CHAPTER 5

SOME DISCRETE PROBABILITY DISTRIBUTIONS

As we had discussed, there are two main types of random variables, namely, *discrete random variables* and *continuous random variables*. In this chapter, we consider only discrete random variables.

A discrete random variable often involve a count of something, such as the number of cars owned by a randomly selected family, the number of people waiting for a haircut in a barbershop, or the number of households in a sample that own a computer. We are going to study the following distributions:

Uniform Discrete Distribution e.g., the number of dots facing up if a balanced die is tossed;

Binomial Distribution e.g., the number of cured patients among all the patients who use the drug in a study involving testing the effectiveness of a new drug;

Hypergeometric Distribution e.g., the number of defective items in the sample, when a sample of items selected from a batch of production is tested;

Negative binomial Distribution e.g., the number of cards that you selected from an ordinary deck of playing cards until you see the fourth ace;

Poisson Distribution e.g., the number of white cells from a fixed amount of an individual's blood sample.

5.1 Uniform Discrete Distribution

Uniform Discrete Distribution

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

$$\mathsf{P}(X=x)=\frac{1}{N}, \quad x=1,2,\ldots,N$$

The mean and the variance of X are given by, respectively,

$$\mu=\frac{N+1}{2} \qquad \text{and} \qquad \sigma^2=\frac{(N+1)(N-1)}{12}$$

EXAMPLE 5.1. Derive the mean and the variance of the uniform discrete distribution.

5.2 Binomial Distributions

The Bernoulli Process

- There are n trials/observations. n is a *fixed* number.
- Each trial/observation results in only *one of just two possible outcomes*, which we call "success" or "failure", for convenience.
- The probability of "success" on each trial or observation is a *constant*, p.
- The trials/observations are *independent*.

Think of tossing a balanced coin 10 times,

- There are 10 observations.
- Each toss falls into one of just two categories – either "Head" or "Tail".
- The probability of a head is $p=0.5$ on each toss for a balanced coin.
- Successive tosses are independent.

Denote X be the number of heads. Then $X\sim\mathrm{B}(10,\ 0.5)$.

EXAMPLE 5.2. Suppose that the probability that a randomly selected car need repair in a one-year period is 0.28. We randomly select 3 cars.

(a) Let RNR denote that the first and third selected cars need repair, and the second doesn't. Calculate $\mathrm{P}(RRR)$, $\mathrm{P}(RNN)$, $\mathrm{P}(NRN)$, $\mathrm{P}(NNR)$, $\mathrm{P}(RRN)$, $\mathrm{P}(RNR)$, $\mathrm{P}(NRR)$ and $\mathrm{P}(NNN)$.

$$\begin{aligned}
\mathrm{P}(RRR) &= (0.28)(0.28)(0.28) \\
&= (0.28)^3 \\
\mathrm{P}(RNN) &= (0.28)(1-0.28)(1-0.28) \\
&= (0.28)(1-0.28)^2 \\
\mathrm{P}(NRN) &= (1-0.28)(0.28)(1-0.28) \\
&= (0.28)(1-0.28)^2 \\
\mathrm{P}(NNR) &= (1-0.28)(1-0.28)(0.28) \\
&= (0.28)(1-0.28)^2 \\
\mathrm{P}(RRN) &= (0.28)(0.28)(1-0.28) \\
&= (0.28)^2(1-0.28) \\
\mathrm{P}(RNR) &= (0.28)(1-0.28)(0.28) \\
&= (0.28)^2(1-0.28) \\
\mathrm{P}(NRR) &= (1-0.28)(0.28)(0.28) \\
&= (0.28)^2(1-0.28) \\
\mathrm{P}(NNN) &= (1-0.28)(1-0.28)(1-0.28) \\
&= (1-0.28)^3
\end{aligned}$$

(b) Let X be the number of cars that require repair. What is the probability distribution of X?

The probabilities are

$$\begin{aligned}
f(0) &= \mathrm{P}(X=0) = \mathrm{P}(NNN) \\
&= (1-0.28)^3 = 0.373248 \\
f(1) &= \mathrm{P}(X=1) = \mathrm{P}(RNN \cup NRN \cup NNR) \\
&= 3(0.28)(1-0.28)^2 = 0.435456 \\
f(2) &= \mathrm{P}(X=2) = \mathrm{P}(RRN \cup RNR \cup NRR) \\
&= 3(0.28)^2(1-0.28) = 0.169344 \\
f(3) &= \mathrm{P}(X=3) = \mathrm{P}(RRR) \\
&= (0.28)^3 = 0.021952
\end{aligned}$$

Thus, the PMF of X is given by

$$f(x) = \binom{3}{x}(0.28)^x(1-0.28)^{3-x},\ x = 0,1,2,3$$

or,

x	0	1	2	3
$f(x)$	0.373248	0.435456	0.169344	0.021952

We say that X follows a binomial distribution with parameters $n = 3$ and $p = 0.28$, i.e. $X \sim \mathrm{B}(3,\ 0.28)$, where $x = 0,1,2,3$.

Binomial Distributions

A Bernoulli trial can result in a success with probability p and a failure with probability $q = 1-p$. Then the probability distribution of the binomial random variable X, the number of successes in n independent trials, is

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1,2,\ldots,n.$$

EXAMPLE 5.3. A pop quiz has 10 independent multiple-choice questions. Each question has 5 possible answers of which only 1 is correct. A student randomly guesses answers to all questions. Let X be the number of correct guesses.

(a) What is the distribution of X?

(b) What is the probability that the student has exactly 6 correct answers?

(c) What is the probability that the student has at most 2 correct answers?

(d) What is the probability that the student has at least 1 correct answer?

Mean and Standard deviation of Binomial distributions

The mean and variance of the binomial distribution $b(x;n,p)$ are

$$\mu = np \quad \text{and} \quad \sigma^2 = npq.$$

EXAMPLE 5.4. Refer to the preceding question. What is the mean of X? What is the standard deviation of X?

EXAMPLE 5.5. A factory employs several thousand workers, of whom 30% are Hispanic. Suppose that 15 members of the union executive committee were chosen from the workers at random. Let X be the number of Hispanics on the committee.

(a) What is the distribution of X?

(b) What is the probability that exactly 3 members of the committee are Hispanic?

(c) What is the probability that 3 or fewer members of the committee are Hispanic?

(d) What are the mean and standard deviation of X?

Example 5.6. According to the *Daily Racing Form*, the probability is about 0.67 that the favorite in a horse race will finish in the money (first, second, or third place). In the next five races, what is the probability that the favorite finishes in the money

(a) exactly twice?

(b) at least four times?

(c) between two and four times, inclusive?

Let X be the number of times the favorite finishes in the money in the next five races. What are the mean and standard deviation of X?

Multinomial Distribution

In a given trail can result in the k outcomes $E_1, E_2, \dots, E_k$ with probabilities $p_1, p_2, \dots, p_k$, then the probability distribution of the random variables $X_1, X_2, \dots, X_k$, representing the number of occurrences for $E_1, E_2, \dots, E_k$ in n independent trials, is

$$f(x_1, x_2, \dots, x_k; p_1, p_2, \dots, p_n, n) = \binom{n}{x_1, x_2, \cdots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

with $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} p_i = 1$.

Example 5.7. According to *USA Today* (March 18, 1997), of 4 million workers in the general workforce, 5.8% tested positive for drugs. Of those testing positive, 22.5% were cocaine users and 54.4% marijuana users. What is the probability that of 10 workers testing positive, 2 are cocaine users, 5 are marijuana users, and 3 are users of other drugs?

5.3 Hypergeometric Distribution

The family of *hypergeometric random variables* is closely related to the family of binomial random variables. The hypergeometric distribution does not require independence and is based on sampling done **without replacement**. The hypergeometric random variables are particularly important in quality control and in the statistical estimation of population proportions.

Example 5.8. A quality assurance engineer of a company that manufactures TV sets inspects finished products in lots of 100. He selected 5 of the 100 TVs at random and inspects them thoroughly. Let X denote the number of defective TVs obtained. If, in fact, 6 of the 100 TVs in the current lot are actually defective,

(a) find the PMF of the random variable X.

(b) what is the probability that exactly one TV is defective?

(c) What is the probability that at least one TV is defective?

In general, we are interested in the probability of selecting x successes from the k items labeled successes and $n-x$ failures from the $N-k$ items labeled failures when a random sample of size n is selected from N items. This is known as a *hypergeometric experiment*, that is, one that possesses the following two properties:

i) A random sample of size n is selected without replacement from N items.

ii) Of the N items, k may be classified as successes and $N-k$ are classified as failures.

The number X of successes of a hypergeometric experiment is called a **hypergeometric random variable**. Accordingly, the probability distribution of the hypergeometric variable is called the *hypergeometric distribution*, and its values are denoted by $\mathrm{h}(x; N, n, k)$, since they depend on the number of successes k in the set N from which we select n items.

Hypergeometric Distribution

The probability distribution of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labeled **success** and $N-k$ labeled **failure**, is

$$\mathrm{h}(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},$$

where, $\max\{0, n-(N-k)\} \le x \le \min\{n, k\}$.

Example 5.9. An Arizona State lottery, called *Lotto*, is played as follows: A player specifies six numbers of her choice from the number of 1–42; these six numbers constitute the player's "tickets" for which the payer pays \$1. In the lottery drawing, six winning numbers are chosen at random without replacement from the numbers 1–42. To win a prize, a *Lotto* ticket must contain three or more of the wining numbers.

(a) Let X be the number of winning numbers on the player's ticket. What is the distribution of X? Determine the PMF of X?

(b) If the player buys one *Lotto* ticket, determine the probability that she wins a prize.

Mean and Standard Deviation of Hypergeometric Distribution

The mean and variance of the hypergeometric distribution $\mathrm{h}(x;N,n,k)$ are

$$\mu = \frac{nk}{N} \quad \text{and} \quad \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N} \cdot \left(1 - \frac{k}{N}\right).$$

EXAMPLE 5.10. Find the mean and standard deviation of the random variable in Example 5.8.

EXAMPLE 5.11. Five cars are selected at random without replacement from an ordinary deck of 52 playing cards.

(a) What is the probability that exactly 3 face cards are obtained?

(b) Identify and provide a formula for the probability distribution of the number of face cards obtained.

(c) How many face cards do you expect to obtain?

Binomial Approximation to the Hypergeometric Distribution

Let k/N play the role of p in the binomial setting ($k = Np, N-k = Nq$). Then

$$\begin{aligned}\lim_{N\to\infty} \mathrm{h}(x;N,n,k) &= \lim_{N\to\infty} \frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}} \\ &= \binom{n}{x} p^x q^{n-x} = \mathrm{b}(x;n,p)\end{aligned}$$

NOTE. The approximation is good when $n/N \le 5\%$.

EXAMPLE 5.12. Refer to Example 5.8. Use the binomial approximation to calculate the probability that exactly one TV is defective. Compare the results.

EXAMPLE 5.13. As reported by Television Bureau of Advertising, Inc., in *Trends in Television,* 84.2% of U.S. households have a VCR. If six U.S. households are randomly selected without replacement, what is the (approximate) probability that the number of household sampled that have a VCR will be exactly four? Why is the probability that you just obtained only approximately correct?

Multivariate Hypergeometric Distribution

If N items can be partitioned into the k cells A_1, A_2, $\ldots A_k$ with a_1, a_2, $\ldots$, a_k elements, respectively, then the probability distribution of the random variables X_1, X_2, $\ldots X_k$, representing the number of elements selected from A_1, A_2, $\ldots A_k$ in a random sample of size n, is

$$f(x_1,x_2,\ldots,x_k;a_1,a_2,\ldots,a_k,N,n) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}}{\binom{N}{n}}$$

where $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} a_i = N$.

EXAMPLE 5.14. From a sack of fruit containing 10 oranges, 4 apples, and 6 bananas, a random sample of 8 pieces of fruit is selected. What is the probability that 4 oranges, 1 apple and 3 bananas are selected?

5.4 Negative Binomial Distribution and Geometric Distribution

We consider the family of *negative binomial random variables*, which includes as a special case the family of geometric random variables.

In the Bernoulli experiment, one of the properties is that the trials will be repeated for n times, where n is fixed and we look for the probability of x successes in n trials. Instead, the **negative binomial experiment** is interested in the probability that the kth success occurs on the xth trial, in which the number of successes is fixed, but the number of trials is not.

EXAMPLE 5.15. Suppose that a coin with probability p of a head is tosses repeatedly. Let X denote the number of tosses until the eighth head is obtained. Determine the PMF of the random variable X.

In general, we consider the **negative binomial experiment** of "the kth success happens on the xth trial", or, "the number x of trials required to produce k successes". This happens if and only if (i) among the first $x-1$ trials, there are exactly $k-1$ successes and $x-k$ failures in some specified order, and (ii) there must be a success on the xth trial. The probabilities of (i) and (ii) are $\binom{x-1}{k-1} p^{k-1} q^{x-k}$ and p, respectively. Since they are independent, we have the following

Negative Binomial Distribution

If repeated independent trials can result in a success with probability p and a failure with probability $q = 1-p$, then the probability distribution of the random variable X, the number of the trial on which the kth success occurs, is

$$b^*(x;k,p) = \binom{x-1}{k-1} p^k q^{x-k}, \qquad x = k, k+1, k+2, \ldots$$

NOTE. The terminology "negative binomial" is used because we can express the PMF above in terms of a binomial coefficient containing a negative term. Specifically, if $X \sim b^*(x;k,p)$, the PMF of X can be written in the alternative form $\binom{-k}{x-k} p^k (p-1)^{x-k}$.

EXAMPLE 5.16. From past experience, it is known that a telemarketer makes a sale with probability 0.2. Assuming that results from one call to the next are independent, determine the probability that the telemarketer makes

(a) the second sale on the fifth call.

(b) the second sale by the fifth call.

(c) the second sale on the fifth call and the fifth sale on the fifteenth call.

EXAMPLE 5.17. According to the *Daily Racing Form*, the probability is about 0.67 that the favorite in a horse race will finish in the money (first, second, or third place). Suppose that you always bet the favorite "across the board," which means that you win something if the favorite finishes in the money. let X denote the number of races that you bet until you win something three times.

(a) Determine and identify the PMF of the random variable X.

(b) Find the probability that the number of races that you bet until you win something three times is exactly four; at least four; at most four.

Mean and Standard Deviation of Negative Binomial Random Variables

If $X \sim b^*(x;k,p)$ then the mean and the variance of X are given by, respectively,

$$\mu = \frac{k}{p} \qquad \text{and} \qquad \sigma^2 = \frac{kq}{p}.$$

EXAMPLE 5.18. Find the mean and the standard deviation of the random variable in the preceding example.

Take $k = 1$ in the negative binomial experiment, we have a geometric random variable where we are interested in the occurrence of the first success on the xth trials.

Geometric Distribution

If repeated independent trials can result in a success with probability p and a failure with probability $q = 1-p$, then the probability distribution of the random variable X, the number of the trial on which the first success occurs, is

$$g(x;p) = pq^{x-1}, \qquad x = 1,2,3,\ldots$$

The mean and the standard deviation of X are given by, respectively,

$$\mu = \frac{1}{p} \qquad \text{and} \qquad \sigma = \sqrt{\frac{1-p}{p}}.$$

EXAMPLE 5.19. Refer to Example 5.17. Let Y denote the number of races that you bet until you win something.

(a) Determine and identify the PMF of the random variable Y.

(b) Find the probability that the number of races that you bet until you win something is exactly three; at least three; at most three.

(c) How many races must you bet to be at least 99% sure of winning something?

Lack-of-memory property

For positive-integer valued random variable, we can express the lack-of-memory property in the form

$$\mathrm{P}(X > m+n | X \geq m) = \mathrm{P}(X > n), \qquad m,n \in \mathscr{N}.$$

Theorem. A positive-integer valued random variable has the lack-of-memory property if and only if it has a geometric distribution.

EXAMPLE 5.20. Let $X \sim g(x;p)$. Determine $\mathrm{P}(X > 9 | X \geq 6)$.

5.5 Poisson Distribution

The Poisson distribution is often used for modelling the behaviour of rare events (with small probabilities). These events involve with the counting processes. Some examples are the number of Emergency Department visits by an infant during the first year of life, the number of arrivals of people in a line, the number of Internet users logging onto a website, and the number of white blood cells found in a cubic centimeter of blood.

The Poisson distribution is a discrete probability distribution that applies to occurrences of some event over a specified interval, such as time, distance, area, volume, etc. Experiments yielding numerical values of a random variable X, the number of outcomes occurring during a given time interval or in a specified region, are called **Poisson experiments**. The number X of outcomes occurring during a Poisson experiment is called a **Poisson random variable**, and its probability distribution is called the **Poisson distribution**.

Poisson Distribution

The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval or specified region denoted by t, is

$$p(x;\lambda t) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}, \qquad x = 0,1,2,\ldots$$

where λ is the average number of outcomes per unit time, distance, area, or volume and $e = 2.71828\ldots$.

NOTE. The mean number of outcomes is computed from $\mu = \lambda t$, where t is the specific "time," "distance," "area," or "volume" of interest. It depends on both λ, the rate of occurrence of outcomes and t, the given time interval or specified region.

Mean and Standard Deviation of Poisson Random Variables

If $X \sim p(x;\lambda t)$ then the mean and the variance of X are given by, respectively,

$$\mu = \lambda t \qquad \text{and} \qquad \sigma^2 = \lambda t.$$

EXAMPLE 5.21. The average number of homes sold by the a Realty company is 2 homes per day. What is the probability that (i) exactly 3 homes, (ii) no homes, and (iii) at least one home will be sold tomorrow?

EXAMPLE 5.22. Suppose the average number of lions seen on a 1-day safari adventure is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari? Let X be the number of lions that tourists will see next week (7 days). What are the mean and the standard deviation of X?

EXAMPLE 5.23. A life insurance salesman sells on the average 3 life insurance policies per week (5 working days).

(a) what is the probability that he will sell at least one policy in a given working day?

(b) what is the probability that he will sell exactly 15 policies in a given months (4 working weeks)?

(c) how many policies would be expected to sell in a year (260 working days)?

Poisson Approximation to the Binomial Distribution

Let X be a binomial random variable with probability distribution $\mathrm{b}(x;n,p)$. When $n \to \infty$, $p \to 0$ and $np \to \mu = \lambda t$ remains constant,

$$\lim_{n\to\infty} \mathrm{b}(x;n,p) = p(x;\mu).$$

NOTE. If p is close to 1, we can still use the Poisson distribution to approximate binomial probabilities by interchanging a "success" and a "failure", thereby changing p to a value close to 0.

EXAMPLE 5.24. The manufacturer of a tricycle for children has received complaints about defective brakes in the product. According to the design of the product and considerable preliminary testing, it had been determined that the probability of the kind of defect in the complaint was 1 in 10,000 (i.e., 0.0001). After a thorough investigation of the complaints, it was determined that during a certain period of time, 200 products were randomly chosen from production and 5 had defective brakes.

(a) Use the binomial distribution to calculate the probability that 5 or more bad defective brakes are found in these 200 randomly selected products. Comment on the 1 in 10,000 claim by the manufacturer.

(b) Repeat part (a) using the Poisson approximation.

Which model to use?

EXAMPLE 5.25. Which probability model would best describe each of the following random variable? Specify the values of the parameters.

(a) At peak periods, 25% of attempted logins to an Internet service provider fails. Login attempts are independent and each has the same probability of failing. Tom logs in repeatedly until he succeeds; X is the number of login attempts until he finally succeeds.

(b) Deal ten cards from a shuffled deck; X is the count of heart cards.

(c) Most calls made at random by sample surveys don't succeed in talking with a live person. Of calls to Winnipeg, only 1/7 succeed. A survey calls 400 randomly selected numbers in Duluth; X is the number that fail to reach a live person.

(d) On a bright October day, Canada geese arrive to foul the pond at an apartment complex at an average rate of 12 geese per hour; X is the number of geese that arrive in the next two hours.