Chapter 3

Discrete Random Variables and Probability Distributions

3.1 Random Variables

Random variable (rv) – is obtained by assigning a numerical value to each outcome in the sample space S of a particular experiment. In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers. Random variables are denoted by uppercase letters such as X and Y.

- A **discrete random variable** is a random variable that takes either a finite number of possible values or at most a countably infinite number of possible values.

- A **continuous random variable** is a random variable that takes an infinite number of possible values that is not countable. For any possible value c, $P(X = c) = 0$.

In this chapter, we examine the basic properties and discuss the most important examples of discrete variables. Chapter 4 focuses on continuous random variables.

Exercise 3.7 For each random variable defined here, describe the set of possible values for the variable, and state whether the variable is discrete.

(a) X = the number of unbroken eggs in a randomly chosen standard egg carton

(b) Y = the number of students on a class list for a particular course who are absent on the first day of classes

(c) U = the number of times a duffer has to swing at a golf ball before hitting it

(d) X = the length of a randomly selected rattlesnake

(e) Y = the pH of a randomly chosen soil sample

(f) X = the tension (psi) at which a randomly selected tennis racket has been strung

3.2 Probability Distributions for Discrete Random Variables

Probability distribution – A table, graph, or mathematical formula that provides the possible values of a random variable X and their corresponding probabilities.

The **probability distribution** or **probability mass function (pmf)** of a discrete variable X is defined for every number x by
$p(x) = P(X = x) = P(\text{all } s \in S; X(s) = x)$,
which is the sum of the probabilities of all sample points in S that are assigned the value x.

- $0 \leq p(x) \leq 1$
- $\sum_x p(x) = 1$

The **cumulative distribution function (cdf)** $F(x)$ of a discrete random variable X is defined for every number x by
$F(x) = P(X \leq x) = \sum\limits_{y:y \leq x} p(y)$.

- $0 \leq F(x) \leq 1$
- $F(a) = 0$, for $a < x_{\min}$, $x_{\min}$ is the smallest possible X value
- $F(b) = 1$, for $b \geq x_{\max}$, $x_{\max}$ is the largest possible X value

Example A fair coin is thrown three times and the sequence of heads (H) and tails (T) is recorded. The sample space is
$S = \{$HHH, HHT, HTT, HTH, TTT, TTH, THH, THT$\}$.

Define random variable X as the total number of heads.
Find probability mass function (pmf) $p(x)$ and construct a table and a **probability histogram** to represent the pmf.
Also, find the cumulative distribution function (cdf) $F(x)$. Use a **step graph** to describe the cdf.

The pmf $p(x)$ can be recovered by calculating the jump value of $F(x)$ at x, i.e., $p(x) = P(X = x) = F(x) - F(x^-)$, where x^- represents the largest possible X value that is strictly less than x.

For any two numbers x_1 and x_2 with $x_1 \leq x_2$, we have $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1^-)$.

Example X is a discrete rv. Complete the following table.

	using pmf $p(x)$	using cdf $F(x)$
$P(X < a)$		
$P(X \leq a)$		
$P(a \leq X \leq b)$		
$P(a < X \leq b)$		
$P(a \leq X < b)$		
$P(a < X < b)$		
$P(X > b)$		
$P(X \geq b)$		

Exercise 3.24 An insurance company offers its policyholders a number of different premium payment options. For a randomly selected policyholder, let X = the number of months between successive payments. The cdf of X is as follows:

$$F(x) = \begin{cases} 0 & x < 1 \\ .30 & 1 \leq x < 3 \\ .40 & 3 \leq x < 4 \\ .45 & 4 \leq x < 6 \\ .60 & 6 \leq x < 12 \\ 1 & 12 \leq x \end{cases}$$

(a) What is the pmf of X?

(b) Using just the cdf, compute $P(3 \leq X \leq 6)$ and $P(4 \leq X)$.

3.3 Expected Values

If X is a discrete random variable with pmf $p(x)$, then the **expected value** or **mean value** of X, denoted by $E(X)$ or μ_X or just μ, is

$$E(X) = \mu_X = \sum_x xP(X = x) = \sum_x xp(x)$$

and for any real-valued function $h(X)$, its expected value is

$$E[h(X)] = \mu_{h(X)} = \sum_x h(x)P(X = x) = \sum_x h(x)p(x)$$

Example A fair coin is thrown three times and the sequence of heads (H) and tails (T) is recorded.

(a) Let random variable X = the total number of heads. Find $E(X)$, which is the expected number of heads you will observe.

(b) If you win \$2 every time obseving a head and have to pay \$1 every time observing a tail. What is the expected profit of this game?

Let X have pmf $p(x)$ with mean μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2 or just σ^2, is defined by

$$\begin{aligned} V(X) = \sigma^2 = E[(X-\mu)^2] &= \sum_x (x-\mu)^2 p(x) \\ = E(X^2) - [E(X)]^2 &= \Big[\sum_x x^2 p(x)\Big] - \mu^2 \end{aligned}$$

The **standard deviation (SD)** of X is $\sigma = \sqrt{\sigma^2}$.

For any real-valued function $h(X)$, the variance of $h(X)$ is

$$V[h(X)] = \sigma^2_{h(X)} = E\{h(x) - E[h(x)]\}^2 = \sum_x \{h(x) - E[h(x)]\}^2 p(x)$$

Example An insurance company will sell you a \$10,000 term life policy for an annual premium of \$300. Based on a period life table from the U.S. government, the probability that you will survive the coming year is 0.999, what is the expected gain and the variance of the gain for the insurance company for the coming year?

Properties of the Expected Value

- If X is a random variable and $Y = aX + b$ for some constants a and b, then $E(Y) = E(aX + b) = aE(X) + b$.

  $\Rightarrow E(aX) = aE(X)$

  $\Rightarrow E(b) = b$

- For any random variables X and Y,
  $E(X + Y) = E(X) + E(Y)$.

  $\Rightarrow E\Big(\sum_i X_i\Big) = \sum_i E(X_i)$

Properties of the Variance

- If X is a random variable and $Y = aX + b$ for some constants a and b, then $V(Y) = V(aX + b) = a^2V(X)$.

  $\Rightarrow V(-X) = V(X)$

  $\Rightarrow V(aX) = a^2V(X)$

  $\Rightarrow V(X + b) = V(X)$

  $\Rightarrow V(b) = 0$

- If random variables X and Y are independent, then
  $V(X + Y) = V(X) + V(Y)$
  $V(X - Y) = V(X) + V(Y)$

Example Given a random variable X with $E(X) = \mu$ and $V(X) = \sigma^2$, let $Y = \frac{X-\mu}{\sigma}$. Find $E(Y)$ and $V(Y)$.

Exercise 3.32 An appliance dealer sells three different models of upright freezers having 13.5, 15.9, and 19.1 cubic feet of storage space, respectively. Let X = the amount of storage space purchased by the next customer to buy a freezer. Suppose that X has pmf

x	$p(x)$
13.5	.2
15.9	.5
19.1	.3

(a) Compute $E(X)$, $E(X^2)$, and $V(X)$.

(b) If the price of a freezer having capacity X cubic feet is $25X - 8.5$, what is the expected price paid by the next customer to buy a freezer?

(c) What is the variance of the price $25X - 8.5$ paid by the next customer?

(d) Suppose that although the rated capacity of a freezer is X, the actual capacity is $h(X) = X - .01X^2$. What is the expected actual capacity of the freezer purchased by the next customer?

3.4 The Binomial Probability Distribution

Bernoulli random variable – a random variable whose only possible values are 0 and 1, with probabilities $1-p$ and p, respectively. Its pmf is thus

$$p(1) = P(X = 1) = p$$
$$p(0) = P(X = 0) = 1 - p = q\ .$$

The mean and variance of the Bernoulli rv are

$$E(X) = p$$
$$V(X) = p(1-p) = pq.$$

Note: Suppose an experiment, which results in a "success" with probability p and a "failure" with probability $1-p$, is performed. The random variable defined to be 1 or 0 according to whether the experiment was a "success" or a "failure" is a Bernoulli random variable.

Suppose for rv X its pmf $p(x)$ depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a **parameter** of the distribution.

Bernoulli random variable deals only with one single experiment resulting in a "success" with probability p and a "failure" with probability $1-p$.

If we repeat it n times, then the total number of successes, X, is a **binomial random variable**, with parameters n and p.
We write $X \sim Bin(n,p)$.
Its pmf is

$$b(x;n,p) = P(X=x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1,\cdots,n \\ 0 & \text{otherwise} \end{cases}$$

Its cdf is

$$B(x;n,p) = P(X \le x) = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y}, \quad x = 0,1,\cdots,n$$

And its mean and variance are

$$E(X) = np$$
$$V(X) = np(1-p) = npq$$

Example It is known that screws produced by a certain company will be defective with probability .01 independently of each other. The company sells the screws in packages of 10 and offers a money-back guarantee that at most 1 of the 10 screws is defective. What proportion of packages sold by this company would need to be refunded?

Exercise 3.54 A particular type of tennis racket comes in a mid size version and an oversize version. Sixty percent of all customers at a certain store want the oversize version.

(a) Among ten randomly selected customers who want this type of racket, what is the probability that at least six want the oversize version?

(b) Among ten randomly selected customers, what is the probability that the number who want the oversize version is within 1 standard deviation of the mean value?

(c) The store currently has seven rackets of each version. What is the probability that all of the next ten customers who want this racket can get the version they want from current stock?

3.5 Hypergeometric and Negative Binomial Distributions

Suppose the population or set to be sampled consists of N individuals, objects, or elements (a finite population).

Each individual can be characterized as a success (S) or a failure (F), and there are M successes in the population.

A sample of n individuals is selected without replacement such that each subset of size n is equally likely to be chosen.

The random variable of interest, X = the number of S's in the sample, is a **hypergeometric random variable** with parameters n, M, and N.
Its pmf is

$$h(x; n, M, N) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

for any integer x satisfying $\max(0, n-N+M) \leq x \leq \min(n, M)$.

And its mean and variance are

$$E(X) = n \cdot \frac{M}{N}$$
$$V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

Exercise 3.68 An electronics store has received a shipment of 20 table radios that have connections for an iPod or iPhone. Twelve of these have two slots (so they can accommodate both devices), and the other eight have a single slot. Suppose that six of the 20 radios are randomly selected to be stored under a shelf where the radios are displayed, and the remaining ones are placed in a storeroom. Let X = the number among the radios stored under the display shelf that have two slots.

(a) What kind of a distribution does X have (name and values of all parameters)?

(b) Compute $P(X = 2)$, $P(X \leq 2)$, and $P(X \geq 2)$.

(c) Calculate the mean value and standard deviation of X.

Suppose that independent Bernoulli trials, each with probability p of being a success, are performed until a total number of r successes occurs, where r is a specified positive integer.

The random variable X = the number of failures that precede the rth success, is called a **negative binomial random variable** with parameters p and r.
We write $X \sim NB(r, p)$.

Its pmf is:

$$nb(x; r, p) = P(X = x) = \binom{x+r-1}{r-1} p^r (1-p)^x, \quad x = 0, 1, 2, \cdots$$

And its mean and variance are

$$E(X) = \frac{r(1-p)}{p}$$

$$V(X) = \frac{r(1-p)}{p^2}$$

Example When a fisherman catches a fish, if it is young with a probability of 0.2, the fisherman returns the fish to the water. On the other hand, an adult fish will be kept. Suppose the fisherman sets a goal of 10 adult fish.

(a) What is the expected number of young fish caught by the fisherman before the 10th adult fish is caught?

(b) What is the expected number of all fish the fisherman needs to catch in order to reach the goal?

Exercise 3.75 Suppose that $p = P(\text{male birth}) = .5$. A couple wishes to have exactly two female children in their family. They will have children until this condition is fulfilled.

(a) What is the probability that the family has x male children?

(b) What is the probability that the family has four children?

(c) What is the probability that the family has at most four children?

(d) How many male children would you expect this family to have? How many children would you expect this family to have?

Geometric rv as a special case of Negative Binomial rv

Suppose we perform the independent Bernoulli trials until one success occurs. The random variable X = the number of failures, is a **geometric random variable** with parameter p.

That is, $X \sim Geometric(p)$ is equivalent to $X \sim NB(1, p)$.

We have the geometric pmfs,

$$g(x; p) = P(X = x) = p(1-p)^x, \qquad x = 0, 1, 2, \cdots$$

And the means and variances are

$$E(X) = \frac{1-p}{p}$$
$$V(X) = \frac{1-p}{p^2}$$

Example When a fisherman catches a fish, if it is young with a probability of 0.2, the fisherman returns the fish to the water. On the other hand, an adult fish will be kept.

(a) What is the expected number of fish caught by the fisherman until the first adult fish is caught?

(b) What is the probability that the fifth fish caught is the first young fish?

(c) If the fisherman catches five fish, what is the probability that there are exactly one young fish?

3.6 The Poisson Probability Distribution

A discrete random variable X is said to have a **Poisson distribution** with parameter $\mu(\mu > 0)$ if the pmf of X is,

$$p(x;\mu) = P(X = x) = \frac{e^{-\mu}\mu^x}{x!}, \quad x = 0, 1, 2, \ldots$$

The mean and variance of the Poisson rv X are

$$E(X) = \mu$$
$$V(X) = \mu$$

Poisson distribution deals with counting the number of times an event occurs in a given interval (time, space, volume, etc.). X = # of occurrences of some event over a given interval.

Poisson distribution can be used to model the number visits to a particular website, the number of pulses of some sort recorded by a counter, the number of accidents in an industrial facility, the number of particles emitted by a radioactive source, or the number of births during a given day.

General assumptions of the Poisson distribution

1. The probability that an event occurs in a given interval of time is proportional to the length of the interval.

2. Events do not happen simultaneously in a sufficient small interval.

3. What happens in one subinterval is independent of what happens in any other non-overlap subinterval.

Exercise 3.86 The number of people arriving for treatment at an emergency room can be modeled by a Poisson process with a rate parameter of five per hour.

(a) What is the probability that exactly four arrivals occur during a particular hour?

(b) What is the probability that at least four people arrive during a particular hour?

(c) How many people do you expect to arrive during a 45-min period?

Poisson Approximation of Binomial

A binomial random variable with parameters (n, p) can be approximated by a Poisson random variable with $\mu = np$, when n is large enough $(n \geq 100)$ and p is small enough so that np approaches a moderate value ($np \leq 10$).

Theoretically, $$b(x; n, p) \xrightarrow[p\to 0]{n\to\infty} p(x; \mu = np).$$

$$X \sim Bin(n, p) \xRightarrow[p\to 0]{n\to\infty} X \sim Poisson(\mu = np)$$

Table 3.2 Comparing the Poisson and Three Binomial Distributions

x	$n = 30, p = .1$	$n = 100, p = .03$	$n = 300, p = .01$	Poisson, $\mu = 3$
0	0.042391	0.047553	0.049041	0.049787
1	0.141304	0.147070	0.148609	0.149361
2	0.227656	0.225153	0.224414	0.224042
3	0.236088	0.227474	0.225170	0.224042
4	0.177066	0.170606	0.168877	0.168031
5	0.102305	0.101308	0.100985	0.100819
6	0.047363	0.049610	0.050153	0.050409
7	0.018043	0.020604	0.021277	0.021604
8	0.005764	0.007408	0.007871	0.008102
9	0.001565	0.002342	0.002580	0.002701
10	0.000365	0.000659	0.000758	0.000810

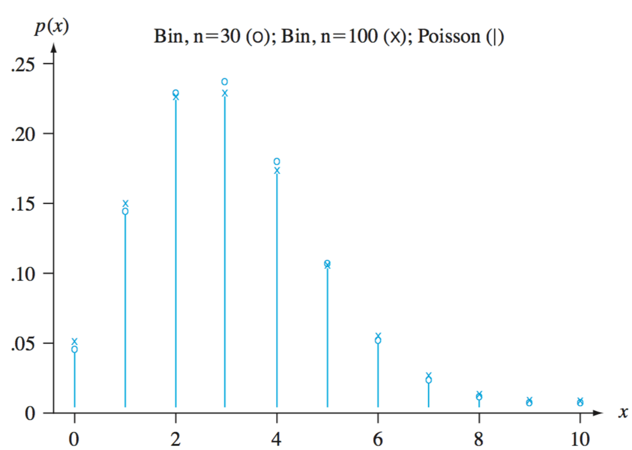

Figure 3.8 Comparing a Poisson and two binomial distributions

Exercise 3.84 Suppose that only .10% of all computers of a certain type experience CPU failure during the warranty period. Consider a sample of 10,000 computers.

(a) What are the expected value and standard deviation of the number of computers in the sample that have the defect?

(b) What is the (approximate) probability that more than 10 sampled computers have the defect?

(c) What is the (approximate) probability that no sampled computers have the defect?