CHAPTER 3: Random Variables and Probability Distributions

Concept of a Random Variable: 3.1

- The outcome of a random experiment need not be a number.
- However, we are usually interested not in the outcome itself, but rather in some measurement of the outcome.

Example: Consider the experiment in which batteries coming off an assembly line were examined until a good one (S) was obtained.

$$S = \{\text{S, FS, FFS}, \ldots\}.$$

We may be interested in the number of batteries examined before the experiment terminates.

A random variable is a function that associate a real number with each element in the sample space.

Example: Tossing two coins

$$S = \{\text{HH, TT, HT, TH}\}$$

Let $X = \#$ of heads observed.

Example: A group of 4 components is known to contain 2 defectives. An inspector tests the components one at the time until the 2 defectives are located. Let X denote the number of the test on which the second defective is found.

Two types of random variables

- A **discrete** random variable is a random variable whose possible values either constitute a finite set or else can be listed in an infinite sequence.
- A random variable is **continuous** if its set of possible values consists of an entire interval on the number line.

  Many random variables, such as weight of an item, length of life of a motor etc., can assume any value in certain intervals.

Probability mass function of a discrete random variable X is defined by

$$f(x) = P(X = x)$$

Example: tossing two coins

$X = \#$ of heads.

$f(0) = P(X = 0) = P(\text{TT}) = 1/4$
$f(1) = P(X = 1) = P(\text{HT, TH}) = 1/2$
$f(2) = P(X = 2) = P(\text{HH}) = 1/4$

Example: An information source produces symbols at random from a five-letter alphabet:

$$S = \{a, b, c, d, e\}.$$

The probabilities of the symbols are

$$p(a) = \frac{1}{2}, \quad p(b) = \frac{1}{4}, \quad p(c) = \frac{1}{8}, \quad p(d) = p(e) = \frac{1}{16}.$$

A data compression system encodes the letters into binary strings as follows:

a	1
b	01
c	001
d	0001
e	0000

Let the random variable Y be equal to the length of the binary string output by the system.

$f(1) = P(Y = 1) =$
$f(2) = p(Y = 2) =$
$f(3) = p(Y = 3) =$
$f(4) = p(Y = 4) =$

$f(x) = P(X = x)$ satisfies the following conditions:
1. $f(x) \geq 0$
2. $\sum f(x) = 1$

Example: A box contains 5 balls numbered $1, 2, 3, 4,$ and 5. Three balls are drawn at random and without replacement from the box. If X is the median of the numbers on the 3 chosen balls, then what is the probability function for X, where nonzero?

Solution

Example: Determine c so that the function $f(x)$ can serve as the probability mass function of a random variable X:

$$f(x) = cx \quad \text{for} \quad x = 1, 2, 3, 4, 5$$

Solution:

<u>The cumulative distribution function:</u> $F(x)$ of a discrete random variable X with probability mass function $f(x)$ is defined for every number x by

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

Example: Assume that

$$f(2) = p(X = 2) = 1/6 \quad f(3) = p(X = 3) = 1/3$$
$$f(4) = p(X = 4) = 1/2$$

Then
$F(2) =$
$F(3) =$
$F(4) =$

$F(x) =$

Example: A mail order computer business has six telephone lines. Let X denote the number of lines in use at a specified time. Suppose that the probability mass function of X is given below

x	0	1	2	3	4	5	6
$p(x)$	0.10	0.15	0.20	0.25	0.20	0.06	0.04

a. Find the cumulative distribution function
b. Find the probability that
{at most 3 lines are in use}.
c. Find the probability that
{at least 4 lines are in use}.

Example:
If X has the cumulative distribution function:

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1/3 & \text{if } 1 \leq x < 4 \\ 1/2 & \text{if } 4 \leq x < 6 \\ 5/6 & \text{if } 6 \leq x < 10 \\ 1 & \text{if } x \geq 10, \end{cases}$$

find the probability mass function.

Solution:

<u>CONTINUOUS PROBABILITY DISTRIBUTION: 3.3</u>

A density curve is a curve that

- is always on or above the horizontal axis, and
- has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.

<u>Definition 3.6 from the book:</u>
The function $f(x)$ is a **probability density function** for the continuous random variable X, defined over the set of real numbers R, if

1. $f(x) \geq 0$, for all $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X < b) = \int_a^b f(x)dx$

Ex. 9 on p. 73: The proportion of people who respond to a certain mail-order solicitation is a continuous random variable X that has density function

$$f(x) = \begin{cases} \frac{2(x+2)}{5} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

(a)Show that $P(0 < X < 1) = 1$

(b) Find the probability that more than 1/4 but fewer than 1/2 of the people contacted will respond to this type of solicitation.

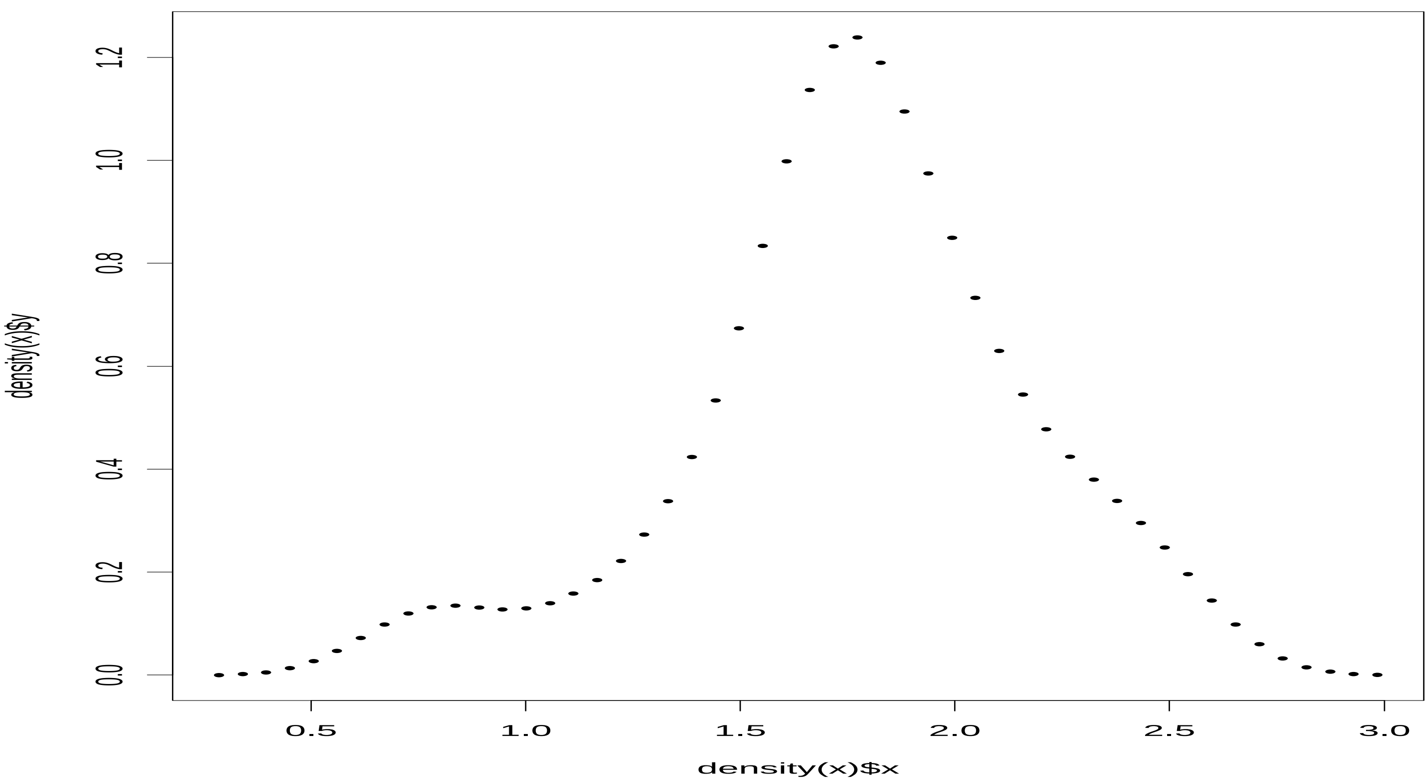

Figure 1: Density histogram for the data in Ex 9 on p. 21

Example The mileage (in thousands of miles) that car owners get with a certain kind of tire is a random variable having the density function

$$f(x) = \begin{cases} \frac{1}{20}e^{-x/20} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0, \end{cases}$$

Find the probabilities that one of these tires will last
(a) at most $10,000$ miles;

(b) anywhere from $16,000$ to $24,000$ miles;

(c) at least $30,000$ miles.

Solution:

Example: The pdf of the samples of speech waveforms is found to decay exponentially at a rate α, so the following pdf is proposed:

$$f(x) = ce^{-\alpha|x|} \qquad -\infty < x < \infty$$

Find the constant C, and then find the probability $P[|X| < \nu]$.

Solution

Example Suppose the reaction temperature X in a certain process has a uniform distribution

$$f(x) = \begin{cases} \frac{1}{10}, & \text{if } -5 \leq x \leq 5, \\ 0 & \text{otherwise,} \end{cases}$$

Compute $P(X < 0)$, $P(-2 < X < 3)$, and $P(-7 < X < 1)$

Solution:

Example: The cumulative distribution function of checkout time duration X is

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{x^2}{4} & \text{if } 0 \leq x < 2, \\ 1 & \text{if } x \geq 2 \end{cases}$$

(a) Use this to compute $P(X \leq 1)$ and
$P(0.5 \leq X \leq 1)$
(b) Find the density function of X

Solution

Review:
1. A sale engineer for a manufacturer of high-speed grinding equipment has just returned from visiting five possible clients. She believes that the following table describes the distribution of the number of sales she will make:

x(#of sales)	0	1	2	3	4	5
$P(X = x)$	0.05	0.3	0.3	0.2	0.1	0.05

a. Compute and plot the cumulative function $F(x) = P(X \leq x)$.
b. Find the probability that the sales engineer will make more than 3 sales.
c. Find the probability that she makes at least 2 sales.

2. The output of a chemical process is continually monitored to ensure that the concentration remains within acceptable limits. Whenever the concentration drifts outside the limits, the process is shut down and re-calibrated. Let X be the number of times in a given week that the process is calibrated. Let X be the of times in a given week that the process is calibrated. The following is the cumulative function $F(X)$,

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.17 & \text{if } 0 \leq x < 1 \\ 0.53 & \text{if } 1 \leq x < 2 \\ 0.84 & \text{if } 2 \leq x < 3 \\ 0.97 & \text{if } 3 \leq x < 4 \\ 1 & \text{if } x \geq 4 \end{cases}$$

a. What is the probability that the process is re-calibrated fewer than two times during a week.
b. What is the probability that the process is re-calibrated more than three times during a week?
c. What is the probability that the process is re-calibrated exactly once during a week?
d. What is the probability that the process is not re-calibrated at all during a week?
e. What is the most probable number of re-calibrations to occur during a week?

3. Specifications call for the thickness of aluminum sheets that are to be made into cans to be between 8 and 11 thousands of an inch. Let X be the thickness of an aluminum sheet. Assume the probability density of X is given by

$$f(x) = \begin{cases} \frac{x}{54} & 6 < x < 12 \\ 0 & \text{otherwise} \end{cases}$$

a. What proportion of sheets will meet the specification?
b. Find the cumulative distribution function of the thickness.
c. Find the median thickness.
d. Find the 10th percentile of the thickness.
e. A particular sheet is 10 thousandths of an inch thick. What proportion of sheets are ticker than this?

4. A college professor never finishes his lecture before the bell rings to end the period and always finishes his lectures within 2 min after the bell rings. Let X =the time that elapses between the bell and the end of the lecture and suppose that the cumulative distribution is

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x^3}{8} & 0 \leq x \leq 2 \\ 1 & x \geq 2 \end{cases}$$

a. What is the probability that the lecture ends within 1 min of the bell ringing?
b. What is the probability that the lecture continues beyond the bell for between 60 and 90 sec?
c. What is the probability that the lecture continues for at least 90 sec beyond the bell?

Questions from Old Exam

1. Consider a random variable X with the following probability mass function

x	-3	0	1	2
$f(x)$	.2	.3	.4	c

Find $P(X > 0.2)$. (hint: you need to find c first).

(a) .4
(b) .8
(c) .7
(d) .5

2. An insurance company offers its policyholder a number of different premium payment options. For a randomly selected policyholder, Let X = the number of months between successive payments. The cumulative distribution function of X is as follows:

$$F(x) = \begin{cases} 0 & x < 1 \\ 0.30 & 1 \le x < 3 \\ 0.40 & 3 \le x < 4 \\ 0.45 & 4 \le x < 6 \\ 0.60 & 6 \le x < 12 \\ 1 & x \ge 12. \end{cases}$$

Compute $P(3 \le X \le 6)$

(a) 0.3
(b) 0.2
(c) none of the above
(d) $\int_3^6 F(x)dx$

3. The probability density function of the time to failure of an electronic component in a copier (in hours) is

$$f(x) = \begin{cases} \frac{1}{3000} e^{-x/3000} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Determine the probability that a component fails in the interval from 1000 to 2000 hours.

(a) 0.2031
(b) $\frac{1}{3000}(e^{-1/3} - e^{-2/3})$
(c) 0.2835
(d) none of the above

4. Suppose the cumulative distribution function of the length (in millimeters) of computer cables is

$$F(x) = \begin{cases} 0 & \text{if } x \le 1200 \\ 0.1x - 120 & \text{if } 1200 < x \le 1210 \\ 1 & \text{if } x > 1210. \end{cases}$$

Which of the following statement is true:
(a) $P(1000 < X < 1208) = 0.8$
(b) $F(x)$ is not a cumulative distribution function since $F(1000) = -20$
(c) $P(1000 < X < 1208) \neq P(X < 1208)$

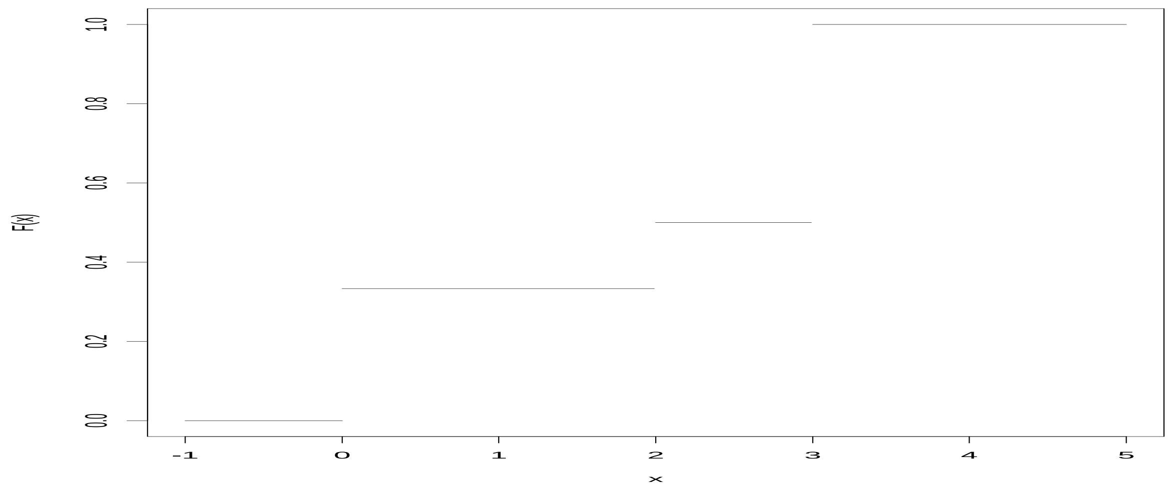

Figure 2: $F(x)$

(d) $P(X = 1208) = 0.8$

5. Let X be a random variable with the following p.m.f.

x	0	3	4.2
$f(x)$	0.4	0.3	0.3

(a) Find $P(X = 3|X \geq 1)$
(b) Find the Cumulative distribution function $F(x)$ for all $-\infty < x < \infty$
(c) Find $P(X < 4.2)$.

6. The cumulative distribution function of X is as follows:

$$F(x) = \begin{cases} 0 & x < -3 \\ 1/7 & -3 \leq x < 3 \\ 6/7 & 3 \leq x < 7 \\ 1 & x \geq 7 \end{cases}$$

(a) Find $P(X \leq 1)$
(b) Find the p.m.f of X

7. For the cumulative distribution function shown in Figure 2, what type of random variable is X? (continuous or discrete). **Explain**

Joint Probability Distributions

Only the discrete case

For a given experiment, we are often interested not only in probability distribution functions of individual random variables but also in the relationship between two or more random variables.

Examples:

- In an experiment into possible causes of cancer, we might be interested in the relationship between the average number of cigarettes smoked daily and the age at which an individual contracts cancer.
- An engineer might be interested in the relationship between the shear strength and the diameter of a spot weld in a fabricated sheet steel specimen.

Let X and Y be two discrete random variables,

$$f(x, y) = P(X = x, Y = y)$$

Example: Suppose that 2 batteries are randomly chosen from a group of 3 new, 4 used but still working, and 5 defective batteries. If we let X and Y denote, respectively, the number of new and used but still working batteries that are chosen, then the joint mass function of X and Y, $f(i, j) = P(X = i, Y = j)$ is given by

$$f(0,0) = \frac{\binom{5}{2}}{\binom{12}{2}} = \frac{10}{66} \quad f(0,1) = \frac{\binom{4}{1}\binom{5}{1}}{\binom{12}{2}} = \frac{20}{66}$$

$$f(0,2) = \frac{\binom{4}{2}}{\binom{12}{2}} = \frac{6}{66} \quad f(1,0) = \frac{\binom{3}{1}\binom{5}{1}}{\binom{12}{2}} = \frac{15}{66}$$

$$f(1,1) = \frac{\binom{3}{1}\binom{4}{1}}{\binom{12}{2}} = \frac{12}{66} \quad f(2,0) = \frac{\binom{3}{2}}{\binom{12}{2}} = \frac{3}{66}$$

			y		Row
	$f(x, y)$	0	1	2	totals
	|0	10/66	20/66	6/66	36/66
x	|1	15/66	12/66		27/66
	|2	3/66			3/66
Column totals		28/66	32/66	6/66	1

Definition 3.8: The function $f(x, y)$ is a **joint probability distribution** or **probability mass function** of the discrete random variables X and Y if

1. $f(x, y) \geq 0$ for all (x, y)
2.
$$\sum_x \sum_y f(x, y) = 1$$
3. $P(X = x, Y = y) = f(x, y)$.

For any region A in the xy plane,

$$P[(X, Y) \in A] = \sum_A \sum f(x, y)$$

Example

In a randomly chosen lot of 1000 bolts, let X be the number that fail to meet a length specification, and let Y be the number that fail to meet a diameter specification. Assume that the joint probability mass function of X and Y is given in the following table.

			y		Row
	$f(x,y)$	0	1	2	totals
	0	0.40	0.12	0.08	0.6
x	1	0.15	0.08	0.03	0.26
	2	0.1	0.03	0.01	0.14
Column totals		0.65	0.23	0.12	1

a. Find $P(X = 0 \ \& \ Y = 2)$
b. Find $P(X > 0 \ \& \ Y le 1)$
c. Find $P(X \leq 1)$
d. Find $P(Y > 0)$
e. Find the probability that all bolts in the lot meet the length specification.
f. Find the probability that all bolts in the lot meet the diameter specification.
g. Find the probability that all bolts in the lot meet both specification.

Ex. 2 on p. 84: If the joint probability distribution function of X and Y is given by

$$f(x,y) = \frac{(x+y)}{30}, \quad \text{for} \quad x = 0,1,2,3 \quad y = 0,1,2.$$

Find
(a)$P(X \leq 2, Y = 1)$; (b) $P(X > 2, Y \leq 1)$

(c) $P(X > Y)$; (d) $P(X + Y = 4)$

Solution:

Definition 3.10: The **marginal distribution** of X alone and of Y alone are

$$g(x) = \sum_y f(x,y) \text{ and } h(y) = \sum_x f(x,y)$$

Example

		x			
$f(x,y)$		0	1	2	3
	0	0	$\frac{1}{30}$	$\frac{2}{30}$	$\frac{3}{30}$
y	1	$\frac{1}{30}$	$\frac{2}{30}$	$\frac{3}{30}$	$\frac{4}{30}$
	2	$\frac{2}{30}$	$\frac{3}{30}$	$\frac{4}{30}$	$\frac{5}{30}$

$P(X=0) =$
$P(X=1) =$
$P(X+2) =$

$P(Y=0) =.$
Thus

x	0	1	2	3
$g(x)$				

y	0	1	2
$h(y)$			

Definition 3.11: The conditional distribution of the random variable Y, given that $X = x$, is

$$f(y|x) = \frac{f(x,y)}{g(x)}, \quad g(x) > 0$$

Similarly, the conditional distribution of the random variable X, given that $Y = y$, is

$$f(x|y) = \frac{f(x,y)}{h(y)}, \quad h(y) > 0$$

Statistical Independence:

Definition: X and Y are said to be **statistically independent** if and only if $f(x|y) = g(x)$.

Result: X and Y are **statistically independent** if and only if $f(x,y) = g(x)h(y)$ for all (x,y).

proof:

$$f(x|y) = \frac{f(x,y)}{h(y)} = g(x),$$

implies that
$f(x,y) = g(x)h(y)$ for all (x,y).

Example: An investor owns two assets. He is interested in the value of his investments in one year. The value of the first asset in one year is a random variable X, and the value of the second asset in one year is a random variable Y. The joint probability mass function of X and Y is given in the following table:

		x		
$f(x,y)$		90	100	110
y	0	0.05	0.27	0.18
	10	0.15	0.33	0.02

(a) Are X and Y independent?

(b) Find $P(X \geq 100)$

(c) Find $P(Y = 0|X = 100)$

Example: A computer program can make calls to two subroutines, A and B. In a randomly chosen run, let X be the number of calls to subroutine A and let Y denote the number of calls to subroutine B. The joint probability mass function of X and Y is given in the following table.

		y		
$f(x,y)$		1	2	3
x	1	0.15	0.10	0.10
	2	0.10	0.20	0.15
	3	0.05	0.05	0.10

a. Find the marginal probability mass function of X.
b. Find the marginal probability mass function of Y.
c. Are X and Y independent? Explain.
d. Find the probability that the number of calls to the two subroutines in 4.
e. Find $P(Y = 1|X = 2)$.
f. Assume that each execution of subroutine A takes 100 ms, and that each execution of subroutine B takes 200 ms. Express the number of milliseconds used in all the calls to the two subroutines in terms of X and Y.