Probability

Jan-April 2014

Chennai Mathematical Institute

Professor Nandini Kannan

Chapter 3

Random Variables

In several of the examples we considered in Chapter 1, the sample spaces consisted of symbols or categories. For example Head and Tail, or Defective and Nondefective. We are often interested in some function of the outcomes as compared to the actual individual outcomes.

Definition: Let $(\mathcal{S}, \mathcal{F}, P)$ be a **probability space**. A **random variable** $X(.)$ is a function from $\mathcal{S}$ into the real numbers.
The function $X(.)$ must be such that

$$A_r = \{s : X(s) \leq r\} \in \mathcal{F} \text{ for any real number } r.$$

In other words, the inverse image of Borel sets in $\mathcal{R}$ are events.
In terms of measure theory, we say that a random variable is a **measurable function**.

Definition: Let $(\mathcal{S}, \mathcal{F}, P)$ be a **probability space**, and X a random variable defined on this space. The random variable X **induces** a probability space $(\mathcal{R}, \mathcal{B}, P_X)$ by the correspondence

$$P_X(A) = P\{s : X(s) \in A\} \forall A \subset \mathcal{B}.$$

The range of X is usually denoted by $\mathcal{X}$. P_X is called the induced probability function and satisfies Kolmogorov's Axioms.

Theorem 3.0.1. *$P_X(.)$ is a probability function.*

Proof:

1. $P_X(A) = P[s : X(s) \in A] \geq 0.$
2. $P_X(\mathcal{X}) = P\{s : X(s) \in \mathcal{X}\} = 1.$
3. If B_i's are disjoint sets in $\mathcal{B}$, then

$$P_X\left[\bigcup_{i=1}^{\infty} B_i\right] = P\left[s : X(s) \in \bigcup_{i=1}^{\infty} B_i\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{\infty} P(s : X(s) \in B_i) \\
&= \sum_{i=1}^{\infty} P_X(B_i).
\end{aligned}
$$

■

Example: Toss a coin twice. Then

$$\mathcal{S} = \{HH, HT, TH, TT\}.$$

Let $X(s)$ = number of heads in s. Then $\mathcal{X} = \{0, 1, 2\}$.

$$\{s : X(s) \le r\} = \begin{cases} \phi, & r < 0; \\ \{TT\}, & 0 \le r < 1; \\ \{TT, HT, TH\}, & 1 \le r < 2; \\ \mathcal{S}, & r \ge 2, \end{cases}$$

i.e. the inverse image is an event.

$$P_X(X = 0) = P[s : X(s) = 0] = P[\{TT\}] = .25.$$

Similarly, $P_X(X = 1) = 0.5, P_X(X = 2) = 0.25$.

■

Definition: Cumulative Distribution Function. Let X be a random variable defined on a probability space. The **cumulative distribution function (cdf)** of X, denoted by $F_X(x)$ is defined by

$$F_X(x) = P_X(X \le x) = P_X((-\infty, x]) \qquad \forall x. \tag{3.1}$$

Theorem 3.0.2. *Let $F_X(x)$ be the cdf of the random variable X. Then the following three conditions hold:*

1. $\lim_{x\to-\infty} F_X(x) = 0; \qquad \lim_{x\to\infty} F_X(x) = 1.$
2. *$F_X(x)$ is a nondecreasing function of x; i.e. for all $a, b \in \mathcal{R}$, $a < b$, we have $F_X(a) \le F_X(b)$.*
3. *$F_X(x)$ is right continuous; i.e.* $\lim_{x\downarrow x_0} F_X(x) = F_X(x_0)$.

Proof:

1.

$$
\begin{aligned}
\lim_{x\to\infty} F_X(x) &= \lim_{x\to\infty} P_X(X \le x) \\
&= P(\mathcal{S}) = 1.
\end{aligned}
$$

Some of the steps need to be justified!

2. If $a < b$, then we know that

$$\{s : X(s) \leq a\} \subset \{s : X(s) \leq b\}.$$

Using the monotonicity property, we have

$$\begin{aligned} & F_X(a) = P_X(X \leq a) = P(s : X(s) \leq a) \\ & \leq P(s : X(s) \leq b) = P_X(X \leq b) = F_X(b). \end{aligned}$$

3. We need to show that

$$\lim_{h \downarrow 0} F_X(x+h) = F_X(x).$$

We have

$$\begin{aligned} \lim_{h \downarrow 0} P_X(x < X \leq x+h) &= \lim_{h \downarrow 0} [F_X(x+h) - F_X(x)] \\ &= \lim_{h \downarrow 0} F_X(x+h) - F_X(x). \end{aligned}$$

Let $\{h_i\}$ be a decreasing sequence of positive real numbers converging to 0. Let

$$A_i = [x < X < x + h_i].$$

Clearly, $A_1 \supset A_2 \supset \ldots$, with $A_n \downarrow \phi$. Using the continuity property of the probability function, we have

$$\lim_{i \to \infty} P_X(x < X \leq x + h_i) = \lim_{i \to \infty} P(A_i) = 0.$$

Therefore

$$\lim_{h \downarrow 0} F_X(x+h) = F_X(x).$$

■

Theorem 3.0.3. *For any random variable,*

$$P_X(X = x) = F_X(x) - F_X(x-),$$

for all $x \in \mathcal{R}$, where $F_X(x-) = \lim_{z \uparrow x} F_X(z)$.

Proof: Define

$$A_n = \left(x - \frac{1}{n}, x\right]$$

$\{A_n\}$ is a decreasing sequence of sets, decreasing to $\{x\}$. Using the continuity property of the Probability function, we have

$$\begin{aligned} P(X = x) &= P\left[\bigcap_{n=1}^{\infty} A_n\right] \\ &= \lim_{n\to\infty} P\left[A_n\right] \\ &= \lim_{n\to\infty} \left[F_X(x) - F_X(x - (1/n))\right] \\ &= F_X(x) - F_X(x-). \end{aligned}$$

■

Example: (contd.) Toss a coin twice. Then

$$\mathcal{S} = \{HH, HT, TH, TT\}.$$

Let $X(s) =$ number of heads in s.

$$F_X(x) = \begin{cases} 0, & x < 0; \\ .25, & 0 \le x < 1; \\ .75, & 1 \le x < 2; \\ 1, & x \ge 2. \end{cases}$$

The cdf is a step function with jumps at $0, 1, 2$. The size of the jumps represent the probability at those points.

$$P_X(X = 0) = P[s : X(s) = 0] = P[\{TT\}] = .25.$$

Similarly, $P_X(X = 1) = 0.5, P_X(X = 2) = 0.25$.
■

Example: Show that $F(.)$ is a cdf, where

$$F_X(x) = 1 - e^{-x}; \qquad x \in (0, \infty).$$

■

Example: Consider the function

$$F_X(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{2}, & x = 0; \\ \frac{1}{2} + \frac{x}{2}, & 0 < x < 1; \\ 1, & x \ge 1. \end{cases}$$

This is a mixture of continuous pieces and jumps. The function has a jump at $x = 0$, and F is continuous in the interval $(0, 1)$.
■

<u>Definition:</u> A random variable X is **discrete** if its cdf is a step function of x.
We may also use the following definition:

<u>Definition:</u> A random variable X is **discrete** if its range is countable.

<u>Definition:</u> The probability mass function (pmf) of a discrete random variable is given by

$$p_X(x) = P(X = x). \tag{3.2}$$

Theorem 3.0.4. *Let X be a discrete rv with pmf $p_X(x)$. Then*

(a) $p_X(x) \geq 0 \qquad \forall x$;

(b) $\sum_x p_X(x) = 1$.

■

The pmf of a discrete random variable is a formula, table or graph that associates a probability with each value of the random variable.

Example: Roll a fair die twice

$$\mathcal{S} = \left\{ \begin{array}{llll} (1,1) & . & . & (1,6) \\ (2,1) & . & . & (2,6) \\ . & . & . & . \\ . & . & . & . \\ (6,1) & . & . & (6,6) \end{array} \right\}$$

Let Z denote the maximum of the two rolls. Find the pmf of Z.

We could also draw what is called a **probability histogram**. The X-axis has the values of the random variable, the Y-axis has the probabilities. Rectangles are constructed with the value being the midpoint and height equal to the corresponding probability.

Example: Consider families with two children.
$\mathcal{S}$ = {GG, GB, BG, BB}. B=boy, G= girl.

Let X be the random variable defined as the number of girls in the family. If we assume births are independent, and that the child is as likely to be a boy as a girl, then each of the four outcomes is equally likely to occur.

The event that X = 0 is equivalent to BB. X = 1 is equivalent to BG or GB and X = 2 is equivalent to GG. The pmf. of X is

k	0	1	2	Total
p(k)	0.25	0.50	0.25	1

3.1 Transformations

Let X be a random variable with cdf $F_X(x)$. Sometimes, we are interested not in X, but a transformed version of X. Let $Y = g(X)$ be a function of X.
Clearly, Y is also a random variable. We can describe the probabilistic behaviour of Y in terms of that of X. We have

$$P[Y \in A] = P[g(X) \in A] \quad \text{for any} \quad A.$$

Let $\mathcal{X}$ be the sample space of X. Then

$$g : \mathcal{X} \to \mathcal{Y}$$

where $\mathcal{Y}$ is the sample space of the random variable Y.
We can define the inverse mapping

$$g^{-1} : \mathcal{Y} \to \mathcal{X}$$

where

$$g^{-1}(A) = \{x \in \mathcal{X}; g(x) \in A\}.$$

Therefore

$$\begin{aligned} P[Y \in A] &= P[g(X) \in A] \\ &= P[\{x \in \mathcal{X}; g(x) \in A\}] \\ &= P[X \in g^{-1}(A)]. \end{aligned}$$

This defines the probability distribution of Y. This distribution satisfies Kolmogorov's axioms.

If X is a discrete r.v. with pmf $p_X(.)$, then Y is also discrete. We have

$$\begin{aligned} p_Y(y) &= P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x) \\ &= \sum_{x \in g^{-1}(y)} p_X(x) \qquad for \;\; y \in \mathcal{Y}. \end{aligned}$$

Here

$$g^{-1}(y) = \{x \in \mathcal{X} : g(x) = y\}.$$

The pmf of Y may be easily obtained by first identifying $g^{-1}(y)$ and then summing the appropriate probabilities.

Example: Let X have a pmf

$$p_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n; 0 \le p \le 1.$$

Find the pmf of $Y = X^2$. We have

$$\mathcal{X} = \{0, 1, \ldots, n\}; \mathcal{Y} = \{0, 1, 4, \ldots, n^2\}.$$

For any $y \in \mathcal{Y}$, $g(x) = x^2 = y$ iff $x = \sqrt{y}$.

$$\begin{aligned} &p_Y(y) = P(Y = y) = P(X = \sqrt{y}) \\ &= \binom{n}{\sqrt{y}} p^{\sqrt{y}} (1-p)^{n-\sqrt{y}}, \qquad y = 0, 1 \ldots, n^2. \end{aligned}$$

■

3.2 Expectation and Moments

The study of the probability distributions of random variables is essentially the study of certain numerical characteristics (parameters) associated with these distributions.
These parameters of the distribution give us some idea about the behaviour of the rv's.
The expected value of a random variable is simply a weighted average: the weights are specified according to the probability distribution. The expected value is a measure of central tendency.

Definition: Let X be a random variable. The expected value or mean of the random variable X is

$$E[g(X)] = \sum_x x p_X(x), \tag{3.3}$$

provided the sum exists. If $E|X| = \infty$, we say that $E(X)$ **does not exist**.

Example: Let X be a discrete random variable with

$$P\left[X = (-1)^{j+1}\frac{3^j}{j}\right] = \frac{2}{3^j}, \qquad j = 1, 2, \dots.$$

$$\sum_{j=1}^{\infty} \frac{2}{3^j} = \frac{2}{3} + \frac{2}{3^2} + \frac{2}{3^3} + \dots = \frac{2}{3}\left[1 + \frac{1}{3} + \frac{1}{3^2} + \dots\right] = 1.$$

We have

$$\sum_{j=1}^{\infty} |x_j| p_j = \sum_{j=1}^{\infty} \frac{3^j}{j}\frac{2}{3^j} = \sum_{j=1}^{\infty} \frac{2}{j} = \infty,$$

which implies $E(X)$ does not exist. However

$$\sum_{j=1}^{\infty} x_j p_j = \sum_{j=1}^{\infty} (-1)^{j+1}\frac{2}{j},$$

is convergent.

■

Theorem 3.2.1. *Let X be a discrete random variable and let $g(X)$ be any real-valued function of X. Then*

$$E[g(X)] = \sum_x g(x) p_X(x),$$

or

$$E[Y] = \sum_y y \; p_Y(y),$$

where $Y = g(X)$ and $p_Y(y)$ is the pmf of Y.

Proof:

■

Corollary 3.2.2. *Let X be a discrete random variable and let a, b be constants. Then*

$$E[aX + b] = aE(X) + b.$$

■

Definition: The r-th **moment** of a random variable X is

$$\mu_r' = E[X^r]. \tag{3.4}$$

The r-th **central moment** is

$$\mu_r = E[(X - \mu)^r], \tag{3.5}$$

where $\mu = \mu_1' = E(X)$ is the expected value or mean of X.

Definition: The **variance** of a random variable is its second central moment.

$$\sigma^2 = Var(X) = E[(X - \mu)^2] = E(X^2) - \mu^2.$$

The positive square root of the variance is called the **standard deviation**.

Remarks:

- The variance and the standard deviation provide a measure of spread of the distribution about its mean.
- $Var(X) = 0$ iff X is a degenerate random variable, i.e. X is constant with probability 1. This implies no variation in X.

Theorem 3.2.3.

$$Var(aX + b) = a^2 Var(X).$$

Proof: We have

$$E(aX + b) = a\mu + b.$$

Therefore

$$Var(aX + b) = E[aX + b - a\mu - b]^2 = E[a(X - \mu)]^2 = a^2 Var(X).$$

■

3.3 Special Families of Discrete Distributions

A parametric family of distributions is a collection of mass functions indexed by one or more parameters. Varying the parameter(s) allows us to change certain characteristics of the distribution while staying with one functional form. Many of these families arise from experiments with special properties.

3.3.1 Discrete Uniform Distribution

Definition: A random variable X has a discrete uniform $(1, N)$ distribution if

$$p_X(k|N) = P(X = k|N) = \frac{1}{N} \qquad k = 1, 2, \ldots, N. \tag{3.6}$$

Here N is some specified integer. The distribution assigns equal probability to all the outcomes.
Clearly, the pmf satisfies both properties listed in Theorem 3.0.4. The notation $p_X(k|N)$ specifies the dependence on the parameter N.

Theorem 3.3.1. *If X is a discrete uniform $(1, N)$ random variable, then*

$$E(X) = \frac{N+1}{2}; \qquad E(X^2) = \frac{(N+1)(2N+1)}{6} \tag{3.7}$$

$$\sigma^2 = Var(X) = \frac{N^2-1}{12}. \tag{3.8}$$

Proof: Using the definition of expectation, we have

$$E(X) = \sum_k k p_X(k) = \sum_{k=1}^{N} k \frac{1}{N} = \frac{N(N+1)}{2N} = \frac{N+1}{2}.$$

$$E(X^2) = \sum_{k=1}^{N} k^2 \frac{1}{N} = \frac{N(N+1)(2N+1)}{6N} = \frac{(N=1)(2N+1)}{6}.$$

The variance is obtained by substituting in the formula

$$\sigma^2 = E(X^2) - [E(X)]^2.$$

■

The distribution may be generalized so that the sample space is any range of integers $N_0, N_0 + 1, \ldots, N_1$. Then

$$p_X(k|N_0, N_1) = \frac{1}{N_1 - N_0 + 1}. \tag{3.9}$$

3.3.2 Hypergeometric Distribution

Consider a population of N items, M of which are of Type I and $N-M$ of Type II. Consider a sample of n items selected at random from this population without replacement. Define the random variable X to be the number of items in the sample that are of Type I. We have

$$p_X(k|M,N,n) = \frac{\begin{pmatrix} M \\ k \end{pmatrix}\begin{pmatrix} N-M \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}. \tag{3.10}$$

The values of the random variable must satisfy the following constraints

$$\begin{aligned} &x \le \min(M,n) \\ &n - x \le N - M \Rightarrow x \ge M - N + n \\ &\Rightarrow x \ge \max(0, M-N+n). \end{aligned}$$

The pmf in (3.10) is called the **Hypergeometric distribution**, and the corresponding random variable X is a **hypergeometric** random variable.

The probabilities defined in (3.10) are clearly non-negative. To show the probabilities sum to 1, we need a combinatorial identity:

$$\sum_{k=0}^{n} \begin{pmatrix} a \\ k \end{pmatrix}\begin{pmatrix} b \\ n-k \end{pmatrix} = \begin{pmatrix} a+b \\ n \end{pmatrix}.$$

Consider the identity

$$(1+t)^a(1+t)^b = (1+t)^{a+b}.$$

Consider the coefficient of t^n in both sides:

$$\begin{pmatrix} a \\ 0 \end{pmatrix}\begin{pmatrix} b \\ n \end{pmatrix} + \begin{pmatrix} a \\ 1 \end{pmatrix}\begin{pmatrix} b \\ n-1 \end{pmatrix} + \ldots + \begin{pmatrix} a \\ n \end{pmatrix}\begin{pmatrix} b \\ 0 \end{pmatrix} = \begin{pmatrix} a+b \\ n \end{pmatrix}.$$

Using this identity, we may show the probabilities in (3.10) sum to 1.

Theorem 3.3.2. *If X is a hypergeometric random variable, then*

$$E(X) = \frac{Mn}{N}; \qquad E[(X(X-1)] = \frac{M(M-1)n(n-1)}{N(N-1)} \tag{3.11}$$

$$\sigma^2 = Var(X) = \frac{Mn}{N}\left[\frac{(N-M)(N-n)}{N(N-1)}\right]. \tag{3.12}$$

Proof: Using the definition of expectation, we have

$$
\begin{aligned}
E(X) &= \sum_{k=0}^{n} k \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \\
&= \sum_{k=1}^{n} k \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \\
&= \sum_{k=1}^{n} \frac{Mn}{N} \frac{\binom{M-1}{k-1}\binom{N-M}{n-k}}{\binom{N-1}{n-1}} \\
&= \frac{Mn}{N} \sum_{j=0}^{n-1} \frac{\binom{M-1}{j}\binom{N-M}{n-1-j}}{\binom{N-1}{n-1}} \\
&= n\,\frac{M}{N},
\end{aligned}
$$

since the terms in the summation are the probabilities for a hypergeometric random variable with parameters $N-1, M-1, n-1$, and therefore must sum to 1.
Similarly,

$$
\begin{aligned}
E[(X)(X-1)] &= \sum_{k=0}^{n} k(k-1) \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \\
&= \sum_{k=2}^{n} M(M-1) \frac{\binom{M-2}{k-2}\binom{N-M}{n-k}}{\binom{N}{n}} \\
&= \frac{M(M-1)}{\frac{N(N-1)}{n(n-1)}} \sum_{k=2}^{n} \frac{\binom{M-2}{k-2}\binom{N-M}{n-k}}{\binom{N-2}{n-2}}
\end{aligned}
$$

$$= \frac{M(M-1)n(n-1)}{N(N-1)},$$

since the terms in the summation are the probabilities for a hypergeometric random variable with parameters $N-2, M-2, n-2$, and therefore must sum to 1.
We know that

$$Var(X) = E(X^2) - [E(X)]^2 = E[X^2 - X] + E(X) - [E(X)]^2.$$

Substituting for $E(X)$ and $E[X(X-1)]$, we have

$$\begin{aligned} Var(X) &= \frac{M(M-1)n(n-1)}{N(N-1)} + \frac{Mn}{N} - \left[\frac{Mn}{N}\right]^2 \\ &= \frac{Mn}{N}\left[\frac{(N-M)(N-n)}{N(N-1)}\right]. \end{aligned}$$

■

Remark: The hypergeometric probabilities satisfy a recurrence relationship:

$$\frac{P(X=k+1)}{P(X=k)} = \frac{(M-k)(n-k)}{(k+1)(N-M-n+k+1)}. \tag{3.13}$$

3.3.3 The Binomial Distribution

Definition: A **Bernoulli trial** is an experiment with only two possible outcomes. We classify these outcomes as *Success* and *Failure*.

Definition: For any Bernoulli trial, we define the random variable X as follows: if the trial results in a Success, $X = 1$; otherwise $X = 0$. The pmf of X is given by

$$P(X=1) = p \qquad P(X=0) = 1-p, \qquad 0 \le p \le 1. \tag{3.14}$$

The random variable X is said to have a **Bernoulli(p)** distribution. We may combine the two probabilities into a single expression:

$$P(X=x) = p^x(1-p)^{1-x}, \qquad x = 0, 1, \qquad 0 \le p \le 1. \tag{3.15}$$

Examples

1. Toss a fair coin: Heads and Tails
2. Test a blood sample for Absence or Presence of a particular disease
3. Test items in a factory: Defective or Nondefective

Theorem 3.3.3. *If X is a Bernoulli(p) random variable, then*

$$E(X) = p \tag{3.16}$$
$$\sigma^2 = Var(X) = p(1-p). \tag{3.17}$$

Proof: We have

$$\mu = E(X) = 1 \times p + 0 \times (1-p) = p.$$

$$E(X^2) = 1^2 \times p + 0^2 \times (1-p) = p.$$

Therefore

$$\sigma^2 = Var(X) = p - p^2 = p(1-p).$$

■

Definition: An experiment is said to be a **Binomial** experiment if it satisfies the following properties:

1. The experiment consists of n identical Bernoulli trials.
2. The probability of success on a single trial is equal to p and remains constant from trial to trial. The probability of failure is then $1-p$ which is denoted by q.
3. The trials are independent.

Let Y be the random variable that records the total number of successes in the n trials. Y can take values $0, 1, \ldots, n$. The random variable Y is called a **Binomial random variable**. We usually write

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

Define

$$X_i = \begin{cases} 1, & \text{if } i\text{th trial results in a success;} \\ 0, & \text{if } i\text{th trial results in a failure.} \end{cases}$$

Then $Y = \displaystyle\sum_{i=1}^{n} X_i$. The pmf of Y is given by

$$p_Y(k|n,p) = P(Y = k|n,p) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0, 1, \ldots, n \tag{3.18}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

In the general formula the term

$$p^k(1-p)^{n-k}$$

is the probability of obtaining one string of exactly k successes and $(n-k)$ failures (assuming the trials are independent). The term

$$\begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{k!(n-k)!}$$

counts the number of distinct strings with exactly k successes and $(n-k)$ failures.

Each term in (3.18) is non-negative. To show the probabilities sum to 1, we use the following:

$$\begin{aligned} \sum_{k=0}^{n} p_Y(k|n,p) &= q^n + \begin{pmatrix} n \\ 1 \end{pmatrix} pq^{n-1} + \ldots + \begin{pmatrix} n \\ n-1 \end{pmatrix} p^{n-1}q + p^n \\ &= (p+q)^n \qquad \text{Binomial Theorem} \\ &= 1. \end{aligned}$$

Tables of Binomial probabilities are available for different values of n and p.

Example: When a customer places an order with an online office supply store, a computerized accounting information system (AIS) automatically checks to see if the customer has exceeded their credit limit. Past records indicate that 5% of customers exceed their credit limits. On a given day, 20 customers place orders.
(a) Is this a binomial experiment?
(b)Find the probability that a majority of customers are over their credit limit.

■

Theorem 3.3.4. *If* $Y \sim Bin(n,p)$, *then*

$$\mu = np \tag{3.19}$$
$$\sigma^2 = n\ p\ q. \tag{3.20}$$

<u>**Proof:**</u>

■

Relation between the Hypergeometric and the Binomial

Consider the hypergeometric distribution. Let $p = M/N$ be the proportion of Type I items in the population. Suppose N is large with n and p remaining fixed, i.e.

$$N \to \infty, \qquad M \to \infty, \qquad \frac{M}{N} \to p.$$

Then the hypergeometric probabilities converge to the binomial probabilities. For large samples, there is practically no difference between sampling with and without replacement. We have

$$P(X = k|N, M, n) = \frac{\begin{pmatrix} M \\ k \end{pmatrix}\begin{pmatrix} N-M \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

$$= \begin{pmatrix} n \\ k \end{pmatrix} \frac{M!(N-M)!(N-n)!}{(M-k)!(N-M-n+k)!N!}$$

$$= \begin{pmatrix} n \\ k \end{pmatrix} \frac{M(M-1)\ldots(M-k+1)(N-M)\ldots(N-M-n+k+1)}{N^k\left[1.\left(1-\frac{1}{N}\right)\ldots\left(1-\frac{(n-1)}{N}\right)\right]}$$

$$\rightarrow \begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k}.$$

Remark: The binomial distribution may be used as an approximation if $n/N \leq 0.05$.

3.3.4 Poisson Distribution

The distribution of count data is often modeled using the **Poisson distribution**.
Let N_t be the total number of occurrences in an interval of length t, i.e. in $(0, t)$.
The Poisson distribution is derived from the **Poisson process** which possesses the following properties, called the **Poisson postulates**:

1. $N_0 = 0$.
2. The probability of exactly one occurrence during a very small interval $[t, t+h]$ of length h, is proportional to the length of the interval, i.e.
$$P(N_h = 1) = \lambda h + o(h).$$
3. The probability of more than one occurrence in the interval $[t, t+h]$ is $o(h)$.
4. The number of occurrences in non-overlapping intervals are independent.

Remark: Here $o(h)$ is a quantity of smaller order of magnitude than h, i.e.

$$\frac{o(h)}{h} \rightarrow 0 \quad \text{as} \quad h \rightarrow 0.$$

Result: Under these postulates, we have

$$P(N_t = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \qquad k = 0, 1, \ldots. \tag{3.21}$$

Proof: We have

$$\begin{aligned}
P_n(t+h) &= P(N_{t+h} = n) \\
&= P(N_t = n, N_{t+h} - N_t = 0) + P(N_t = n-1, N_{t+h} - N_t = 1) \\
&= \sum_{k=2}^{n} P(N_t = n-k, N_{t+h} - N_t = k) \\
&= P_n(t)P_0(h) + P_{n-1}(t)P_1(h) + o(h) \qquad \text{using P1, P3} \\
&= (1-\lambda h)P_n(t) + \lambda h P_{n-1}(t) + o(h) \qquad \text{using P2.}
\end{aligned}$$

Therefore

$$\frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h}.$$

Letting $h \to 0$, we have

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t). \tag{3.22}$$

For $n = 0$, we get

$$P_0(t+h) = P_0(t)[1 - \lambda h] + o(h).$$

Rearranging the terms as before, and letting $h \to 0$, we have

$$\begin{aligned} &P_0'(t) = -\lambda P_0(t) \\ &\Rightarrow P_0(t) = Ke^{-\lambda t}. \end{aligned}$$

Since $P_0(0) = 1$, we have

$$P_0(t) = e^{-\lambda t}.$$

Using the differential equation in (3.22) and the initial conditions, we get

$$\begin{aligned} P_1'(t) &= -\lambda P_1(t) + \lambda P_0(t) \\ &= -\lambda P_1(t) + \lambda e^{-\lambda t}. \end{aligned}$$

This implies

$$P_1(t) = \lambda t e^{-\lambda t}.$$

Proceeding in a similar fashion, we can derive the general case:

$$\begin{aligned} &P_n'(t) + \lambda P_n(t) = \lambda P_{n-1}(t) \\ &\Rightarrow e^{\lambda t}[P_n'(t) + \lambda P_n(t)] = \lambda e^{\lambda t} P_{n-1}(t) \\ &\Rightarrow \frac{d}{dt}[e^{\lambda t} P_n(t)] = \lambda e^{\lambda t} P_{n-1}(t) \\ = \quad &\lambda e^{\lambda t} \left[\frac{e^{-\lambda t}(\lambda t)^{n-1}}{(n-1)!} \right] \\ &\Rightarrow P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \qquad n = 0, 1, \ldots. \end{aligned}$$

■

Definition: A random variable is said to have a Poisson distribution $P(\lambda)$ if

$$p_X(x|\lambda) = P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots; \lambda > 0. \tag{3.23}$$

The probabilities are clearly non-negative. Further

$$\begin{aligned}\sum_{k=0}^{\infty} p_X(x|\lambda) &= \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{k=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda}e^{\lambda} = 1,\end{aligned}$$

since the infinite sum is the Taylor expansion for e^{λ}.

Example: Ernest Rutherford conducted a series of experiments on radioactive decay. In one of these, a radioactive substance was observed in N = 2608 time intervals of 7.5 seconds each, and the number of decay particles reaching a counter during each period was recorded. The data can be fit to a Poisson distribution.

Theorem 3.3.5. *If If $X \sim P(\lambda)$, then*

$$\mu = \lambda \tag{3.24}$$
$$\sigma^2 = \lambda. \tag{3.25}$$

Proof:

■

Remark: We have the following recursion formula for the Poisson probabilities:

$$P(X = x) = \frac{\lambda}{x}\, P(X = x - 1). \tag{3.26}$$

Example: The number of orders for pizza at Pizza Hut has a Poisson distribution with an average of 12 per hour.
a. What is the probability that exactly five orders are received during a particular hour?
b. If the phones are down for 30 minutes, what is the probability that no orders will be missed?

■

Poisson as a limiting form of the Binomial

The binomial distribution can be approximated by the Poisson distribution for large n and small p for $\lambda = np$ of moderate magnitude.

Theorem 3.3.6. *Let $X \sim Bin(n, p)$. Then*

$$\binom{n}{x} p^x q^{n-x} \to \frac{e^{-\lambda}\lambda^x}{x!}$$

as $n \to \infty, p \to 0$ with $np \to \lambda$.

Proof: We have

$$P(X = 0) = (1-p)^n = \left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda} \quad \text{as} \quad n \to \infty.$$

For the binomial distribution, we have the recursion

$$P(X = x) = \frac{n-x+1}{x}\frac{p}{1-p}P(X = x-1).$$

Therefore

$$\frac{P(X = x)}{P(X = x-1)} = \frac{np - (x-1)p}{x\ q} = \frac{\lambda - (x-1)p}{x\ q} \approx \frac{\lambda}{x}$$

for small p. Therefore

$$P(X = x) \approx \frac{\lambda}{x} P(X = x - 1).$$

Using the value for $P(X = 0)$, we can see that

$$P(X = 1) = \lambda e^{-\lambda} \tag{3.27}$$

$$P(X = 2) = \left(\frac{\lambda}{2}\right) \lambda e^{-\lambda} = \frac{e^{-\lambda}\lambda^2}{2!} \tag{3.28}$$

$$\dots$$

■

Example: The IT office collects data on the number of major network errors experienced each day on a local area network. Network errors occur infrequently. Past data indicate the probability of a major error is 0.001. What is the probability that in a period of a year (365 days) only one network error will occur?

■

3.3.5 Negative Binomial Distribution

Consider a sequence of independent Bernoulli trials. Define the random variable X to be the trial at which the r-th success occurs, where r is a fixed integer. We have

$$P(X = x|r, p) = \left(\begin{array}{c} x - 1 \\ r - 1 \end{array} \right) p^r q^{x-r}, \;\; x = r, r + 1, r + 2 \dots \tag{3.29}$$

We may also define Y to be the random variable that counts the number of failures observed before the r-th success. Then

$$Y = X - r.$$

The pmf of Y is given by

$$p_Y(y|r,p) = P(Y = y|r,p) = \begin{pmatrix} r+y-1 \\ y \end{pmatrix} p^r q^y, \qquad y = 0, 1, 2 \dots \tag{3.30}$$

The random variable X (or Y) is said to have a **Negative Binomial** distribution, and we write

$$Y \sim NB(r,p).$$

For the pmf in (3.29), the term

$$p^r q^{x-r}$$

is the probability of obtaining a success on the $x-$th trial preceded by $r-1$ successes and $x-r$ failures in some specified order. The term

$$\begin{pmatrix} x-1 \\ r-1 \end{pmatrix}$$

counts the number of possible strings.

Example: The percentage of individuals in the population possessing a rare blood type is 2 %. Individuals arrive at a blood bank and are tested. The moment two matches are located, the testing stops.
Find the probability that exactly 5 individuals are tested.

■

Let us define

$$\binom{x}{r} = \frac{x(x-1)\dots(x-r+1)}{r!}$$

for all values of x and all positive integers r. We use the convention that

$$\binom{x}{0} = 1.$$

We have

$$\binom{-1}{r} = \frac{(-1)(-1-1)\dots(-1-r+1)}{r!} = (-1)^r.$$

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

Consider

$$\begin{aligned}\binom{r+y-1}{y} &= \frac{(r+y-1)!}{y!(r-1)!} \\ &= \frac{(r+y-1)(r+y-2)\dots r}{y!} \\ &= \frac{(-r)(-r-1)\dots(-r-y+1)(-1)^y}{y!}\end{aligned}$$

$$= \binom{-r}{y}(-1)^y.$$

The negative binomial pmf is then given by

$$P(Y = y|r, p) = \binom{-r}{y}(-1)^y p^r q^y, \qquad y = 0, 1, 2 \ldots$$

which resembles the binomial distribution.

Result: The binomial theorem or formula $(1 + t)^a$ for any number a and all values $-1 < t < 1$ is

$$(1+t)^a = 1 + \binom{a}{1} t + \ldots + \binom{a}{k} t^k + \ldots$$

If a is a positive integer, all terms containing powers higher than t^a vanish.

Using this result, we have

$$\begin{aligned}
\sum_{y=0}^{\infty} P(Y = y|r, p) &= \sum_{y=0}^{\infty} \binom{-r}{y}(-1)^y p^r q^y \\
&= p^r \sum_{y=0}^{\infty} \binom{-r}{y}(-q)^y \\
&= p^r \left[1 + \binom{-r}{1}(-q) + \ldots + \binom{-r}{k}(-q)^k + \ldots\right] \\
&= p^r (1-q)^{-r} = 1.
\end{aligned}$$

Theorem 3.3.7. *Let $Y \sim NB(r, p)$. Then*

$$\mu = \frac{rq}{p} \tag{3.31}$$

$$\sigma^2 = \frac{rq}{p^2}. \tag{3.32}$$

Proof:

■

Definition: If $r = 1$, we get the **geometric random variable**, X the number of trials required for a single success. The probability of X taking the value x is given by

$$P(X = x|p) = pq^{x-1}, \qquad x = 1, 2, 3 \ldots \tag{3.33}$$

We can also define the geometric random variable as Y, the number of failures preceding the first success. We have $Y = X - 1$.

$$P(Y = y|p) = pq^{y}, \qquad x = 0, 1, 2, \ldots \tag{3.34}$$

Example: A process that fills packages is stopped whenever a package is detected whose weight falls outside the specification. Assume the probability of falling outside the specification limits is 0.01. Find the probability that 10 packets are filled before the process stops.

■

Theorem 3.3.8. *If X has a geometric distribution, then for any two positive integers s, t, with $s > t$,*

$$P(X > s|X > t) \;=\; P[X > s - t]. \tag{3.35}$$

This is the lack of memory property.

Proof: We have

$$\begin{aligned} P(X > n) &= P(\text{no successes in } n \text{ trials}) \\ &= (1 - p)^n. \end{aligned}$$

$$\begin{aligned} P(X > s|X > t) &= \frac{P(X > s)}{P(X > t)} = (1 - p)^{s-t} \\ &= P(X > s - t). \end{aligned}$$

■

The converse of the result is also true.

Theorem 3.3.9. *If X is an integer valued random variable satisfying*

$$P(X > s|X > t) \;=\; P[X > s - t].$$

for any two positive integers s, t, with $s > t$, then X is a geometric random variable.

Proof: Let

$$P(X = x) = p_x \qquad x = 1, 2, \ldots.$$

$$P(X > s - t) = \sum_{k=s-t+1}^{\infty} p_k = q_{s-t} \text{ say.}$$

$$P(X > s | X > t) = \frac{P(X > s)}{P(X > t)} = \frac{q_s}{q_t}.$$

Using the lack of memory property, we have

$$q_s = q_t \; q_{s-t}.$$

If $s - t = 1$, then

$$q_{t+1} = q_t \; q_1 \qquad q_1 = \sum_{k=2}^{\infty} p_k = 1 - p_1.$$

Therefore

$$q_{t+1} = q_t \; (1 - p_1).$$

Substituting values for t, we have

$$\begin{aligned} q_2 &= q_1 \; (1 - p_1) = (1 - p_1)^2 \\ q_3 &= q_2 \; (1 - p_1) = (1 - p_1)^3 \\ &\vdots \\ q_k &= (1 - p_1)^k. \end{aligned}$$

Therefore

$$\begin{aligned} p_k &= q_{k-1} - q_k = (1 - p_1)^{k-1} - (1 - p_1)^k \\ &= p_1(1 - p_1)^k, \qquad k - 1, 2, \ldots, \end{aligned}$$

which is the geometric pmf.

■

Result: If $Y \sim Bin(n, p)$ and $X \sim NB(r, p)$, then

$$P(X \leq n) = P(Y \geq r), \tag{3.36}$$

i.e. if there are r or more successes in the first n trials, then at most n trials were required to obtain the first r of these successes.