\mathbf{X} = (X_1,\ldots,X_k)$ denote a vector of k random variables (a random vector) and let $\mathbf{x} = (x_1,\ldots,x_k)$ denote a potential value for $\mathbf{X}$. With this notation $f_{\mathbf{X}}(\mathbf{x})$ denotes the joint p.m.f. of $\mathbf{X} = (X_1,\ldots,X_k)$ and $\Omega_{\mathbf{X}}$ the corresponding joint sample space.

Multinomial distribution.

For $n \in \mathcal{Z}^+$, $k \geq 3$, and $0 < p_i < 1$ for $i = 1,\ldots,k$ with $p_1 + \cdots + p_k = 1$, the multinomial distribution with parameters n and $p_1,\ldots,p_k$ is the joint distribution of $\mathbf{X} = (X_1,\ldots,X_k)$ which has joint p.m.f.

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{n!}{x_1!x_2!\cdots x_n!} p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}\mathbf{1}_{\Omega_{\mathbf{X}}}(\mathbf{x}),$$

where the joint sample space $\Omega_{\mathbf{X}}$ is the set of nonnegative integer valued vectors $\mathbf{x}$ for which $x_1 + \cdots + x_k = n$.

Note that if $k = 2$, then the multinomial distribution with parameters n, p_1, and $p_2 = 1-p_1$ is degenerate, since $X_2 = n-X_1$, and the distribution reduces to the distribution of X_1 which is binomial with parameters n and $p = p_1$, *i.e.* under these conditions $\Pr(X_1 = x_1, X_2 = n - x_1) = \Pr(X_1 = x_1) = \binom{n}{x_1} p_1^{x_1}(1-p_1)^{n-x_1}$.

Example. If a fair die is tossed 10 times and we let X_1 denote the number of ones, X_2 the number of twos, and X_3 the number of values other than one or two, then the joint distribution of X_1, X_2, and X_3 is the multinomial distribution with $n = 10$, $p_1 = p_2 = \frac{1}{6}$, and $p_3 = \frac{4}{6}$. In this example $k = 3$ and this distribution is an example of a trinomial distribution.

Multiple hypergeometric distribution.

For $n \in \mathcal{Z}^+$, $k \geq 3$, and positive integers $N_1, \ldots, N_k$ with $N_1 + \cdots + N_k = N$, the multiple hypergeometric distribution with parameters n and $N_1, \ldots, N_k$ is the joint distribution of $\mathbf{X} = (X_1, \ldots, X_k)$ which has joint p.m.f.

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_k}{x_k}}{\binom{N}{n}}.$$

This expression only makes sense when $0 \leq x_i \leq N_i$ for $i = 1, \ldots, k$; however, this expression is valid for all integral values of the x_i provided we adopt the usual convention that $\binom{N_i}{x} = 0$ whenever $x > N_i$.

Note that if $k = 2$, then the multiple hypergeometric distribution with parameters n, N_1, and N_2 is degenerate, since $X_2 = n - X_1$, and the distribution reduces to the distribution of X_1 which is hypergeometric with parameters n, N_1, and $N - N_1$.

Example. If a poker hand is dealt randomly and we let X_1 denote the number of aces in the hand, X_2 the number of kings, and X_3 the number of cards which are neither aces nor kings, then the joint distribution of X_1, X_2, and X_3 is the multiple hypergeometric distribution with parameters $n = 5$, $N_1 = N_2 = 4$, and $N_3 = 44$. In this example $k = 3$ and this distribution is an example of a double hypergeometric distribution.

5.4 Sums of independent nonnegative integer valued random variables.

Let X and Y denote independent nonnegative integer valued r.v.'s with p.m.f.'s f_X and f_Y and p.g.f.'s P_X and P_Y and consider the distribution of the r.v. $Z = X + Y$. Note that Z is also a nonnegative integer valued r.v. The event $[Z = z]$ can be expressed as the union of the disjoint events $[X = 0, Y = z]$, $[X = 1, Y = z - 1]$, $\ldots$, $[X = z, Y = 0]$. Thus

$$f_Z(z) = f_X(0)f_Y(z) + f_X(1)f_Y(z-1) + \cdots + f_X(z)f_Y(0).$$

A p.m.f. f_Z of this form is said to be the convolution of the p.m.f.'s f_X and f_Y and this relationship is often denoted by $f_Z = f_X * f_Y$. It is straightforward to verify that $f_Z = f_X * f_Y$ is equivalent to $P_Z(t) = P_X(t)P_Y(t)$. This important property is summarized in the following theorem for ease of reference.

Theorem 5.2. *If X and Y are independent nonnegative integer valued r.v.'s with p.g.f.'s P_X and P_Y, then the p.g.f. of their sum $X+Y$ is the product of P_X and P_Y, i.e., $P_{X+Y}(t) = P_X(t)P_Y(t)$.*

The convolution can be generalized to convolutions of three or more nonnegative integer valued r.v.'s in the obvious way. For example, if X_1, X_2, and X_3 are three such r.v.'s, then $f_{X_1} * f_{X_2} * f_{X_3} = [f_{X_1} * f_{X_2}] * f_{X_3}$ and $P_{X_1+X_2+X_3}(t) = P_{X_1}(t)P_{X_2}(t)P_{X_3}(t)$. Convolutions are of particular interest when the distributions of the component r.v.'s are identical.

Applications to the binomial, negative binomial, and Poisson distributions are provided in the following theorems.

Theorem 5.3. *If $X_1, \ldots, X_k$ are independent r.v.'s and, for $i = 1, \ldots, k$, X_i has the binomial distribution with parameters n_i and p, then $X_1 + \cdots + X_k$ has the binomial distribution with parameters $n_1 + \cdots + n_k$ and p.*

Theorem 5.4. *If $X_1, \ldots, X_k$ are independent r.v.'s and, for $i = 1, \ldots, k$, X_i has the negative binomial distribution with parameters r_i and p, then $X_1 + \cdots + X_k$ has the negative binomial distribution with parameters $r_1 + \cdots + r_k$ and p.*

Theorem 5.5. *If $X_1, \ldots, X_k$ are independent r.v.'s and, for $i = 1, \ldots, k$, X_i has the Poisson distribution with parameter λ_i, then $X_1 + \cdots + X_k$ has the Poisson distribution with parameter $\lambda_1 + \cdots + \lambda_k$.*