LECTURE NOTES

ON

PROBABILITY AND STATISTICS

2018 – 2019

II B. Tech I Semester

(R-17)

Dr.B.ARUNA KUMARI, **Associate Professor**

CHADALAWADA RAMANAMMA ENGINEERING COLLEGE

(AUTONOMOUS)

Chadalawada Nagar, Renigunta Road, Tirupati – 517 506

Department of Freshman Engineering

UNIT-3

PROBABILITY

INTRODUCTION:

Probability theory was originated from gambling theory. A large number of problems exist even today which are based on the game of chance, such as coin tossing, dice throwing and playing cards.

The probability is defined in two different ways,

- Mathematical (or a priori) definition
- Statistical (or empirical) definition

SOME IMPORTANT TERMS &CONCEPTS:

- RANDOM EXPERIMENTS:
  Experiments of any type where the outcome cannot be predicted are called random experiments.
- SAMPLE SPACE:
  A set of all possible outcomes from an experiment is called a sample space.
  Eg: Consider a random experiment E of throwing 2 coins at a time. The possible outcomes are HH, TT, HT, TH.
  These 4 outcomes constitute a sample space denoted by, S ={ HH, TT, HT, TH}.
- TRAIL & EVENT:
  Consider an experiment of throwing a coin. When tossing a coin, we may get a head(H) or tail(T). Here tossing of a coin is a trail and getting a hand or tail is an event.
  In otherwords, "Every non-empty subset of A of the sample space S is called an event".
- NULL EVENT:
  An event having no sample point is called a null event and is denoted by $\emptyset$.
- EXHAUSTIVE EVENTS:
  The total number of possible outcomes in any trail is known as exhaustive events.
  Eg: In throwing a die the possible outcomes are getting 1 or 2 or 3 or 4 or 5 or 6. Hence we have 6 exhaustive events in throwing a die.
- MUTUALLY EXCLUSIVE EVENTS:
  Two events are said to be mutually exclusive when the occurrence of one affects the occurrence of the other. In otherwords, if A & B are mutually exclusive events and if A happens then B will not happen and viceversa.
  Eg: In tossing a coin the events head or tail are mutually exclusive, since both tail & head cannot appear in the same time.

- EQUALLY LIKELY EVENTS:

  Two events are said to be equally likely if one of them cannot be expected in the preference to the other.

  Eg: In throwing a coin, the events head & tail have equal chances of occurrence.
- INDEPENDENT & DEPENDENT EVENTS:

  Two events are said to be independent when the actual happening of one doesnot influence in any way the happening of the other. Events which are not independent are called dependent events.

  Eg: If we draw a card in a pack of well shuffled cards and again draw a card from the rest of pack of cards (containing 51 cards), then the second draw is dependent on the first. But if on the other hand, we draw a second card from the pack by replacing the first card drawn, the second draw is known as independent of the first.
- FAVOURABLE EVENTS:

  Mathematical or classical or a priori definition of probability,

  Probability (of happening an event E) = $\frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$

  $= \frac{m}{n}$

  Where m = Number of favourable cases

  n = Total number of exhaustive cases.

PROBLEMS:

1. In tossing a coin, what is the prob. of getting a head.

Sol: Total no. of events = {H, T}= 2

Favourable event = {H}= 1

Probability $= \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$

$= \frac{1}{2}$

2. In throwing a die, the prob. of getting 2.

Sol: Total no. of events = {1,2,3,4,5,6}= 6

Favourable event = {2}= 1

Probability $= \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$

$= \frac{1}{6}$

3. Find the prob. of throwing 7 with two dice.

Sol: Total no. of possible ways of throwing a dice twice = 36 ways

Number of ways of getting 7 is, (1,6), (2,5), (3,4), (4,3), (5,2), (6,1) = 6

$$\text{Probability} = \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$$
$$= \frac{6}{36}$$
$$= \frac{1}{6}$$

4. A bag contains 6 red & 7 black balls. Find the prob. of drawing a red ball.

Sol: Total no. of possible ways of getting 1 ball = 6 + 7

Number of ways of getting 1 red ball = 6

$$\text{Probability} = \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$$
$$= \frac{6}{13}$$

5. Find the prob. of a card drawn at random from an ordinary pack, is a diamond.

Sol: Total no. of possible ways of getting 1 card = 52

Number of ways of getting 1 diamond card is 6

$$\text{Probability} = \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$$
$$= \frac{13}{52}$$
$$= \frac{1}{4}$$

6. From a pack of 52 cards, 1 card is drawn at random. Find the prob. of getting a queen.

Sol: A queen may be chosen in 4 ways.

Total no. of ways of selecting 1 card = 52

$$\text{Probability} = \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$$
$$= \frac{4}{52} = \frac{1}{13}$$

7. Find the prob. of throwing: (a) 4, (b) an odd number, (c) an even number with an ordinary die (six faced).

Sol: a) When throwing a die there is only one way of getting 4.

$$\text{Probability} = \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$$
$$= \frac{1}{6}$$

b) Number of ways of falling an odd number is 1, 3, 5 = 3

$$\text{Probability} = \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases} = \frac{3}{6} = \frac{1}{2}$$

c) Number of ways of falling an even number is 2, 4, 6 = 3

Probability $= \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases} = \frac{3}{6} = \frac{1}{2}$

8. From a group of 3 Indians, 4 Pakistanis, and 5 Americans, a sub-committee of four people is selected by lots. Find the probability that the sub-committee will consist of

   i) 2 Indians and 2 Pakistanis.

   ii) 1 Indians, 1 Pakistanis and 2 Americans.

   iii) 4 Americans.

Sol: Total no. of people = 3 + 4 + 5 = 12

∴ 4 people can be chosen from 12 people = $12C_4$ ways

$= \frac{12 \times 11 \times 10 \times 9}{1 \times 2 \times 3 \times 4} = 495$ ways

i) 2 Indians can be chosen from 3 Indians = $3C_2$ ways

2 Pakistanis can be chosen from 4 Pakistanis = $4C_2$ ways

∴ No. of favourable cases = $3C_2 \times 4C_2$

∴ Prob. $= \frac{3C_2 \times 4C_2}{495} = \frac{2}{55}$

ii) 1 Indian can be chosen from 3 Indians = $3C_1$ ways

1 Pakistani can be chosen from 4 Pakistanis = $4C_1$ ways

2 Americans can be chosen from 5 Americans = $5C_2$ ways

Favourable events = $3C_1 \times 4C_1 \times 5C_2$

∴ Prob. $= \frac{3C_1 \times 4C_2 \times 5C_2}{495} = \frac{8}{33}$

iii) 4 Americans can be chosen from 5 Americans = $5C_4$ ways

∴ Prob. $= \frac{5C_4}{495} = \frac{1}{99}$

9. A bag contains 7 white, 6 red & 5 black balls. Two balls are drawn at random. Find the prob. that they both will be white.

Sol: Total no. of balls = 7 + 6 + 5

= 18

From there 18 balls, 2 balls can be drawn in $18C_2$ ways

i.e) $\frac{18 \times 17}{1 \times 2}$ = 153

2 white balls can be drawn from 7 white balls = $7C_2$ ways

= 21

∴ Favourable cases = 21

P(drawing 2 white balls) = $\frac{21}{153} = \frac{7}{51}$

10. A bag contains 10 white, 6 red, 4 black & 7 blue balls. 5 balls are drawn at random. What is the prob. that 2 of them are red and one is black?

Sol: Total no. of balls = 10 + 6 + 4 + 7 =27

5 balls can be drawn from these 27 balls = $27C_5$ ways

$= \frac{27 \times 26 \times 25 \times 24 \times 23}{1 \times 2 \times 3 \times 4 \times 5}$

= 80730 ways

Total no. of exhaustive events = 80730

2 red balls can be drawn from 6 red balls = $6C_2$ ways

$= \frac{6 \times 5}{1 \times 2}$ = 15 ways

1 black balls can be drawn from 4 black balls = $4C_1$ ways

= 4

∴ No. of favourable cases = 15 × 4 = 60

Probability $= \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$

$= \frac{60}{80730} = \frac{6}{8073}$

11. What is the prob. of having a king and a queen, when 2 cards are drawn from a pack of 52 cards?

Sol: 2 cards can be drawn from a pack of 52 cards = $52C_2$ ways

$= \frac{52 \times 51}{1 \times 2}$ = 1326 ways

1 queen card can be drawn from 4 queen cards = $4C_1$ ways

1 king card can be drawn from 4 king cards = $4C_1$ ways

Favourable cases = 4 × 4 = 16 ways

P(drawing 1 queen & 1 king card) = $\frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$

$$= \frac{16}{1326} = \frac{8}{663}$$

12. What is the prob. that out of 6 cards taken from a full pack, 3 will be black and 3 will be red?

Sol: A full pack contains 52cards. Out of 52 cards, 26 cards are red & 26 black cards .

6 cards can be chosen from 52 cards = $52C_6$ ways

3 black cards can be chosen from 26 black cards = $26C_3$ ways

3 red cards can be chosen from 26 red cards = $26C_3$ ways

Favourable cases = $26C_3 \times 26C_3$

Probability = $\frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$

$$= \frac{26C_3 \times 26C_3}{52C_6}$$

13. Find the prob. that a hand at bridge will consist of 3 spades, 5 hearts, 2 diamonds & 3 clubs?

Sol: Total no. of balls = 3 + 5 + 2 + 3 = 13

From 52 cards, 13 cards are chosen in $52C_{13}$ ways

In a pack of 52 cards, there are 13 cards of each type.

3 spades can be chosen from 13 spades = $13C_3$ ways

5 hearts can be chosen from 13 hearts = $13C_5$ ways

2 diamonds can be chosen from 13 diamonds = $13C_2$ ways

3 clubs can be chosen from 13 clubs = $13C_3$ ways

Hence the total no. of favourable cases are = $13C_3 \times 13C_5 \times 13C_2 \times 13C_3$

Probability $= \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases}$

$= \frac{13C_3 \times 13C_5 \times 13C_2 \times 13C_3}{52C_{13}}$

<u>OPERATIONS ON SETS:</u>

If A & B are any two sets, then

i) UNION OF TWO SETS

$A \cup B = \{x: x \in A\ (or)\ x \in B\}$

In general, $A_1 \cup A_2 \cup \ldots\ldots \cup A_n = \{x: x \in A_1\ or\ x \in A_2\ or \ldots\ldots\ldots or\ x \in A_n\}$

i.e) $\bigcup_{i=1}^{n} A_i = \{x: x \in A_i,$ for atleast one i$\}$

ii) INTERSECTION OF TWO SETS

$A \cap B = \{x: x \in A\ \&\ x \in B\}$

In general, $A_1 \cap A_2 \cap \ldots\ldots \cap A_n = \{x: x \in A_1\ and\ x \in A_2\ and \ldots\ldots\ldots and\ x \in A_n\}$

i.e) $\bigcap_{i=1}^{n} A_i = \{x: x \in A_i,$ for all i = 1,2,3…..n$\}$

iii) COMPLEMENT OF A SET

$A'\ or\ \bar{A} = \{x: x \notin A\}$

iv) DIFFERENCE OF TWO SETS

A – B = $\{x: x \in A\ but\ x \notin B\}$

COMMUTATIVE LAW:

$A \cup B = B \cup A$ & $A \cap B = B \cap A$

ASSOCIATIVE LAW:

$(A \cup B) \cup C = A \cup (B \cup C)$ & $(A \cap B) \cap C = A \cap (B \cap C)$

DISTRIBUTIVE LAW:

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

COMPLEMENTARY LAW:

$$A \cup A' = S \ \& \ A \cap A' = \emptyset$$

AXIOMATIC APPROACH TO PROBABILITY:

It is a rule which associates to each event a real number P (A) which satisfies the following three axioms.

AXIOM I : For any event A, P (A) ≥ 0.

AXIOM II : P (S) =1

AXIOM III: If $A_1, A_2, \ldots, A_n$ are finite number of disjoint event of S, then

$$P(A_1, A_2, \ldots, A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

$$= \sum P(A_i)$$

THEOREMS ON PROBABILITY:

THEOREM 1: Probability of an impossible event is zero. i.e) $P(\emptyset) = 0$

THEOREM 2: Probability of the complementary event $\bar{A}$ of A is given by, $P(\bar{A}) = 1 - P(A)$.

THEOREM 3: For any two events A & B, $P(\bar{A} \cap B) = P(B) - P(A \cap B)$.

THEOREM 4: If A and B are two events such that $A \subset B$, then $P(B \cap \bar{A}) = P(B) - P(A)$.

THEOREM 5: If $B \subset A$, then $P(A) \geq P(B)$.

THEOREM 6: If $A \cap B = \emptyset$, then $P(A) \leq P(\bar{B})$.

LAW OF ADDITION OF PROBABILITIES:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, where A & B are any two events and are not disjoint.

PROBLEMS:

1. If from a pack of cards a single card is drawn. What is the prob. that it is either a spade or a king?

Sol: $P(A) = P(\text{a spade card}) = \frac{13}{52}$

$P(B) = P(\text{a king card}) = \frac{4}{52}$

P (either a spade or a king card) = P (A or B)

$= P(A \cup B)$

$= P(A) + P(B) - P(A \cap B)$

$= \frac{13}{52} + \frac{4}{52} - \left[\frac{13}{52} \times \frac{4}{52}\right]$

$= \frac{4}{13}$

2. A person is known to hit the target in 3 out of 4 shots, whereas another person is known to hit the target in 2 out of 3 shots. Find the probability of the targets being hit at all when they both person try.

Sol: The prob. that the first person hit the target = P (A) = $\frac{3}{4}$

The prob. that the second person hit the target = P (B) = $\frac{2}{3}$

The two events are not mutually exclusive, since both persons hit the same target.

$P(A \text{ or } B) = P(A \cup B)$

$= P(A) + P(B) - P(A \cap B)$

$= \frac{3}{4} + \frac{2}{3} - \left[\frac{3}{4} \times \frac{2}{3}\right]$

$= \frac{11}{12}$

MULTIPLICATION LAW OF PROBABILITY (INDEPENDENT EVENTS):

If A & B are two independent events, then

$P(A \cap B)$ = P (Both A & B will happen)

$= P(A) \times P(B)$

PROBLEMS:

1. If P (A) = 0.35, P (B) = 0.73, $P(A \cap B)$ = 0.14. Find $P(\bar{A} \cup \bar{B})$

Sol: Using Demargon's Law,

$\bar{A} \cup \bar{B} = \overline{A \cup B}$

$P(\bar{A} \cup \bar{B}) = P(\overline{A \cup B})$

$$P(\bar{A} \cup \bar{B}) = 1 - P(A \cap B)$$

$$= 1 - 0.14 = 0.86$$

2. A bag contains 8 white and 10 black balls. Two balls are drawn in succession. What is the prob. that first is white and second is black.

Sol: Total no. of balls = 8 + 10 = 18

P (drawing one white ball from 8 balls) = $\frac{8}{18}$

P (drawing one black ball from 10 balls) = $\frac{10}{18}$

P (drawing first white & second black) = $\frac{8}{18} \times \frac{10}{18}$

$= \frac{20}{81}$

3. Two persons A & B appear in an interview for 2 vacancies for the same post. The probability of A's selection is $\frac{1}{7}$ and that of B's selection is $\frac{1}{5}$. What is the probability that, i) both of them will be selected, ii) none of them will be selected.

Sol: P (A selected) = $\frac{1}{7}$

P (B selected) = $\frac{1}{5}$

P (A will not be selected) = $1 - \frac{1}{7} = \frac{6}{7}$

P (B will not be selected) = $1 - \frac{1}{5} = \frac{4}{5}$

i) P (Both of them will be selected) = P (A) × P (B)

$= \frac{1}{7} \times \frac{1}{5}$

$= \frac{1}{35}$

ii) P (none of them will be selected) = P (A) × P (B)

$= \frac{6}{7} \times \frac{4}{5}$

$= \frac{24}{35}$

4. A problem in mathematics is given to 3 students A, B, C whose chances of solving it are $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ respectively. What is the prob. that the problem will be solved?

Sol: P (A will not solve the problem) = 1- $\frac{1}{2} = \frac{1}{2}$

P (B will not solve the problem) = 1- $\frac{1}{3} = \frac{2}{3}$

P (C will not solve the problem) = 1- $\frac{1}{4} = \frac{3}{4}$

P (all three will not solve the problem) = $\frac{1}{2} \times \frac{2}{3} \times \frac{3}{4}$

$= \frac{1}{4}$

$\therefore$ P (all the three will solve the problem) = 1- $\frac{1}{4} = \frac{3}{4}$

5. What is the chance of getting two sixes in two rolling of a single die?

Sol: P (getting a six in first rolling) = $\frac{1}{6}$

P (getting a six in second rolling) = $\frac{1}{6}$

Since two rolling are independent.

$\therefore$ P (getting two sixes in 2 rolls) = $\frac{1}{6} \times \frac{1}{6}$

$= \frac{1}{36}$

6. An article manufactured by a company consists of two parts A & B. In the process of manufacture of part A, 9 out of 100 are likely to be defective. Similarly, 5 0ut of 100 are likely to be defective in the manufacture of part B. Calculate the prob. that the assembled article will not be (assuming that the events of finding the part A non-defective and that of B are independent).

Sol: Prob. that part A will be defective = $\frac{9}{100}$

$\therefore$ P (A will not be defective) = 1- $\frac{9}{100}$

$= \frac{100-9}{100}$

$= \frac{91}{100}$

Prob. that part B will be defective = $\frac{5}{100}$

$\therefore$ P (A will not be defective) = 1- $\frac{5}{100}$

$= \frac{100-5}{100}$

$= \frac{95}{100}$

$\therefore$ P (the assembled article will not be defective) = P (A will not be defective) $\times$ P (B will not be defective)

$= \frac{91}{100} \times \frac{95}{100}$

= 0.86

7. From a bag containing 4 white and 6 black balls, two balls are drawn at random. If the balls are drawn one after the other without replacement, find the probability that
   i) both balls are white.
   ii) both balls are black.
   iii) the first ball is white and the second ball is black.
   iv) one ball is white and the other is black.

Sol: Total no. of balls = 4 + 6 = 10

i) P (first ball is white) = $\frac{4}{10}$

P (second ball is white) = $\frac{3}{9}$

$\therefore$ P (both balls are white) = $\frac{4}{10} \times \frac{3}{9}$

$= \frac{2}{15}$

ii) P (first ball is black) = $\frac{6}{10}$

P (second ball is black) = $\frac{5}{9}$

$\therefore$ P (both balls are black) = $\frac{6}{10} \times \frac{5}{9}$

$= \frac{1}{3}$

iii) P (first ball is white) = $\frac{4}{10}$

P (second ball is black) = $\frac{6}{9}$

$\therefore$ P (first ball is white & second ball is black) = $\frac{4}{10} \times \frac{6}{9}$

$= \frac{4}{15}$

iv) a) P (first ball is white & second ball is black) = $\frac{4}{10} \times \frac{6}{9}$

$= \frac{24}{90}$

b) P (first ball is black & second ball is white) = $\frac{6}{10} \times \frac{4}{9}$

$= \frac{24}{90}$

Hence both events (a) & (b) are mutually exclusive.

$\therefore$ P (one ball is white & the other is black) = $\frac{24}{90} + \frac{24}{90}$

$= \frac{8}{15}$

8. Find the probability in each of the below four cases, if the balls are drawn one after the other with replacement. A bag containing 4 white & 6 black balls, 2 balls are drawn at random.
   i) both balls are white.
   ii) both balls are black.
   iii) the first ball is white and the second ball is black.
   iv) one ball is white and the other is black.

Sol: Total no. of balls = 4 + 6 = 10

i) P (first ball is white) = $\frac{4}{10}$

P (second ball is white) = $\frac{4}{10}$

$\therefore$ P (both balls are white) = $\frac{4}{10} \times \frac{4}{10}$

$= \frac{4}{25}$

ii) P (first ball is black) = $\frac{6}{10}$

P (second ball is black) = $\frac{6}{10}$

∴ P (both balls are black) = $\frac{6}{10} \times \frac{6}{10}$

= $\frac{9}{25}$

iii) P (first ball is white) = $\frac{4}{10}$

P (second ball is black) = $\frac{6}{10}$

∴ P (first ball is white & second ball is black) = $\frac{4}{10} \times \frac{6}{10}$

= $\frac{6}{25}$

iv) P (first ball is white & second ball is black) = $\frac{4}{10} \times \frac{6}{10}$

= $\frac{6}{25}$

CONDITIONAL PROBABILITY:

The conditional probability of event A, when the event B has already happened is defined as,

$$P\,(A/B) = \frac{P\,(A \cap B)}{P\,(B)} \quad , P\,(B) \neq 0 \text{ (OR) } P\,(A \cap B) = P\,(A/B)\,.\,P\,(B)$$

If A & B are mutually exclusive events then,

$$P\,(B/A) = \frac{P\,(A \cap B)}{P\,(A)} \quad , P\,(A) \neq 0$$

PROBLEMS:

1. A bag contains 3 red & 4 white balls. Two draws are made without replacement. What is the prob. that both the balls are red.

Sol: P (drawing a red ball in the first draw) = $\frac{3}{7}$

i.e) P (A) = $\frac{3}{7}$

P (drawing a red ball in the first draw given that first ball drawn is red) = $\frac{2}{6}$

i.e) $P\,(B/A)$ = $\frac{2}{6}$

$$\therefore P\,(A \cap B) = \; P\,(B/A) \; \times P\,(A)$$

$$= \frac{2}{6} \times \frac{3}{7}$$

$$= \frac{1}{7}$$

2. Find the prob. of drawing a queen and a king from a pack of cards in two consecutive draws, the cards drawn not being replaced.

Sol: P (drawing a queen card) = $\frac{4}{52}$

i.e) P (A) = $\frac{4}{52}$

P (drawing a king after a queen has been drawn) = $\frac{4}{51}$

i.e) $P\,(B/A) = \frac{4}{51}$

$$\therefore P\,(A \cap B) = \; P\,(B/A) \; \times P\,(A)$$

$$= \frac{4}{51} \times \frac{4}{52}$$

$$= \frac{4}{663}$$

3. In a box there are 100 resistors having resistance and tolerance as shown in the following table. Let a resistor be selected from the box and assume each resistor has the same likelihood of being chosen. Define three events A as draw a 47Ω resistor, B as draw a resistor with 5% tolerance and C as draw a 100Ω resistor. Find $P\,(A/B), P\,(A/C), P\,(B/C)$.

Resistance Ω	5%	10%	Total
22	10	14	24
47	28	16	44
100	24	8	32
Total	62	38	100

Sol: P (A) = P (47Ω) = $\frac{44}{100}$

P (B) = P (5%) = $\frac{62}{100}$

$$P\ (C) = P\ (100\Omega) = \frac{32}{100}$$

The joint probabilities are,

$$P\ (A \cap B) = P\ (47\Omega \cap 5\%)$$

$$= \frac{28}{100}$$

$$P\ (A \cap C) = P\ (47\Omega \cap 100\Omega)$$

$$= 0$$

$$P\ (B \cap C) = P\ (5\% \cap 100\Omega)$$

$$= \frac{24}{100}$$

$$\therefore P\ (A/B) = \frac{P\ (A \cap B)}{P\ (B)} = \frac{28/100}{62/100}$$

$$= \frac{28}{62}$$

$$P\ (A/C) = \frac{P\ (A \cap C)}{P\ (C)} = \frac{0}{32/100}$$

$$= 0$$

$$P\ (B/C) = \frac{P\ (B \cap C)}{P\ (C)} = \frac{24/100}{32/100}$$

$$= \frac{24}{32}$$

4. The Hindu newspaper publishes three columns entitled politics (A), books(B), cinema(C). Reading habits of a randomly selected reader with respect to three columns are,

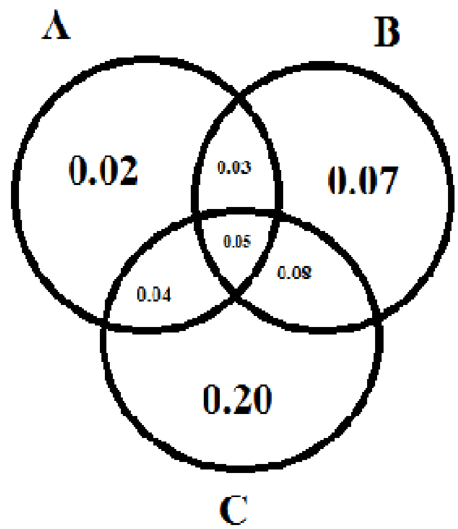
Read Regularly	A	B	C	$A \cap B$	$A \cap C$	$B \cap C$	$A \cap B \cap C$
Probability	0.14	0.23	0.37	0.08	0.09	0.13	0.05

Find P (A/B), P (A/B∪C), P (A/reads atleast one), P (A∪B /C).

Sol: $$P\ (A/B) = \frac{P\ (A \cap B)}{P\ (B)}$$

$$= \frac{0.08}{0.23}$$

$$= 0.348$$



$$P\,(A/B\ \cup C) = \frac{P\,[A\cap(B\cup C)]}{P\,(B\cup C)}$$

$$= \frac{0.04+0.05+0.03}{0.47}$$

$$= 0.255$$

$$P\,(\text{A / reads atleast one}) = P\,[A / (A\cup B\cup C)]$$

$$= \frac{P\,[A\cap(A\cup B\cup C)]}{P\,(A\cup B\cup C)}$$

$$= \frac{P\,(A)}{P\,(A\cup B\cup C)}$$

$$= \frac{0.14}{0.49}$$

$$= 0.286$$

$$P\,(A\cup B\ /\ C) = \frac{P\,[(A\cup B)\cap C]}{P\,(C)}$$

$$= \frac{0.04+0.05+0.08}{0.37}$$

$$= 0.459$$

CHAPTER 2

Random Variables and Probability Distributions

Random Variables

Suppose that to each point of a sample space we assign a number. We then have a *function* defined on the sample space. This function is called a *random variable* (or *stochastic variable*) or more precisely a *random function* (*stochastic function*). It is usually denoted by a capital letter such as X or Y. In general, a random variable has some specified physical, geometrical, or other significance.

EXAMPLE 2.1 Suppose that a coin is tossed twice so that the sample space is $S = \{HH, HT, TH, TT\}$. Let X represent the number of heads that can come up. With each sample point we can associate a number for X as shown in Table 2-1. Thus, for example, in the case of HH (i.e., 2 heads), $X = 2$ while for TH (1 head), $X = 1$. It follows that X is a random variable.

Table 2-1

Sample Point	HH	HT	TH	TT
X	2	1	1	0

It should be noted that many other random variables could also be defined on this sample space, for example, the square of the number of heads or the number of heads minus the number of tails.

A random variable that takes on a finite or countably infinite number of values (see page 4) is called a *discrete random variable* while one which takes on a noncountably infinite number of values is called a *nondiscrete random variable*.

Discrete Probability Distributions

Let X be a discrete random variable, and suppose that the possible values that it can assume are given by $x_1, x_2, x_3, \ldots$, arranged in some order. Suppose also that these values are assumed with probabilities given by

$$P(X = x_k) = f(x_k) \qquad k = 1, 2, \ldots \tag{1}$$

It is convenient to introduce the *probability function*, also referred to as *probability distribution*, given by

$$P(X = x) = f(x) \tag{2}$$

For $x = x_k$, this reduces to (1) while for other values of x, $f(x) = 0$.

In general, $f(x)$ is a probability function if

1. $f(x) \geq 0$
2. $\sum_x f(x) = 1$

where the sum in 2 is taken over all possible values of x.

EXAMPLE 2.2 Find the probability function corresponding to the random variable X of Example 2.1. Assuming that the coin is fair, we have

$$P(HH) = \frac{1}{4} \qquad P(HT) = \frac{1}{4} \qquad P(TH) = \frac{1}{4} \qquad P(TT) = \frac{1}{4}$$

Then

$$P(X = 0) = P(TT) = \frac{1}{4}$$

$$P(X = 1) = P(HT \cup TH) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(X = 2) = P(HH) = \frac{1}{4}$$

The probability function is thus given by Table 2-2.

Table 2-2

x	0	1	2
$f(x)$	1/4	1/2	1/4

Distribution Functions for Random Variables

The *cumulative distribution function*, or briefly the *distribution function*, for a random variable X is defined by

$$F(x) = P(X \le x) \tag{3}$$

where x is any real number, i.e., $-\infty < x < \infty$.

The distribution function $F(x)$ has the following properties:

1. $F(x)$ is nondecreasing [i.e., $F(x) \le F(y)$ if $x \le y$].
2. $\lim_{x \to -\infty} F(x) = 0$; $\lim_{x \to \infty} F(x) = 1$.
3. $F(x)$ is continuous from the right [i.e., $\lim_{h \to 0^+} F(x + h) = F(x)$ for all x].

Distribution Functions for Discrete Random Variables

The distribution function for a discrete random variable X can be obtained from its probability function by noting that, for all x in $(-\infty, \infty)$,

$$F(x) = P(X \le x) = \sum_{u \le x} f(u) \tag{4}$$

where the sum is taken over all values u taken on by X for which $u \le x$.

If X takes on only a finite number of values $x_1, x_2, \ldots, x_n$, then the distribution function is given by

$$F(x) = \begin{cases} 0 & -\infty < x < x_1 \\ f(x_1) & x_1 \le x < x_2 \\ f(x_1) + f(x_2) & x_2 \le x < x_3 \\ \vdots & \vdots \\ f(x_1) + \cdots + f(x_n) & x_n \le x < \infty \end{cases} \tag{5}$$

EXAMPLE 2.3 (a) Find the distribution function for the random variable X of Example 2.2. (b) Obtain its graph.

(a) The distribution function is

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ \frac{1}{4} & 0 \le x < 1 \\ \frac{3}{4} & 1 \le x < 2 \\ 1 & 2 \le x < \infty \end{cases}$$

(b) The graph of $F(x)$ is shown in Fig. 2-1.

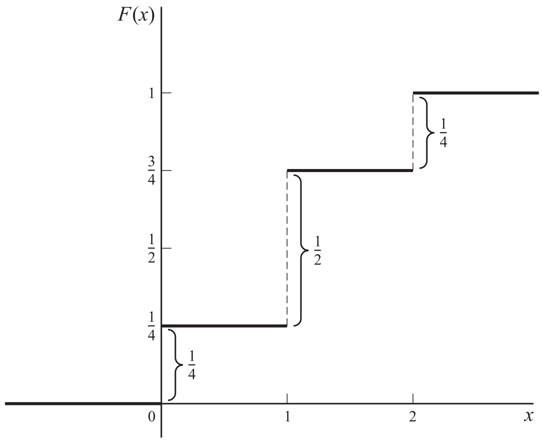


Fig. 2-1

The following things about the above distribution function, which are true in general, should be noted.

1. The magnitudes of the jumps at 0, 1, 2 are $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ which are precisely the probabilities in Table 2-2. This fact enables one to obtain the probability function from the distribution function.
2. Because of the appearance of the graph of Fig. 2-1, it is often called a *staircase function* or *step function*. The value of the function at an integer is obtained from the higher step; thus the value at 1 is $\frac{3}{4}$ and not $\frac{1}{4}$. This is expressed mathematically by stating that the distribution function is *continuous from the right* at 0, 1, 2.
3. As we proceed from left to right (i.e. going *upstairs*), the distribution function either remains the same or increases, taking on values from 0 to 1. Because of this, it is said to be a *monotonically increasing function*.

It is clear from the above remarks and the properties of distribution functions that the probability function of a discrete random variable can be obtained from the distribution function by noting that

$$f(x) = F(x) - \lim_{u \to x^-} F(u). \tag{6}$$

Continuous Random Variables

A nondiscrete random variable X is said to be *absolutely continuous*, or simply *continuous*, if its distribution function may be represented as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\,du \qquad (-\infty < x < \infty) \tag{7}$$

where the function $f(x)$ has the properties

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

It follows from the above that if X is a continuous random variable, then the probability that X takes on any one particular value is zero, whereas the *interval probability* that X lies *between two different values*, say, a and b, is given by

$$P(a < X < b) = \int_{a}^{b} f(x)\,dx \tag{8}$$

EXAMPLE 2.4 If an individual is selected at random from a large group of adult males, the probability that his height X is precisely 68 inches (i.e., 68.000 . . . inches) would be zero. However, there is a probability greater than zero than X is between 67.000 . . . inches and 68.500 . . . inches, for example.

A function $f(x)$ that satisfies the above requirements is called a *probability function* or *probability distribution* for a continuous random variable, but it is more often called a *probability density function* or simply *density function*. Any function $f(x)$ satisfying Properties 1 and 2 above will automatically be a density function, and required probabilities can then be obtained from (8).

EXAMPLE 2.5 (a) Find the constant c such that the function

$$f(x) = \begin{cases} cx^2 & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

is a density function, and (b) compute $P(1 < X < 2)$.

(a) Since $f(x)$ satisfies Property 1 if $c \geq 0$, it must satisfy Property 2 in order to be a density function. Now

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^3 cx^2\,dx = \frac{cx^3}{3}\bigg|_0^3 = 9c$$

and since this must equal 1, we have $c = 1/9$.

(b)
$$P(1 < X < 2) = \int_1^2 \frac{1}{9}x^2\,dx = \frac{x^3}{27}\bigg|_1^2 = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}$$

In case $f(x)$ is continuous, which we shall assume unless otherwise stated, the probability that X is equal to any particular value is zero. In such case we can replace either or both of the signs $<$ in (8) by $\leq$. Thus, in Example 2.5,

$$P(1 \leq X \leq 2) = P(1 \leq X < 2) = P(1 < X \leq 2) = P(1 < X < 2) = \frac{7}{27}$$

EXAMPLE 2.6 (a) Find the distribution function for the random variable of Example 2.5. (b) Use the result of (a) to find $P(1 < x \leq 2)$.

(a) We have

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\,du$$

If $x < 0$, then $F(x) = 0$. If $0 \leq x < 3$, then

$$F(x) = \int_0^x f(u)\,du = \int_0^x \frac{1}{9}u^2\,du = \frac{x^3}{27}$$

If $x \geq 3$, then

$$F(x) = \int_0^3 f(u)\,du + \int_3^x f(u)\,du = \int_0^3 \frac{1}{9}u^2\,du + \int_3^x 0\,du = 1$$

Thus the required distribution function is

$$F(x) = \begin{cases} 0 & x < 0 \\ x^3/27 & 0 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

Note that $F(x)$ increases monotonically from 0 to 1 as is required for a distribution function. It should also be noted that $F(x)$ in this case is continuous.

(b) We have

$$P(1 < X \leq 2) = P(X \leq 2) - P(X \leq 1)$$
$$= F(2) - F(1)$$
$$= \frac{2^3}{27} - \frac{1^3}{27} = \frac{7}{27}$$

as in Example 2.5.

The probability that X is between x and $x + \Delta x$ is given by

$$P(x \leq X \leq x + \Delta x) = \int_x^{x+\Delta x} f(u)\,du \tag{9}$$

so that if Δx is small, we have approximately

$$P(x \leq X \leq x + \Delta x) = f(x)\Delta x \tag{10}$$

We also see from (7) on differentiating both sides that

$$\frac{dF(x)}{dx} = f(x) \tag{11}$$

at all points where $f(x)$ is continuous; i.e., the derivative of the distribution function is the density function.

It should be pointed out that random variables exist that are neither discrete nor continuous. It can be shown that the random variable X with the following distribution function is an example.

$$F(x) = \begin{cases} 0 & x < 1 \\ \dfrac{x}{2} & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

In order to obtain (11), we used the basic property

$$\frac{d}{dx}\int_a^x f(u)\,du = f(x) \tag{12}$$

which is one version of the Fundamental Theorem of Calculus.

Graphical Interpretations

If $f(x)$ is the density function for a random variable X, then we can represent $y = f(x)$ graphically by a curve as in Fig. 2-2. Since $f(x) \geq 0$, the curve cannot fall below the x axis. The entire area bounded by the curve and the x axis must be 1 because of Property 2 on page 36. Geometrically the probability that X is between a and b, i.e., $P(a < X < b)$, is then represented by the area shown shaded, in Fig. 2-2.

The distribution function $F(x) = P(X \leq x)$ is a monotonically increasing function which increases from 0 to 1 and is represented by a curve as in Fig. 2-3.

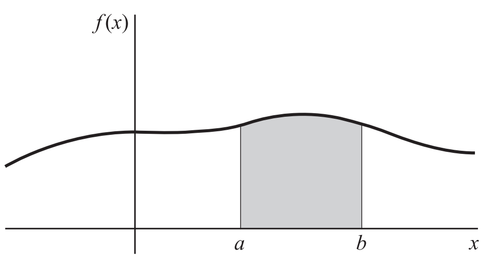


Fig. 2-2

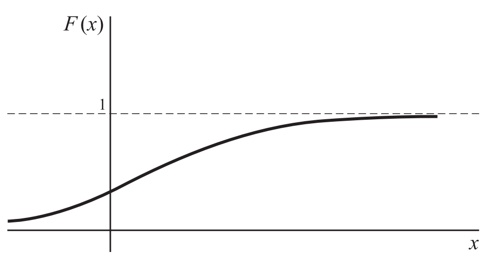


Fig. 2-3

Joint Distributions

The above ideas are easily generalized to two or more random variables. We consider the typical case of two random variables that are either both discrete or both continuous. In cases where one variable is discrete and the other continuous, appropriate modifications are easily made. Generalizations to more than two variables can also be made.

1. DISCRETE CASE. If X and Y are two discrete random variables, we define the *joint probability function* of X and Y by

$$P(X = x, Y = y) = f(x, y) \tag{13}$$

where 1. $f(x, y) \geq 0$

2. $\sum_x \sum_y f(x, y) = 1$

i.e., the sum over all values of x and y is 1.

Suppose that X can assume any one of m values $x_1, x_2, \ldots, x_m$ and Y can assume any one of n values $y_1, y_2, \ldots, y_n$. Then the probability of the event that $X = x_j$ and $Y = y_k$ is given by

$$P(X = x_j, Y = y_k) = f(x_j, y_k) \tag{14}$$

A joint probability function for X and Y can be represented by a *joint probability table* as in Table 2-3. The probability that $X = x_j$ is obtained by adding all entries in the row corresponding to x_i and is given by

$$P(X = x_j) = f_1(x_j) = \sum_{k=1}^{n} f(x_j, y_k) \tag{15}$$

Table 2-3

$X \backslash Y$	y_1	y_2	$\ldots$	y_n	Totals $\downarrow$	
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	$\ldots$	$f(x_1, y_n)$	$f_1(x_1)$	
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	$\ldots$	$f(x_2, y_n)$	$f_1(x_2)$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	
x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	$\ldots$	$f(x_m, y_n)$	$f_1(x_m)$	
Totals $\rightarrow$	$f_2(y_1)$	$f_2(y_2)$	$\ldots$	$f_2(y_n)$	1	$\leftarrow$ Grand Total

For $j = 1, 2, \ldots, m$, these are indicated by the entry totals in the extreme right-hand column or margin of Table 2-3. Similarly the probability that $Y = y_k$ is obtained by adding all entries in the column corresponding to y_k and is given by

$$P(Y = y_k) = f_2(y_k) = \sum_{j=1}^{m} f(x_j, y_k) \tag{16}$$

For $k = 1, 2, \ldots, n$, these are indicated by the entry totals in the bottom row or margin of Table 2-3.

Because the probabilities (15) and (16) are obtained from the margins of the table, we often refer to $f_1(x_j)$ and $f_2(y_k)$ [or simply $f_1(x)$ and $f_2(y)$] as the *marginal probability functions* of X and Y, respectively.

It should also be noted that

$$\sum_{j=1}^{m} f_1(x_j) = 1 \quad \sum_{k=1}^{n} f_2(y_k) = 1 \tag{17}$$

which can be written

$$\sum_{j=1}^{m} \sum_{k=1}^{n} f(x_j, y_k) = 1 \tag{18}$$

This is simply the statement that the total probability of all entries is 1. The *grand total* of 1 is indicated in the lower right-hand corner of the table.

The *joint distribution function* of X and Y is defined by

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{u \leq x} \sum_{v \leq y} f(u, v) \tag{19}$$

In Table 2-3, $F(x, y)$ is the sum of all entries for which $x_j \leq x$ and $y_k \leq y$.

2. CONTINUOUS CASE. The case where both variables are continuous is obtained easily by analogy with the discrete case on replacing sums by integrals. Thus the *joint probability function* for the random variables X and Y (or, as it is more commonly called, the *joint density function* of X and Y) is defined by

1. $f(x, y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$

Graphically $z = f(x, y)$ represents a surface, called the *probability surface*, as indicated in Fig. 2-4. The total volume bounded by this surface and the xy plane is equal to 1 in accordance with Property 2 above. The probability that X lies between a and b while Y lies between c and d is given graphically by the shaded volume of Fig. 2-4 and mathematically by

$$P(a < X < b, c < Y < d) = \int_{x=a}^{b} \int_{y=c}^{d} f(x, y)\, dx\, dy \tag{20}$$

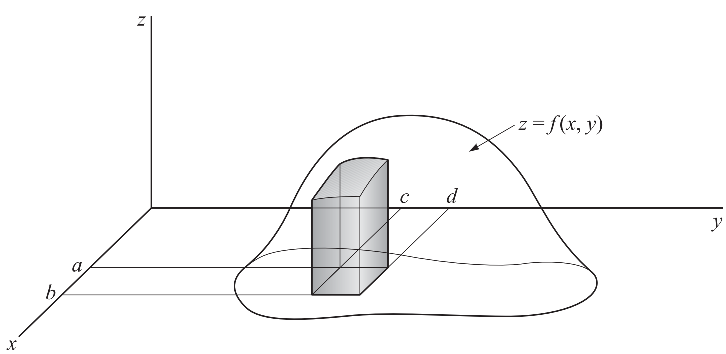


Fig. 2-4

More generally, if A represents any event, there will be a region $\mathcal{R}_A$ of the xy plane that corresponds to it. In such case we can find the probability of A by performing the integration over $\mathcal{R}_A$, i.e.,

$$P(A) = \iint_{\mathcal{R}_A} f(x, y)\, dx\, dy \tag{21}$$

The *joint distribution function* of X and Y in this case is defined by

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u, v)\, du\, dv \tag{22}$$

It follows in analogy with (11), page 38, that

$$\frac{\partial^2 F}{\partial x\,\partial y} = f(x, y) \tag{23}$$

i.e., the density function is obtained by differentiating the distribution function with respect to x and y.

From (22) we obtain

$$P(X \le x) = F_1(x) = \int_{u=-\infty}^{x}\int_{v=-\infty}^{\infty} f(u, v)\,du\,dv \tag{24}$$

$$P(Y \le y) = F_2(y) = \int_{u=-\infty}^{\infty}\int_{v=-\infty}^{y} f(u, v)\,du\,dv \tag{25}$$

We call (24) and (25) the *marginal distribution functions*, or simply the *distribution functions*, of X and Y, respectively. The derivatives of (24) and (25) with respect to x and y are then called the *marginal density functions,* or simply the *density functions*, of X and Y and are given by

$$f_1(x) = \int_{v=-\infty}^{\infty} f(x, v)\,dv \qquad f_2(y) = \int_{u=-\infty}^{\infty} f(u, y)\,du \tag{26}$$

Independent Random Variables

Suppose that X and Y are discrete random variables. If the events $X = x$ and $Y = y$ are independent events for all x and y, then we say that X and Y are *independent random variables*. In such case,

$$P(X = x, Y = y) = P(X = x)P(Y = y) \tag{27}$$

or equivalently

$$f(x, y) = f_1(x)f_2(y) \tag{28}$$

Conversely, if for all x and y the joint probability function $f(x, y)$ can be expressed as the product of a function of x alone and a function of y alone (which are then the marginal probability functions of X and Y), X and Y are independent. If, however, $f(x, y)$ cannot be so expressed, then X and Y are *dependent*.

If X and Y are continuous random variables, we say that they are *independent random variables* if the events $X \le x$ and $Y \le y$ are independent events for all x and y. In such case we can write

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y) \tag{29}$$

or equivalently

$$F(x, y) = F_1(x)F_2(y) \tag{30}$$

where $F_1(z)$ and $F_2(y)$ are the (marginal) distribution functions of X and Y, respectively. Conversely, X and Y are independent random variables if for all x and y, their joint distribution function $F(x, y)$ can be expressed as a product of a function of x alone and a function of y alone (which are the marginal distributions of X and Y, respectively). If, however, $F(x, y)$ cannot be so expressed, then X and Y are dependent.

For continuous independent random variables, it is also true that the joint density function $f(x, y)$ is the product of a function of x alone, $f_1(x)$, and a function of y alone, $f_2(y)$, and these are the (marginal) density functions of X and Y, respectively.

Change of Variables

Given the probability distributions of one or more random variables, we are often interested in finding distributions of other random variables that depend on them in some specified manner. Procedures for obtaining these distributions are presented in the following theorems for the case of discrete and continuous variables.

1. DISCRETE VARIABLES

Theorem 2-1 Let X be a discrete random variable whose probability function is $f(x)$. Suppose that a discrete random variable U is defined in terms of X by $U = \phi(X)$, where to each value of X there corresponds one and only one value of U and conversely, so that $X = \psi(U)$. Then the probability function for U is given by

$$g(u) = f[\psi(u)] \tag{31}$$

Theorem 2-2 Let X and Y be discrete random variables having joint probability function $f(x, y)$. Suppose that two discrete random variables U and V are defined in terms of X and Y by $U = \phi_1(X, Y)$, $V = \phi_2(X, Y)$, where to each pair of values of X and Y there corresponds one and only one pair of values of U and V and conversely, so that $X = \psi_1(U, V)$, $Y = \psi_2(U, V)$. Then the joint probability function of U and V is given by

$$g(u, v) = f[\psi_1(u, v), \psi_2(u, v)] \tag{32}$$

2. CONTINUOUS VARIABLES

Theorem 2-3 Let X be a continuous random variable with probability density $f(x)$. Let us define $U = \phi(X)$ where $X = \psi(U)$ as in Theorem 2-1. Then the probability density of U is given by $g(u)$ where

$$g(u)|du| = f(x)|dx| \tag{33}$$

or

$$g(u) = f(x)\left|\frac{dx}{du}\right| = f[\psi(u)]\,|\psi'(u)| \tag{34}$$

Theorem 2-4 Let X and Y be continuous random variables having joint density function $f(x, y)$. Let us define $U = \phi_1(X, Y)$, $V = \phi_2(X, Y)$ where $X = \psi_1(U, V)$, $Y = \psi_2(U, V)$ as in Theorem 2-2. Then the joint density function of U and V is given by $g(u, v)$ where

$$g(u, v)|du\, dv| = f(x, y)|dx\, dy| \tag{35}$$

or

$$g(u, v) = f(x, y)\left|\frac{\partial(x, y)}{\partial(u, v)}\right| = f[\psi_1(u, v), \psi_2(u, v)]\,|J| \tag{36}$$

In (36) the *Jacobian determinant*, or briefly *Jacobian*, is given by

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} \tag{37}$$

Probability Distributions of Functions of Random Variables

Theorems 2-2 and 2-4 specifically involve joint probability functions of two random variables. In practice one often needs to find the probability distribution of some specified function of several random variables. Either of the following theorems is often useful for this purpose.

Theorem 2-5 Let X and Y be continuous random variables and let $U = \phi_1(X, Y)$, $V = X$ (the second choice is arbitrary). Then the density function for U is the marginal density obtained from the joint density of U and V as found in Theorem 2-4. A similar result holds for probability functions of discrete variables.

Theorem 2-6 Let $f(x, y)$ be the joint density function of X and Y. Then the density function $g(u)$ of the random variable $U = \phi_1(X, Y)$ is found by differentiating with respect to u the distribution function given by

$$G(u) = P[\phi_1(X, Y) \le u] = \iint_{\mathcal{R}} f(x, y)\,dx\,dy \tag{38}$$

Where $\mathcal{R}$ is the region for which $\phi_1(x, y) \le u$.

Convolutions

As a particular consequence of the above theorems, we can show (see Problem 2.23) that the density function of the sum of two continuous random variables X and Y, i.e., of $U = X + Y$, having joint density function $f(x, y)$ is given by

$$g(u) = \int_{-\infty}^{\infty} f(x, u - x)\,dx \tag{39}$$

In the special case where X and Y are independent, $f(x, y) = f_1(x)f_2(y)$, and (39) reduces to

$$g(u) = \int_{-\infty}^{\infty} f_1(x)\,f_2(u - x)\,dx \tag{40}$$

which is called the *convolution* of f_1 and f_2, abbreviated, $f_1 * f_2$.

The following are some important properties of the convolution:

1. $f_1 * f_2 = f_2 * f_1$
2. $f_1 *(f_2 * f_3) = (f_1 * f_2) * f_3$
3. $f_1 *(f_2 + f_3) = f_1 * f_2 + f_1 * f_3$

These results show that f_1, f_2, f_3 obey the *commutative*, *associative*, and *distributive laws* of algebra with respect to the operation of convolution.

Conditional Distributions

We already know that if $P(A) > 0$,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{41}$$

If X and Y are discrete random variables and we have the events (A: $X = x$), (B: $Y = y$), then (41) becomes

$$P(Y = y \mid X = x) = \frac{f(x, y)}{f_1(x)} \tag{42}$$

where $f(x, y) = P(X = x, Y = y)$ is the joint probability function and $f_1(x)$ is the marginal probability function for X. We define

$$f(y \mid x) \equiv \frac{f(x, y)}{f_1(x)} \tag{43}$$

and call it the *conditional probability function of Y given X*. Similarly, the conditional probability function of X given Y is

$$f(x \mid y) \equiv \frac{f(x, y)}{f_2(y)} \tag{44}$$

We shall sometimes denote $f(x \mid y)$ and $f(y \mid x)$ by $f_1(x \mid y)$ and $f_2(y \mid x)$, respectively.

These ideas are easily extended to the case where X, Y are continuous random variables. For example, the *conditional density function of Y given X* is

$$f(y \mid x) \equiv \frac{f(x, y)}{f_1(x)} \tag{45}$$

where $f(x, y)$ is the joint density function of X and Y, and $f_1(x)$ is the marginal density function of X. Using (45) we can, for example, find that the probability of Y being between c and d given that $x < X < x + dx$ is

$$P(c < Y < d \mid x < X < x + dx) = \int_c^d f(y \mid x)\,dy \tag{46}$$

Generalizations of these results are also available.

Applications to Geometric Probability

Various problems in probability arise from geometric considerations or have geometric interpretations. For example, suppose that we have a target in the form of a plane region of area K and a portion of it with area K_1, as in Fig. 2-5. Then it is reasonable to suppose that the probability of hitting the region of area K_1 is proportional to K_1. We thus define

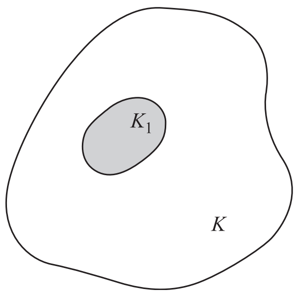


Fig. 2-5

$$P(\text{hitting region of area } K_1) = \frac{K_1}{K} \tag{47}$$

where it is assumed that the probability of hitting the target is 1. Other assumptions can of course be made. For example, there could be less probability of hitting outer areas. The type of assumption used defines the probability distribution function.

SOLVED PROBLEMS

Discrete random variables and probability distributions

2.1. Suppose that a pair of fair dice are to be tossed, and let the random variable X denote the sum of the points. Obtain the probability distribution for X.

The sample points for tosses of a pair of dice are given in Fig. 1-9, page 14. The random variable X is the sum of the coordinates for each point. Thus for (3, 2) we have $X = 5$. Using the fact that all 36 sample points are equally probable, so that each sample point has probability $1/36$, we obtain Table 2-4. For example, corresponding to $X = 5$, we have the sample points (1, 4), (2, 3), (3, 2), (4, 1), so that the associated probability is $4/36$.

Table 2-4

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

2.2. Find the probability distribution of boys and girls in families with 3 children, assuming equal probabilities for boys and girls.

Problem 1.37 treated the case of n mutually independent trials, where each trial had just two possible outcomes, A and A', with respective probabilities p and $q = 1 - p$. It was found that the probability of getting exactly x A's in the n trials is ${}_nC_x p^x q^{n-x}$. This result applies to the present problem, under the assumption that successive births (the "trials") are independent as far as the sex of the child is concerned. Thus, with A being the event "a boy," $n = 3$, and $p = q = \frac{1}{2}$, we have

$$P(\text{exactly } x \text{ boys}) = P(X = x) = {}_3C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{3-x} = {}_3C_x\left(\frac{1}{2}\right)^3$$

where the random variable X represents the number of boys in the family. (Note that X is defined on the sample space of 3 trials.) The probability function for X,

$$f(x) = {}_3C_x\left(\frac{1}{2}\right)^3$$

is displayed in Table 2-5.

Table 2-5

x	0	1	2	3
$f(x)$	1/8	3/8	3/8	1/8

Discrete distribution functions

2.3. (a) Find the distribution function $F(x)$ for the random variable X of Problem 2.1, and (b) graph this distribution function.

(a) We have $F(x) = P(X \le x) = \sum_{u \le x} f(u)$. Then from the results of Problem 2.1, we find

$$F(x) = \begin{cases} 0 & -\infty < x < 2 \\ 1/36 & 2 \le x < 3 \\ 3/36 & 3 \le x < 4 \\ 6/36 & 4 \le x < 5 \\ \vdots & \vdots \\ 35/36 & 11 \le x < 12 \\ 1 & 12 \le x < \infty \end{cases}$$

(b) See Fig. 2-6.

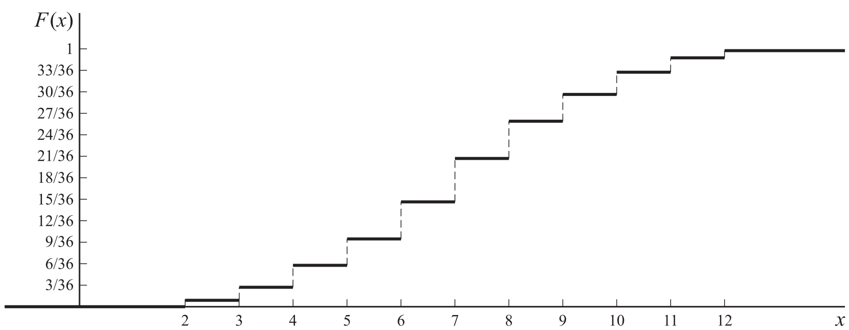


Fig. 2-6

2.4. (a) Find the distribution function $F(x)$ for the random variable X of Problem 2.2, and (b) graph this distribution function.

(a) Using Table 2-5 from Problem 2.2, we obtain

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1/8 & 0 \le x < 1 \\ 1/2 & 1 \le x < 2 \\ 7/8 & 2 \le x < 3 \\ 1 & 3 \le x < \infty \end{cases}$$

(b) The graph of the distribution function of (a) is shown in Fig. 2-7.

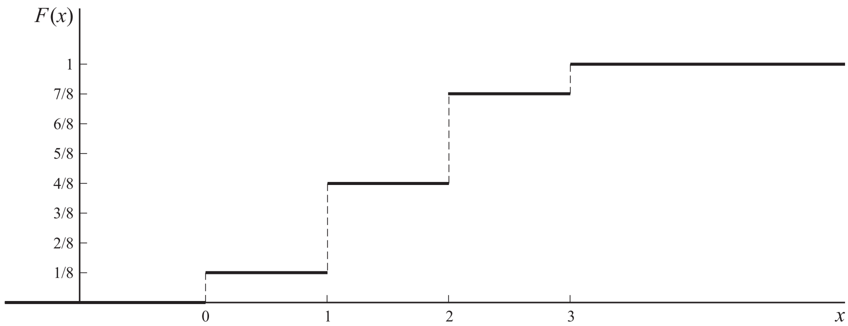


Fig. 2-7

Continuous random variables and probability distributions

2.5. A random variable X has the density function $f(x) = c/(x^2 + 1)$, where $-\infty < x < \infty$. (*a*) Find the value of the constant c. (b) Find the probability that X^2 lies between 1/3 and 1.

(a) We must have $\int_{-\infty}^{\infty} f(x)\,dx = 1$, i.e.,

$$\int_{-\infty}^{\infty} \frac{c\,dx}{x^2+1} = c \tan^{-1} x \Big|_{-\infty}^{\infty} = c\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1$$

so that $c = 1/\pi$.

(b) If $\frac{1}{3} \le X^2 \le 1$, then either $\frac{\sqrt{3}}{3} \le X \le 1$ or $-1 \le X \le -\frac{\sqrt{3}}{3}$. Thus the required probability is

$$\frac{1}{\pi}\int_{-1}^{-\sqrt{3}/3} \frac{dx}{x^2+1} + \frac{1}{\pi}\int_{\sqrt{3}/3}^{1} \frac{dx}{x^2+1} = \frac{2}{\pi}\int_{\sqrt{3}/3}^{1} \frac{dx}{x^2+1}$$

$$= \frac{2}{\pi}\left[\tan^{-1}(1) - \tan^{-1}\left(\frac{\sqrt{3}}{3}\right)\right]$$

$$= \frac{2}{\pi}\left(\frac{\pi}{4} - \frac{\pi}{6}\right) = \frac{1}{6}$$

2.6. Find the distribution function corresponding to the density function of Problem 2.5.

$$F(x) = \int_{-\infty}^{x} f(u)\,du = \frac{1}{\pi}\int_{-\infty}^{x} \frac{du}{u^2+1} = \frac{1}{\pi}\left[\tan^{-1} u \Big|_{-\infty}^{x}\right]$$

$$= \frac{1}{\pi}[\tan^{-1} x - \tan^{-1}(-\infty)] = \frac{1}{\pi}\left[\tan^{-1} x + \frac{\pi}{2}\right]$$

$$= \frac{1}{2} + \frac{1}{\pi}\tan^{-1} x$$

2.7. The distribution function for a random variable X is

$$F(x) = \begin{cases} 1 - e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Find (a) the density function, (b) the probability that $X > 2$, and (c) the probability that $-3 < X \leq 4$.

(a) $$f(x) = \frac{d}{dx}F(x) = \begin{cases} 2e^{-2x} & x > 0 \\ 0 & x < 0 \end{cases}$$

(b) $$P(X > 2) = \int_2^{\infty} 2e^{-2u}\,du = -e^{-2u}\Big|_2^{\infty} = e^{-4}$$

Another method

By definition, $P(X \leq 2) = F(2) = 1 - e^{-4}$. Hence,

$$P(X > 2) = 1 - (1 - e^{-4}) = e^{-4}$$

(c) $$\begin{aligned} P(-3 < X \leq 4) &= \int_{-3}^{4} f(u)\,du = \int_{-3}^{0} 0\,du + \int_0^4 2e^{-2u}\,du \\ &= -e^{-2u}\Big|_0^4 = 1 - e^{-8} \end{aligned}$$

Another method

$$\begin{aligned} P(-3 < X \leq 4) &= P(X \leq 4) - P(X \leq -3) \\ &= F(4) - F(-3) \\ &= (1 - e^{-8}) - (0) = 1 - e^{-8} \end{aligned}$$

Joint distributions and independent variables

2.8. The joint probability function of two discrete random variables X and Y is given by $f(x, y) = c(2x + y)$, where x and y can assume all integers such that $0 \leq x \leq 2$, $0 \leq y \leq 3$, and $f(x, y) = 0$ otherwise.

(a) Find the value of the constant c. (c) Find $P(X \geq 1, Y \leq 2)$.
(b) Find $P(X = 2, Y = 1)$.

(a) The sample points (x, y) for which probabilities are different from zero are indicated in Fig. 2-8. The probabilities associated with these points, given by $c(2x + y)$, are shown in Table 2-6. Since the grand total, $42c$, must equal 1, we have $c = 1/42$.

Table 2-6

X \ Y	0	1	2	3	Totals ↓
0	0	c	$2c$	$3c$	$6c$
1	$2c$	$3c$	$4c$	$5c$	$14c$
2	$4c$	$5c$	$6c$	$7c$	$22c$
Totals →	$6c$	$9c$	$12c$	$15c$	$42c$

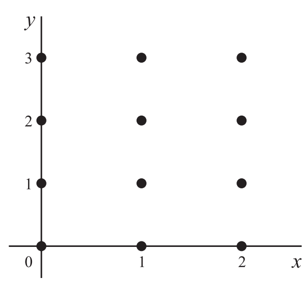


Fig. 2-8

(b) From Table 2-6 we see that

$$P(X = 2, Y = 1) = 5c + \frac{5}{42}$$

(c) From Table 2-6 we see that

$$\begin{aligned} P(X \geq 1, Y \leq 2) &= \sum_{x \geq 1} \sum_{y \leq 2} f(x, y) \\ &= (2c + 3c + 4c)(4c + 5c + 6c) \\ &= 24c = \frac{24}{42} = \frac{4}{7} \end{aligned}$$

as indicated by the entries shown shaded in the table.

2.9. Find the marginal probability functions (a) of X and (b) of Y for the random variables of Problem 2.8.

(a) The marginal probability function for X is given by $P(X = x) = f_1(x)$ and can be obtained from the margin totals in the right-hand column of Table 2-6. From these we see that

$$P(X = x) = f_1(x) = \begin{cases} 6c = 1/7 & x = 0 \\ 14c = 1/3 & x = 1 \\ 22c = 11/21 & x = 2 \end{cases}$$

Check: $\frac{1}{7} + \frac{1}{3} + \frac{11}{21} = 1$

(b) The marginal probability function for Y is given by $P(Y = y) = f_2(y)$ and can be obtained from the margin totals in the last row of Table 2-6. From these we see that

$$P(Y = y) = f_2(y) = \begin{cases} 6c = 1/7 & y = 0 \\ 9c = 3/14 & y = 1 \\ 12c = 2/7 & y = 2 \\ 15c = 5/14 & y = 3 \end{cases}$$

Check: $\frac{1}{7} + \frac{3}{14} + \frac{2}{7} + \frac{5}{14} = 1$

2.10. Show that the random variables X and Y of Problem 2.8 are dependent.

If the random variables X and Y are independent, then we must have, for all x and y,

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

But, as seen from Problems 2.8(b) and 2.9,

$$P(X = 2, Y = 1) = \frac{5}{42} \qquad P(X = 2) = \frac{11}{21} \qquad P(Y = 1) = \frac{3}{14}$$

so that

$$P(X = 2, Y = 1) \neq P(X = 2)P(Y = 1)$$

The result also follows from the fact that the joint probability function $(2x + y)/42$ cannot be expressed as a function of x alone times a function of y alone.

2.11. The joint density function of two continuous random variables X and Y is

$$f(x, y) = \begin{cases} cxy & 0 < x < 4, 1 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of the constant c.

(b) Find $P(1 < X < 2, 2 < Y < 3)$.

(c) Find $P(X \geq 3, Y \leq 2)$.

(a) We must have the total probability equal to 1, i.e.,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1$$

Using the definition of $f(x, y)$, the integral has the value

$$\int_{x=0}^{4}\int_{y=1}^{5} cxy\,dx\,dy = c\int_{x=0}^{4}\left[\int_{y=1}^{5} xy\,dy\right]dx$$

$$= c\int_{z=0}^{4} \frac{xy^2}{2}\bigg|_{y=1}^{5} dx = c\int_{x=0}^{4}\left(\frac{25x}{2} - \frac{x}{2}\right)dx$$

$$= c\int_{x=0}^{4} 12x\,dx = c(6x^2)\bigg|_{x=0}^{4} = 96c$$

Then $96c = 1$ and $c = 1/96$.

(b) Using the value of c found in (a), we have

$$P(1 < X < 2, 2 < Y < 3) = \int_{x=1}^{2}\int_{y=2}^{3} \frac{xy}{96}\,dx\,dy$$

$$= \frac{1}{96}\int_{x=1}^{2}\left[\int_{y=2}^{3} xy\,dy\right]dx = \frac{1}{96}\int_{x=1}^{2} \frac{xy^2}{2}\bigg|_{y=2}^{3} dx$$

$$= \frac{1}{96}\int_{x=1}^{2} \frac{5x}{2}\,dx = \frac{5}{192}\left(\frac{x^2}{2}\right)\bigg|_{1}^{2} = \frac{5}{128}$$

(c)

$$P(X \geq 3, Y \leq 2) = \int_{x=3}^{4}\int_{y=1}^{2} \frac{xy}{96}\,dx\,dy$$

$$= \frac{1}{96}\int_{x=3}^{4}\left[\int_{y=1}^{2} xy\,dy\right]dx = \frac{1}{96}\int_{x=3}^{4} \frac{xy^2}{2}\bigg|_{y=1}^{2} dx$$

$$= \frac{1}{96}\int_{x=3}^{4} \frac{3x}{2}\,dx = \frac{7}{128}$$

2.12. Find the marginal distribution functions (a) of X and (b) of Y for Problem 2.11.

(a) The marginal distribution function for X if $0 \leq x < 4$ is

$$F_1(x) = P(X \leq x) = \int_{u=-\infty}^{x}\int_{v=-\infty}^{\infty} f(u, v)\,du\,dv$$

$$= \int_{u=0}^{x}\int_{v=1}^{5} \frac{uv}{96}\,du\,dv$$

$$= \frac{1}{96}\int_{u=0}^{x}\left[\int_{v=1}^{5} uv\,dv\right]du = \frac{x^2}{16}$$

For $x \geq 4$, $F_1(x) = 1$; for $x < 0$, $F_1(x) = 0$. Thus

$$F_1(x) = \begin{cases} 0 & x < 0 \\ x^{2/16} & 0 \leq x < 4 \\ 1 & x \geq 4 \end{cases}$$

As $F_1\,(x)$ is continuous at $x = 0$ and $x = 4$, we could replace $<$ by $\leq$ in the above expression.

(b) The marginal distribution function for Y if $1 \le y < 5$ is

$$F_2(y) = P(Y \le y) = \int_{u=-\infty}^{\infty}\int_{v=1}^{y} f(u, v)\,du\,dv$$

$$= \int_{u=0}^{4}\int_{v=1}^{y} \frac{uv}{96}\,du\,dv = \frac{y^2 - 1}{24}$$

For $y \ge 5$, $F_2(y) = 1$. For $y < 1$, $F_2(y) = 0$. Thus

$$F_2(y) = \begin{cases} 0 & y < 1 \\ (y^2 - 1)/24 & 1 \le y < 5 \\ 1 & y \ge 5 \end{cases}$$

As $F_2(y)$ is continuous at $y = 1$ and $y = 5$, we could replace $<$ by $\le$ in the above expression.

2.13. Find the joint distribution function for the random variables X, Y of Problem 2.11.

From Problem 2.11 it is seen that the joint density function for X and Y can be written as the product of a function of x alone and a function of y alone. In fact, $f(x, y) = f_1(x)f_2(y)$, where

$$f_1(x) = \begin{cases} c_1 x & 0 < x < 4 \\ 0 & \text{otherwise} \end{cases} \qquad f_2(y) = \begin{cases} c_2 y & 1 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

and $c_1c_2 = c = 1/96$. It follows that X and Y are independent, so that their joint distribution function is given by $F(x, y) = F_1(x)F_2(y)$. The marginal distributions $F_1(x)$ and $F_2(y)$ were determined in Problem 2.12, and Fig. 2-9 shows the resulting piecewise definition of $F(x, y)$.

2.14. In Problem 2.11 find $P(X + Y < 3)$.

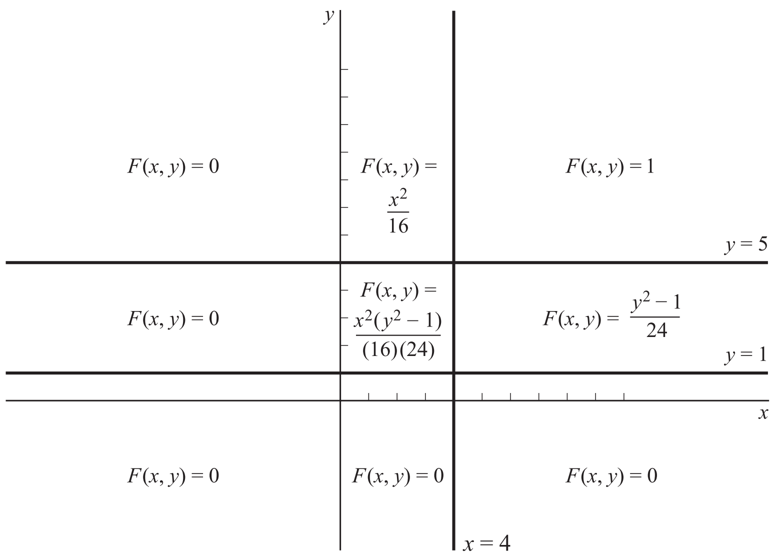


Fig. 2-9

In Fig. 2-10 we have indicated the square region $0 < x < 4$, $1 < y < 5$ within which the joint density function of X and Y is different from zero. The required probability is given by

$$P(X + Y < 3) = \iint_{\mathcal{R}} f(x, y)\,dx\,dy$$

where $\mathcal{R}$ is the part of the square over which $x + y < 3$, shown shaded in Fig. 2-10. Since $f(x, y) = xy/96$ over $\mathcal{R}$, this probability is given by

$$\int_{x=0}^{2}\int_{y=1}^{3-x} \frac{xy}{96}\,dx\,dy$$

$$= \frac{1}{96}\int_{x=0}^{2}\left[\int_{y=1}^{3-x} xy\,dy\right]dx$$

$$= \frac{1}{96}\int_{x=0}^{2} \frac{xy^2}{2}\bigg|_{y=1}^{3-x} dx = \frac{1}{192}\int_{x=0}^{2}[x(3-x)^2 - x] = \frac{1}{48}$$

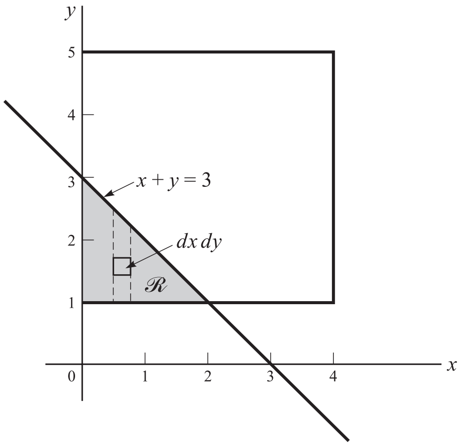


Fig. 2-10

Change of variables

2.15. Prove Theorem 2-1, page 42.

The probability function for U is given by

$$g(u) = P(U = u) = P[\phi(X) = u] = P[X = \psi(u)] = f[\psi(u)]$$

In a similar manner Theorem 2-2, page 42, can be proved.

2.16. Prove Theorem 2-3, page 42.

Consider first the case where $u = \phi(x)$ or $x = \psi(u)$ is an increasing function, i.e., u increases as x increases (Fig. 2-11). There, as is clear from the figure, we have

$$(1) \qquad P(u_1 < U < u_2) = P(x_1 < X < x_2)$$

or

$$(2) \qquad \int_{u_1}^{u_2} g(u)\,du = \int_{x_1}^{x_2} f(x)\,dx$$

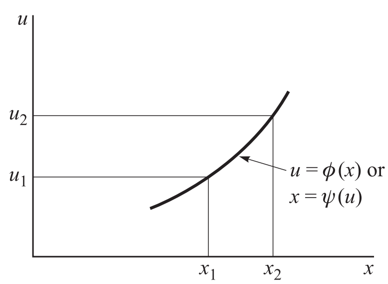


Fig. 2-11

Letting $x = \psi(u)$ in the integral on the right, (2) can be written

$$\int_{u_1}^{u_2} g(u)\,du = \int_{u_1}^{u_2} f[\psi(u)]\,\psi'(u)\,du$$

This can hold for all u_1 and u_2 only if the integrands are identical, i.e.,

$$g(u) = f[\psi(u)]\psi'(u)$$

This is a special case of (34), page 42, where $\psi'(u) > 0$ (i.e., the slope is positive). For the case where $\psi'(u) \leq 0$, i.e., u is a decreasing function of x, we can also show that (34) holds (see Problem 2.67). The theorem can also be proved if $\psi'(u) \geq 0$ or $\psi'(u) < 0$.

2.17. Prove Theorem 2-4, page 42.

We suppose first that as x and y increase, u and v also increase. As in Problem 2.16 we can then show that

$$P(u_1 < U < u_2, v_1 < V < v_2) = P(x_1 < X < x_2, y_1 < Y < y_2)$$

or
$$\int_{v_1}^{u_2}\int_{v_1}^{v_2} g(u, v)\,du\,dv = \int_{x_1}^{x_2}\int_{y_1}^{y_2} f(x, y)\,dx\,dy$$

Letting $x = \psi_1(u, v)$, $y = \psi_2(u, v)$ in the integral on the right, we have, by a theorem of advanced calculus,

$$\int_{v_1}^{u_2}\int_{v_1}^{v_2} g(u, v)\,du\,dv = \int_{u_1}^{u_2}\int_{v_1}^{v_2} f[\psi_1(u, v), \psi_2(u, v)]J\,du\,dv$$

where
$$J = \frac{\partial(x, y)}{\partial(u, v)}$$

is the *Jacobian*. Thus

$$g(u, v) = f[\psi_1(u, v), \psi_2(u, v)]J$$

which is (36), page 42, in the case where $J > 0$. Similarly, we can prove (36) for the case where $J < 0$.

2.18. The probability function of a random variable X is

$$f(x) = \begin{cases} 2^{-x} & x = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Find the probability function for the random variable $U = X^4 + 1$.

Since $U = X^4 + 1$, the relationship between the values u and x of the random variables U and X is given by $u = x^4 + 1$ or $x = \sqrt[4]{u - 1}$, where $u = 2, 17, 82, \ldots$ and the real positive root is taken. Then the required probability function for U is given by

$$g(u) = \begin{cases} 2^{-\sqrt[4]{u-1}} & u = 2, 17, 82, \ldots \\ 0 & \text{otherwise} \end{cases}$$

using Theorem 2-1, page 42, or Problem 2.15.

2.19. The probability function of a random variable X is given by

$$f(x) = \begin{cases} x^2/81 & -3 < x < 6 \\ 0 & \text{otherwise} \end{cases}$$

Find the probability density for the random variable $U = \frac{1}{3}(12 - X)$.

We have $u = \frac{1}{3}(12 - x)$ or $x = 12 - 3u$. Thus to each value of x there is one and only one value of u and conversely. The values of u corresponding to $x = -3$ and $x = 6$ are $u = 5$ and $u = 2$, respectively. Since $\psi'(u) = dx/du = -3$, it follows by Theorem 2-3, page 42, or Problem 2.16 that the density function for U *is*

$$g(u) = \begin{cases} (12 - 3u)^2/27 & 2 < u < 5 \\ 0 & \text{otherwise} \end{cases}$$

Check: $$\int_2^5 \frac{(12 - 3u)^2}{27}\,du = -\frac{(12 - 3u)^3}{243}\bigg|_2^5 = 1$$

2.20. Find the probability density of the random variable $U = X^2$ where X is the random variable of Problem 2.19.

We have $u = x^2$ or $x = \pm\sqrt{u}$. Thus to each value of x there corresponds one and only one value of u, but to each value of $u \neq 0$ there correspond *two* values of x. The values of x for which $-3 < x < 6$ correspond to values of u for which $0 \le u < 36$ as shown in Fig. 2-12.

As seen in this figure, the interval $-3 < x \le 3$ corresponds to $0 \le u \le 9$ while $3 < x < 6$ corresponds to $9 < u < 36$. In this case we cannot use Theorem 2-3 directly but can proceed as follows. The distribution function for U is

$$G(u) = P(U \le u)$$

Now if $0 \le u \le 9$, we have

$$G(u) = P(U \le u) = P(X^2 \le u) = P(-\sqrt{u} \le X \le \sqrt{u})$$

$$= \int_{-\sqrt{u}}^{\sqrt{u}} f(x)\,dx$$

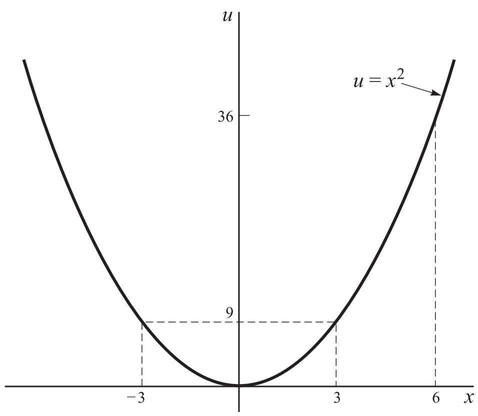


Fig. 2-12

But if $9 < u < 36$, we have

$$G(u) = P(U \le u) = P(-3 < X < \sqrt{u}) = \int_{-3}^{\sqrt{u}} f(x)\,dx$$

Since the density function $g(u)$ is the derivative of $G(u)$, we have, using (12),

$$g(u) = \begin{cases} \dfrac{f(\sqrt{u}) + f(-\sqrt{u})}{2\sqrt{u}} & 0 \le u \le 9 \\ \dfrac{f(\sqrt{u})}{2\sqrt{u}} & 9 < u < 36 \\ 0 & \text{otherwise} \end{cases}$$

Using the given definition of $f(x)$, this becomes

$$g(u) = \begin{cases} \sqrt{u}/81 & 0 \le u \le 9 \\ \sqrt{u}/162 & 9 < u < 36 \\ 0 & \text{otherwise} \end{cases}$$

Check:

$$\int_0^9 \frac{\sqrt{u}}{81}\,du + \int_9^{36} \frac{\sqrt{u}}{162}\,du = \frac{2u^{3/2}}{243}\bigg|_0^9 + \frac{u^{3/2}}{243}\bigg|_9^{36} = 1$$

2.21. If the random variables X and Y have joint density function

$$f(x, y) = \begin{cases} xy/96 & 0 < x < 4,\ 1 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

(see Problem 2.11), find the density function of $U = X + 2Y$.

Method 1

Let $u = x + 2y$, $v = x$, the second relation being chosen arbitrarily. Then simultaneous solution yields $x = v$, $y = \frac{1}{2}(u - v)$. Thus the region $0 < x < 4$, $1 < y < 5$ corresponds to the region $0 < v < 4$, $2 < u - v < 10$ shown shaded in Fig. 2-13.

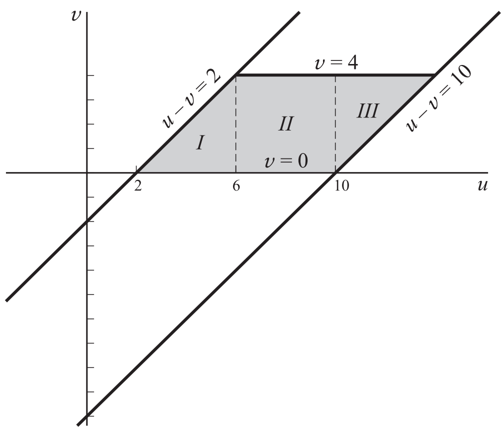


Fig. 2-13

The Jacobian is given by

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \dfrac{1}{2} & -\dfrac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

Then by Theorem 2-4 the joint density function of U and V is

$$g(u, v) = \begin{cases} v(u - v)/384 & 2 < u - v < 10, 0 < v < 4 \\ 0 & \text{otherwise} \end{cases}$$

The marginal density function of U is given by

$$g_1(u) = \begin{cases} \displaystyle\int_{v=0}^{u-2} \frac{v(u - v)}{384}\, dv & 2 < u < 6 \\ \displaystyle\int_{v=0}^{4} \frac{v(u - v)}{384}\, dv & 6 < u < 10 \\ \displaystyle\int_{v=u-10}^{4} \frac{v(u - v)}{384}\, dv & 10 < u < 14 \\ 0 & \text{otherwise} \end{cases}$$

as seen by referring to the shaded regions *I*, *II*, *III* of Fig. 2-13. Carrying out the integrations, we find

$$g_1(u) = \begin{cases} (u - 2)^2(u + 4)/2304 & 2 < u < 6 \\ (3u - 8)/144 & 6 < u < 10 \\ (348u - u^3 - 2128)/2304 & 10 < u < 14 \\ 0 & \text{otherwise} \end{cases}$$

A check can be achieved by showing that the integral of $g_1(u)$ is equal to 1.

Method 2

The distribution function of the random variable $X + 2Y$ is given by

$$P(X + 2Y \leq u) = \iint\limits_{x+2y \leq u} f(x, y)\,dx\,dy = \iint\limits_{\substack{x+2y \leq u \\ 0<x<4 \\ 1<y<5}} \frac{xy}{96}\,dx\,dy$$

For $2 < u < 6$, we see by referring to Fig. 2-14, that the last integral equals

$$\int_{x=0}^{u-2}\int_{y=1}^{(u-x)/2} \frac{xy}{96}\,dx\,dy = \int_{x=0}^{u-2}\left[\frac{x(u - x)^2}{768} - \frac{x}{192}\right]dx$$

The derivative of this with respect to u is found to be $(u - 2)^2(u + 4)/2304$. In a similar manner we can obtain the result of Method 1 for $6 < u < 10$, etc.

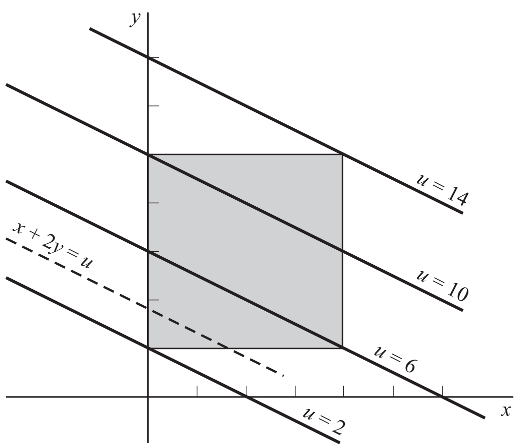


Fig. 2-14

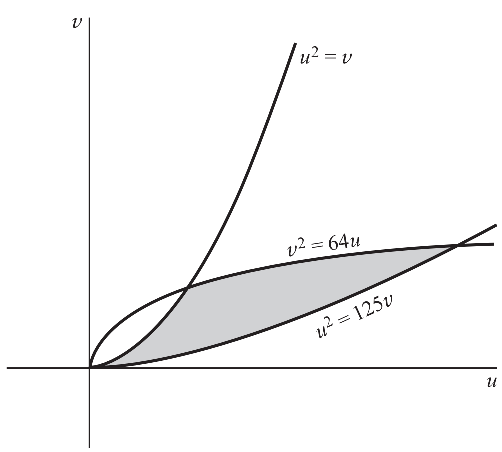


Fig. 2-15

2.22. If the random variables X and Y have joint density function

$$f(x, y) = \begin{cases} xy/96 & 0 < x < 4, 1 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

(see Problem 2.11), find the joint density function of $U = XY^2$, $V = X^2Y$.

Consider $u = xy^2$, $v = x^2y$. Dividing these equations, we obtain $y/x = u/v$ so that $y = ux/v$. This leads to the simultaneous solution $x = v^{2/3} u^{-1/3}$, $y = u^{2/3} v^{-1/3}$. The image of $0 < x < 4$, $1 < y < 5$ in the uv-plane is given by

$$0 < v^{2/3}u^{-1/3} < 4 \qquad 1 < u^{2/3}v^{-1/3} < 5$$

which are equivalent to

$$v^2 < 64u \qquad v < u^2 < 125v$$

This region is shown shaded in Fig. 2-15.

The Jacobian is given by

$$J = \begin{vmatrix} -\frac{1}{3}v^{2/3}u^{-4/3} & \frac{2}{3}v^{-1/3}u^{-1/3} \\ \frac{2}{3}u^{-1/3}v^{-1/3} & -\frac{1}{3}u^{2/3}v^{-4/3} \end{vmatrix} = -\frac{1}{3}u^{-2/3}v^{-2/3}$$

Thus the joint density function of U and V is, by Theorem 2-4,

$$g(u, v) = \begin{cases} \dfrac{(v^{2/3}u^{-1/3})(u^{2/3}v^{-1/3})}{96}\left(\tfrac{1}{3}u^{-2/3}v^{-2/3}\right) & v^2 < 64u, v < u^2 < 125v \\ 0 & \text{otherwise} \end{cases}$$

or

$$g(u, v) = \begin{cases} u^{-1/3}v^{-1/3}/288 & v^2 < 64u, \quad v < u^2 < 125v \\ 0 & \text{otherwise} \end{cases}$$

Convolutions

2.23. Let X and Y be random variables having joint density function $f(x, y)$. Prove that the density function of $U = X + Y$ is

$$g(u) = \int_{-\infty}^{\infty} f(v, u - v)\,dv$$

Method 1

Let $U = X + Y$, $V = X$, where we have arbitrarily added the second equation. Corresponding to these we have $u = x + y$, $v = x$ or $x = v$, $y = u - v$. The Jacobian of the transformation is given by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1$$

Thus by Theorem 2-4, page 42, the joint density function of U and V is

$$g(u, v) = f(v, u - v)$$

It follows from (26), page 41, that the marginal density function of U is

$$g(u) = \int_{-\infty}^{\infty} f(v, u - v)\,dv$$

Method 2

The distribution function of $U = X + Y$ is equal to the double integral of $f(x, y)$ taken over the region defined by $x + y \leq u$, i.e.,

$$G(u) = \iint\limits_{x+y \leq u} f(x, y)\, dx\, dy$$

Since the region is below the line $x + y = u$, as indicated by the shading in Fig. 2-16, we see that

$$G(u) = \int_{x=-\infty}^{\infty} \left[\int_{y=-\infty}^{u-x} f(x, y)\, dy \right] dx$$

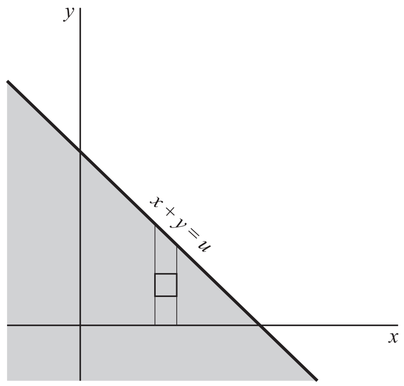


Fig. 2-16

The density function of U is the derivative of $G(u)$ with respect to u and is given by

$$g(u) = \int_{-\infty}^{\infty} f(x, u - x)\, dx$$

using (12) first on the x integral and then on the y integral.

2.24. Work Problem 2.23 if X and Y are independent random variables having density functions $f_1(x)$, $f_2(y)$, respectively.

In this case the joint density function is $f(x, y) = f_1(x) f_2(y)$, so that by Problem 2.23 the density function of $U = X + Y$ is

$$g(u) = \int_{-\infty}^{\infty} f_1(v) f_2(u - v) dv = f_1 * f_2$$

which is the *convolution* of f_1 and f_2.

2.25. If X and Y are independent random variables having density functions

$$f_1(x) = \begin{cases} 2e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad f_2(y) = \begin{cases} 3e^{-3y} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

find the density function of their sum, $U = X + Y$.

By Problem 2.24 the required density function is the convolution of f_1 and f_2 and is given by

$$g(u) = f_1 * f_2 = \int_{-\infty}^{\infty} f_1(v) f_2(u - v)\, dv$$

In the integrand f_1 vanishes when $v < 0$ and f_2 vanishes when $v > u$. Hence

$$g(u) = \int_0^u (2e^{-2v})(3e^{-3(u-v)})\, dv$$

$$= 6e^{-3u} \int_0^u e^v\, dv = 6e^{-3u}(e^u - 1) = 6(e^{-2u} - e^{3u})$$

if $u \geq 0$ and $g(u) = 0$ if $u < 0$.

Check: $$\int_{-\infty}^{\infty} g(u)\,du = 6\int_{0}^{\infty}(e^{-2u} - e^{-3u})\,du = 6\left(\frac{1}{2} - \frac{1}{3}\right) = 1$$

2.26. Prove that $f_1 * f_2 = f_2 * f_1$ (Property 1, page 43).

We have

$$f_1 * f_2 = \int_{v=-\infty}^{\infty} f_1(v)f_2(u - v)\,dv$$

Letting $w = u - v$ so that $v = u - w$, $dv = -dw$, we obtain

$$f_1 * f_2 = \int_{w=\infty}^{-\infty} f_1(u - w)f_2(w)(-dw) = \int_{w=-\infty}^{\infty} f_2(w)f_1(u - w)\,dw = f_2 * f_1$$

Conditional distributions

2.27. Find (a) $f(y\,|\,2)$, (b) $P(Y = 1\,|\,X = 2)$ for the distribution of Problem 2.8.

(a) Using the results in Problems 2.8 and 2.9, we have

$$f(y\,|\,x) = \frac{f(x, y)}{f_1(x)} = \frac{(2x + y)/42}{f_1(x)}$$

so that with $x = 2$

$$f(y\,|\,2) = \frac{(4 + y)/42}{11/21} = \frac{4 + y}{22}$$

(b) $$P(Y = 1\,|\,X = 2) = f(1\,|\,2) = \frac{5}{22}$$

2.28. If X and Y have the joint density function

$$f(x, y) = \begin{cases} \frac{3}{4} + xy & 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

find (a) $f(y\,|\,x)$, (b) $P(Y > \frac{1}{2}\,|\,\frac{1}{2} < X < \frac{1}{2} + dx)$.

(a) For $0 < x < 1$,

$$f_1(x) = \int_0^1 \left(\frac{3}{4} + xy\right)dy = \frac{3}{4} + \frac{x}{2}$$

and $$f(y\,|\,x) = \frac{f(x, y)}{f_1(x)} = \begin{cases} \dfrac{3 + 4xy}{3 + 2x} & 0 < y < 1 \\ 0 & \text{other } y \end{cases}$$

For other values of x, $f(y\,|\,x)$ is not defined.

(b) $$P(Y > \tfrac{1}{2}\,|\,\tfrac{1}{2} < X < \tfrac{1}{2} + dx) = \int_{1/2}^{\infty} f(y\,|\,\tfrac{1}{2})\,dy = \int_{1/2}^{1} \frac{3 + 2y}{4}\,dy = \frac{9}{16}$$

2.29. The joint density function of the random variables X and Y is given by

$$f(x, y) = \begin{cases} 8xy & 0 \leq x \leq 1,\ 0 \leq y \leq x \\ 0 & \text{otherwise} \end{cases}$$

Find (a) the marginal density of X, (b) the marginal density of Y, (c) the conditional density of X, (d) the conditional density of Y.

The region over which $f(x, y)$ is different from zero is shown shaded in Fig. 2-17.

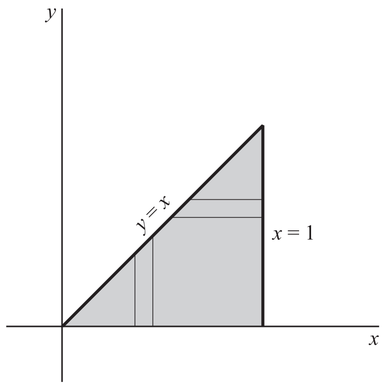


Fig. 2-17

(a) To obtain the marginal density of X, we fix x and integrate with respect to y from 0 to x as indicated by the vertical strip in Fig. 2-17. The result is

$$f_1(x) = \int_{y=0}^{x} 8xy\,dy = 4x^3$$

for $0 < x < 1$. For all other values of x, $f_1(x) = 0$.

(b) Similarly, the marginal density of Y is obtained by fixing y and integrating with respect to x from $x = y$ to $x = 1$, as indicated by the horizontal strip in Fig. 2-17. The result is, for $0 < y < 1$,

$$f_2(y) = \int_{x=y}^{1} 8xy\,dx = 4y(1 - y^2)$$

For all other values of y, $f_2(y) = 0$.

(c) The conditional density function of X is, for $0 < y < 1$,

$$f_1(x \mid y) = \frac{f(x, y)}{f_2(y)} = \begin{cases} 2x/(1 - y^2) & y \le x \le 1 \\ 0 & \text{other } x \end{cases}$$

The conditional density function is not defined when $f_2(y) = 0$.

(d) The conditional density function of Y is, for $0 < x < 1$,

$$f_2(y \mid x) = \frac{f(x, y)}{f_1(x)} = \begin{cases} 2y/x^2 & 0 \le y \le x \\ 0 & \text{other } y \end{cases}$$

The conditional density function is not defined when $f_1(x) = 0$.

Check:

$$\int_0^1 f_1(x)\,dx = \int_0^1 4x^3\,dx = 1, \quad \int_0^1 f_2(y)\,dy = \int_0^1 4y(1 - y^2)\,dy = 1$$

$$\int_y^1 f_1(x \mid y)\,dx = \int_y^1 \frac{2x}{1 - y^2}\,dx = 1$$

$$\int_0^x f_2(y \mid x)\,dy = \int_0^x \frac{2y}{x^2}\,dy = 1$$

2.30. Determine whether the random variables of Problem 2.29 are independent.

In the shaded region of Fig. 2-17, $f(x, y) = 8xy$, $f_1(x) = 4x^3$, $f_2(y) = 4y(1 - y^2)$. Hence $f(x, y) \ne f_1(x) f_2(y)$, and thus X and Y are dependent.

It should be noted that it does not follow from $f(x, y) = 8xy$ that $f(x, y)$ can be expressed as a function of x alone times a function of y alone. This is because the restriction $0 \le y \le x$ occurs. If this were replaced by some restriction on y not depending on x (as in Problem 2.21), such a conclusion would be valid.

Applications to geometric probability

2.31. A person playing darts finds that the probability of the dart striking between r and $r + dr$ is

$$P(r \le R \le r + dr) = c\left[1 - \left(\frac{r}{a}\right)^2\right]dr$$

Here, R is the distance of the hit from the center of the target, c is a constant, and a is the radius of the target (see Fig. 2-18). Find the probability of hitting the bull's-eye, which is assumed to have radius b. Assume that the target is always hit.

The density function is given by

$$f(r) = c\left[1 - \left(\frac{r}{a}\right)^2\right]$$

Since the target is always hit, we have

$$c\int_0^a \left[1 - \left(\frac{r}{a}\right)^2\right]dr = 1$$

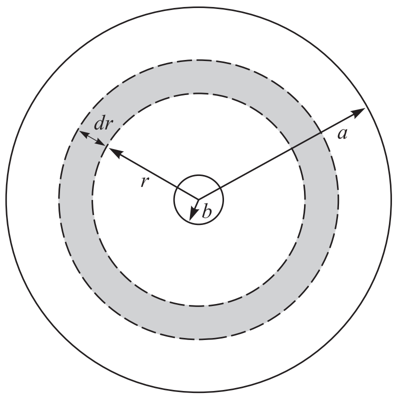


Fig. 2-18

from which $c = 3/2a$. Then the probability of hitting the bull's-eye is

$$\int_0^b f(r)\,dr = \frac{3}{2a}\int_0^b \left[1 - \left(\frac{r}{a}\right)^2\right]dr = \frac{b(3a^2 - b^2)}{2a^3}$$

2.32. Two points are selected at random in the interval $0 \le x \le 1$. Determine the probability that the sum of their squares is less than 1.

Let X and Y denote the random variables associated with the given points. Since equal intervals are assumed to have equal probabilities, the density functions of X and Y are given, respectively, by

(1) $$f_1(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad f_2(y) = \begin{cases} 1 & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Then since X and Y are independent, the joint density function is given by

(2) $$f(x, y) = f_1(x)f_2(y) = \begin{cases} 1 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

It follows that the required probability is given by

(3) $$P(X^2 + Y^2 \le 1) = \iint_{\mathcal{R}} dx\,dy$$

where $\mathcal{R}$ is the region defined by $x^2 + y^2 \le 1, x \ge 0, y \ge 0$, which is a quarter of a circle of radius 1 (Fig. 2-19). Now since (3) represents the area of $\mathcal{R}$, we see that the required probability is $\pi/4$.

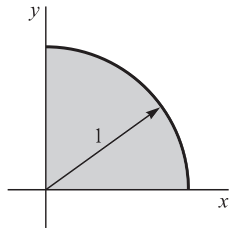


Fig. 2-19

Miscellaneous problems

2.33. Suppose that the random variables X and Y have a joint density function given by

$$f(x, y) = \begin{cases} c(2x + y) & 2 < x < 6, 0 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) the constant c, (b) the marginal distribution functions for X and Y, (c) the marginal density functions for X and Y, (d) $P(3 < X < 4, Y > 2)$, (e) $P(X > 3)$, (f) $P(X + Y > 4)$, (g) the joint distribution function, (h) whether X and Y are independent.

(a) The total probability is given by

$$\int_{x=2}^{6}\int_{y=0}^{5} c(2x + y)\,dx\,dy = \int_{x=2}^{6} c\left(2xy + \frac{y^2}{2}\right)\bigg|_0^5 dx$$

$$= \int_{x=2}^{6} c\left(10x + \frac{25}{2}\right)dx = 210c$$

For this to equal 1, we must have $c = 1/210$.

(b) The marginal distribution function for X is

$$F_1(x) = P(X \le x) = \int_{u=-\infty}^{x}\int_{v=-\infty}^{\infty} f(u, v)\,du\,dv$$

$$= \begin{cases} \displaystyle\int_{u=-\infty}^{x}\int_{v=-\infty}^{\infty} 0\,du\,dv = 0 & x < 2 \\ \displaystyle\int_{u=2}^{x}\int_{v=0}^{5} \frac{2u + v}{210}\,du\,dv = \frac{2x^2 + 5x - 18}{84} & 2 \le x < 6 \\ \displaystyle\int_{u=2}^{6}\int_{v=0}^{5} \frac{2u + v}{210}\,du\,dv = 1 & x \ge 6 \end{cases}$$

The marginal distribution function for Y is

$$F_2(y) = P(Y \le y) = \int_{u=-\infty}^{\infty}\int_{v=-\infty}^{y} f(u, v)\,du\,dv$$

$$= \begin{cases} \displaystyle\int_{u=-\infty}^{\infty}\int_{v=-8}^{y} 0\,du\,dv = 0 & y < 0 \\ \displaystyle\int_{u=0}^{6}\int_{v=0}^{y} \frac{2u + v}{210}\,du\,dv = \frac{y^2 + 16y}{105} & 0 \le y < 5 \\ \displaystyle\int_{u=2}^{6}\int_{v=0}^{5} \frac{2u + v}{210}\,du\,dv = 1 & y \ge 5 \end{cases}$$

(c) The marginal density function for X is, from part (b),

$$f_1(x) = \frac{d}{dx}F_1(x) = \begin{cases} (4x + 5)/84 & 2 < x < 6 \\ 0 & \text{otherwise} \end{cases}$$

The marginal density function for Y is, from part (b),

$$f_2(y) = \frac{d}{dy}F_2(y) = \begin{cases} (2y + 16)/105 & 0 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

(d)
$$P(3 < X < 4, Y > 2) = \frac{1}{210}\int_{x=3}^{4}\int_{y=2}^{5}(2x + y)\,dx\,dy = \frac{3}{20}$$

(e)
$$P(X > 3) = \frac{1}{210}\int_{x=3}^{6}\int_{y=0}^{5}(2x + y)\,dx\,dy = \frac{23}{28}$$

(f)
$$P(X + Y > 4) = \iint_{\mathcal{R}} f(x, y)\,dx\,dy$$

where $\mathcal{R}$ is the shaded region of Fig. 2-20. Although this can be found, it is easier to use the fact that

$$P(X + Y > 4) = 1 - P(X + Y \le 4) = 1 - \iint_{\mathcal{R}} f(x, y)\,dx\,dy$$

where $\mathcal{R}'$ is the cross-hatched region of Fig. 2-20. We have

$$P(X + Y \le 4) = \frac{1}{210}\int_{x=2}^{4}\int_{y=0}^{4-x}(2x + y)\,dx\,dy = \frac{2}{35}$$

Thus $P(X + Y > 4) = 33/35$.

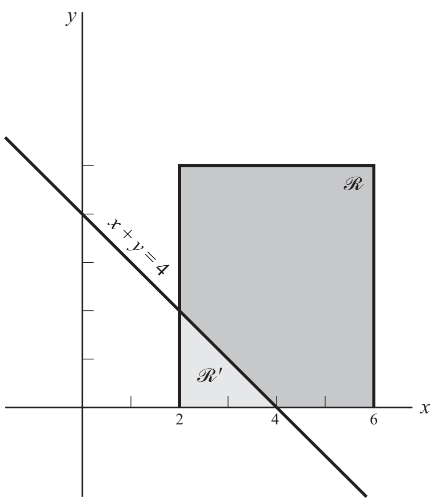


Fig. 2-20

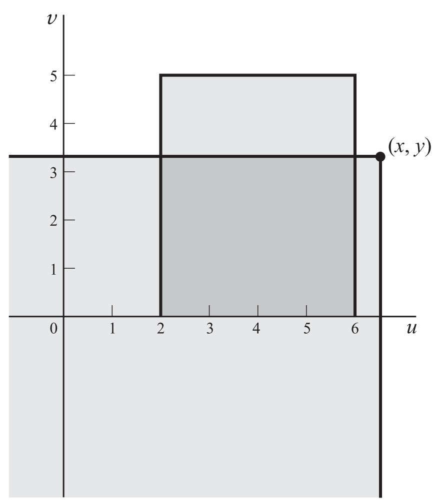


Fig. 2-21

(g) The joint distribution function is

$$F(x, y) = P(X \le x, Y \le y) = \int_{u=-\infty}^{x}\int_{v=-\infty}^{y} f(u, v)\,du\,dv$$

In the uv plane (Fig. 2-21) the region of integration is the intersection of the quarter plane $u \le x$, $v \le y$ and the rectangle $2 < u < 6$, $0 < v < 5$ [over which $f(u, v)$ is nonzero]. For (x, y) located as in the figure, we have

$$F(x, y) = \int_{u=2}^{6}\int_{v=0}^{y}\frac{2u + v}{210}\,du\,dv = \frac{16y + y^2}{105}$$

When (x, y) lies inside the rectangle, we obtain another expression, etc. The complete results are shown in Fig. 2-22.

(h) The random variables are dependent since

$$f(x, y) \neq f_1(x)f_2(y)$$

or equivalently, $F(x, y) \neq F_1(x)F_2(y)$.

2.34. Let X have the density function

$$f(x) = \begin{cases} 6x(1 - x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find a function $Y = h(X)$ which has the density function

$$g(y) = \begin{cases} 12y^3(1 - y^2) & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

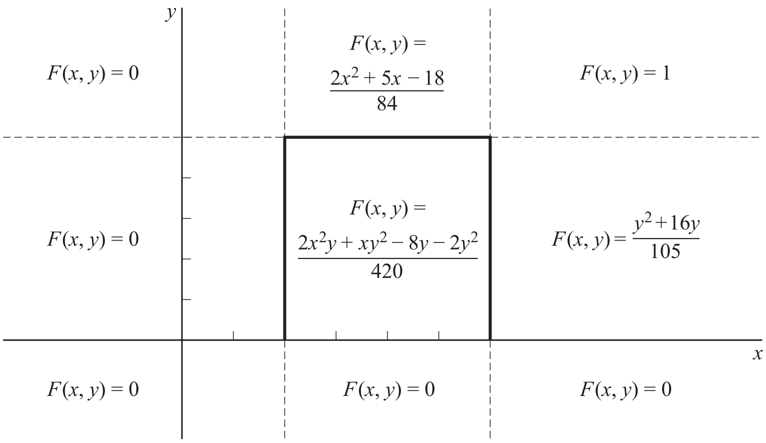


Fig. 2-22

We assume that the unknown function h is such that the intervals $X \leq x$ and $Y \leq y + h(x)$ correspond in a one-one, continuous fashion. Then $P(X \leq x) = P(Y \leq y)$, i.e., the distribution functions of X and Y must be equal. Thus, for $0 < x, y < 1$,

$$\int_0^x 6u(1 - u)\,du = \int_0^y 12v^3(1 - v^2)\,dv$$

or

$$3x^2 - 2x^3 = 3y^4 - 2y^6$$

By inspection, $x = y^2$ or $y = h(x) = +\sqrt{x}$ is a solution, and this solution has the desired properties. Thus $Y = +\sqrt{X}$.

2.35. Find the density function of $U = XY$ if the joint density function of X and Y is $f(x, y)$.

Method 1

Let $U = XY$ and $V = X$, corresponding to which $u = xy$, $v = x$ or $x = v$, $y = u/v$. Then the Jacobian is given by

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ v^{-1} & -uv^{-2} \end{vmatrix} = -v^{-1}$$

Thus the joint density function of U and V is

$$g(u, v) = \frac{1}{|v|} f\left(v, \frac{u}{v}\right)$$

from which the marginal density function of U is obtained as

$$g(u) = \int_{-\infty}^{\infty} g(u, v)\, dv = \int_{-\infty}^{\infty} \frac{1}{|v|} f\left(v, \frac{u}{v}\right) dv$$

Method 2

The distribution function of U is

$$G(u) = \iint\limits_{xy \le u} f(x, y)\, dx\, dy$$

For $u \ge 0$, the region of integration is shown shaded in Fig. 2-23. We see that

$$G(u) = \int_{-\infty}^{0} \left[\int_{u/x}^{\infty} f(x, y)\, dy \right] dx + \int_{0}^{\infty} \left[\int_{-\infty}^{u/x} f(x, y)\, dy \right] dx$$

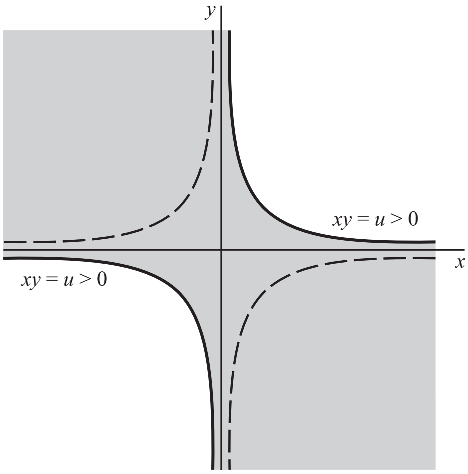


Fig. 2-23

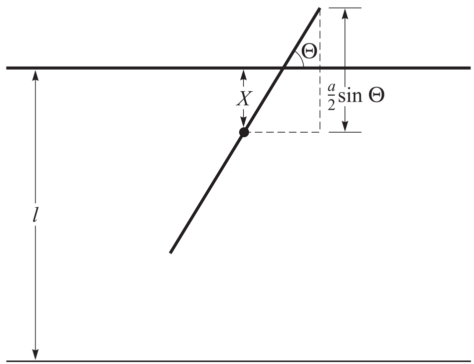


Fig. 2-24

Differentiating with respect to u, we obtain

$$g(u) = \int_{-\infty}^{0} \left(\frac{-1}{x}\right) f\left(x, \frac{u}{x}\right) dx + \int_{0}^{\infty} \frac{1}{x} f\left(x, \frac{u}{x}\right) dx = \int_{-\infty}^{\infty} \frac{1}{|x|} f\left(x, \frac{u}{x}\right) dx$$

The same result is obtained for $u < 0$, when the region of integration is bounded by the dashed hyperbola in Fig. 2-24.

2.36. A floor has parallel lines on it at equal distances l from each other. A needle of length $a < l$ is dropped at random onto the floor. Find the probability that the needle will intersect a line. (This problem is known as *Buffon's needle problem.*)

Let X be a random variable that gives the distance of the midpoint of the needle to the nearest line (Fig. 2-24). Let Θ be a random variable that gives the acute angle between the needle (or its extension) and the line. We denote by x and θ any particular values of X and Θ. It is seen that X can take on any value between 0 and $l/2$, so that $0 \le x \le l/2$. Also Θ can take on any value between 0 and $\pi/2$. It follows that

$$P(x < X \le x + dx) = \frac{2}{l}\, dx \qquad P(\theta \le \Theta + d\theta) = \frac{2}{\pi}\, d\theta$$

i.e., the density functions of X and Θ are given by $f_1(x) = 2/l$, $f_2(\theta) = 2/\pi$. As a check, we note that

$$\int_0^{l/2} \frac{2}{l}\, dx = 1 \qquad \int_0^{\pi/2} \frac{2}{\pi}\, d\theta = 1$$

Since X and Θ are independent the joint density function is

$$f(x, \theta) = \frac{2}{l} \cdot \frac{2}{\pi} = \frac{4}{l\pi}$$

From Fig. 2-24 it is seen that the needle actually hits a line when $X \le (a/2) \sin \Theta$. The probability of this event is given by

$$\frac{4}{l\pi} \int_{\theta=0}^{\pi/2} \int_{x=0}^{(a/2)\sin\theta} dx\, d\theta = \frac{2a}{l\pi}$$

When the above expression is equated to the frequency of hits observed in actual experiments, accurate values of π are obtained. This indicates that the probability model described above is appropriate.

2.37. Two people agree to meet between 2:00 P.M. and 3:00 P.M., with the understanding that each will wait no longer than 15 minutes for the other. What is the probability that they will meet?

Let X and Y be random variables representing the times of arrival, measured in fractions of an hour after 2:00 P.M., of the two people. Assuming that equal intervals of time have equal probabilities of arrival, the density functions of X and Y are given respectively by

$$f_1(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(y) = \begin{cases} 1 & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Then, since X and Y are independent, the joint density function is

$$(1) \qquad f(x, y) = f_1(x) f_2(y) = \begin{cases} 1 & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Since 15 minutes $= \frac{1}{4}$ hour, the required probability is

$$(2) \qquad P\left(|X - Y| \le \frac{1}{4}\right) = \iint_{\mathcal{R}} dx\, dy$$

where 5 is the region shown shaded in Fig. 2-25. The right side of (2) is the area of this region, which is equal to $1 - (\frac{3}{4})(\frac{3}{4}) = \frac{7}{16}$, since the square has area 1, while the two corner triangles have areas $\frac{1}{2}(\frac{3}{4})(\frac{3}{4})$ each. Thus the required probability is $7/16$.

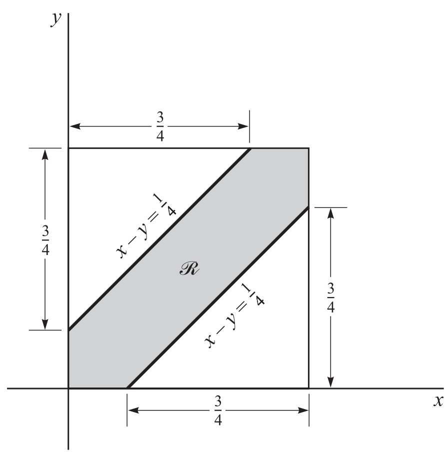


Fig. 2-25

SUPPLEMENTARY PROBLEMS

Discrete random variables and probability distributions

2.38. A coin is tossed three times. If X is a random variable giving the number of heads that arise, construct a table showing the probability distribution of X.

2.39. An urn holds 5 white and 3 black marbles. If 2 marbles are to be drawn at random without replacement and X denotes the number of white marbles, find the probability distribution for X.

2.40. Work Problem 2.39 if the marbles are to be drawn with replacement.

2.41. Let Z be a random variable giving the number of heads minus the number of tails in 2 tosses of a fair coin. Find the probability distribution of Z. Compare with the results of Examples 2.1 and 2.2.

2.42. Let X be a random variable giving the number of aces in a random draw of 4 cards from an ordinary deck of 52 cards. Construct a table showing the probability distribution of X.

Discrete distribution functions

2.43. The probability function of a random variable X is shown in Table 2-7. Construct a table giving the distribution function of X.

Table 2-7

x	1	2	3
$f(x)$	1/2	1/3	1/6

Table 2-8

x	1	2	3	4
$F(x)$	1/8	3/8	3/4	1

2.44. Obtain the distribution function for (a) Problem 2.38, (b) Problem 2.39, (c) Problem 2.40.

2.45. Obtain the distribution function for (a) Problem 2.41, (b) Problem 2.42.

2.46. Table 2-8 shows the distribution function of a random variable X. Determine (a) the probability function, (b) $P(1 \le X \le 3)$, (c) $P(X \ge 2)$, (d) $P(X < 3)$, (e) $P(X > 1.4)$.

Continuous random variables and probability distributions

2.47. A random variable X has density function

$$f(x) = \begin{cases} ce^{-3x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

Find (a) the constant c, (b) $P(1 < X < 2)$, (c) $P(X \ge 3)$, (d) $P(X < 1)$.

2.48. Find the distribution function for the random variable of Problem 2.47. Graph the density and distribution functions, describing the relationship between them.

2.49. A random variable X has density function

$$f(x) = \begin{cases} cx^2 & 1 \le x \le 2 \\ cx & 2 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) the constant c, (b) $P(X > 2)$, (c) $P(1/2 < X < 3/2)$.

2.50. Find the distribution function for the random variable X of Problem 2.49.

2.51. The distribution function of a random variable X is given by

$$F(x) = \begin{cases} cx^3 & 0 \le x < 3 \\ 1 & x \ge 3 \\ 0 & x < 0 \end{cases}$$

If $P(X = 3) = 0$, find (a) the constant c, (b) the density function, (c) $P(X > 1)$, (d) $P(1 < X < 2)$.

2.52. Can the function

$$F(x) = \begin{cases} c(1 - x^2) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

be a distribution function? Explain.

2.53. Let X be a random variable having density function

$$f(x) = \begin{cases} cx & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) the value of the constant c, (b) $P(\frac{1}{2} < X < \frac{3}{2})$, (c) $P(X > 1)$, (d) the distribution function.

Joint distributions and independent variables

2.54. The joint probability function of two discrete random variables X and Y is given by $f(x, y) = cxy$ for $x = 1, 2, 3$ and $y = 1, 2, 3$, and equals zero otherwise. Find (a) the constant c, (b) $P(X = 2, Y = 3)$, (c) $P(1 \le X \le 2, Y \le 2)$, (d) $P(X \ge 2)$, (e) $P(Y < 2)$, (f) $P(X = 1)$, (g) $P(Y = 3)$.

2.55. Find the marginal probability functions of (a) X and (b) Y for the random variables of Problem 2.54. (c) Determine whether X and Y are independent.

2.56. Let X and Y be continuous random variables having joint density function

$$f(x, y) = \begin{cases} c(x^2 + y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Determine (a) the constant c, (b) $P(X < \frac{1}{2}, Y > \frac{1}{2})$, (c) $P(\frac{1}{4} < X < \frac{3}{4})$, (d) $P(Y < \frac{1}{2})$, (e) whether X and Y are independent.

2.57. Find the marginal distribution functions (a) of X and (b) of Y for the density function of Problem 2.56.

Conditional distributions and density functions

2.58. Find the conditional probability function (a) of X given Y, (b) of Y given X, for the distribution of Problem 2.54.

2.59. Let

$$f(x, y) = \begin{cases} x + y & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the conditional density function of (a) X given Y, (b) Y given X.

2.60. Find the conditional density of (a) X given Y, (b) Y given X, for the distribution of Problem 2.56.

2.61. Let

$$f(x, y) = \begin{cases} e^{-(x+y)} & x \ge 0, y \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

be the joint density function of X and Y. Find the conditional density function of (a) X given Y, (b) Y given X.

Change of variables

2.62. Let X have density function

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

Find the density function of $Y = X^2$.

2.63. (a) If the density function of X is $f(x)$ find the density function of X^3. (b) Illustrate the result in part (a) by choosing

$$f(x) = \begin{cases} 2e^{-2x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

and check the answer.

2.64. If X has density function $f(x) = 2(\pi)^{-1/2}e^{-x^2/2}$, $-\infty < x < \infty$, find the density function of $Y = X^2$.

2.65. Verify that the integral of $g_1(u)$ in Method 1 of Problem 2.21 is equal to 1.

2.66. If the density of X is $f(x) = 1/\pi(x^2 + 1)$, $-\infty < x < \infty$, find the density of $Y = \tan^{-1} X$.

2.67. Complete the work needed to find $g_1(u)$ in Method 2 of Problem 2.21 and check your answer.

2.68. Let the density of X be

$$f(x) = \begin{cases} 1/2 & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the density of (a) $3X - 2$, (b) $X^3 + 1$.

2.69. Check by direct integration the joint density function found in Problem 2.22.

2.70. Let X and Y have joint density function

$$f(x, y) = \begin{cases} e^{-(x+y)} & x \ge 0, y \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

If $U = X/Y$, $V = X + Y$, find the joint density function of U and V.

2.71. Use Problem 2.22 to find the density function of (a) $U = XY^2$, (b) $V = X^2Y$.

2.72. Let X and Y be random variables having joint density function $f(x, y) = (2\pi)^{-1} e^{-(x^2+y^2)}$, $-\infty < x < \infty$, $-\infty < y < \infty$. If R and Θ are new random variables such that $X = R \cos \Theta$, $Y = R \sin \Theta$, show that the density function of R is

$$g(r) = \begin{cases} re^{-r^2/2} & r \ge 0 \\ 0 & r < 0 \end{cases}$$

2.73. Let
$$f(x, y) = \begin{cases} 1 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$
be the joint density function of X and Y. Find the density function of $Z = XY$.

Convolutions

2.74. Let X and Y be identically distributed independent random variables with density function
$$f(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & \text{otherwise} \end{cases}$$
Find the density function of $X + Y$ and check your answer.

2.75. Let X and Y be identically distributed independent random variables with density function
$$f(t) = \begin{cases} e^{-t} & t \ge 0 \\ 0 & \text{otherwise} \end{cases}$$
Find the density function of $X + Y$ and check your answer.

2.76. Work Problem 2.21 by first making the transformation $2Y = Z$ and then using convolutions to find the density function of $U = X + Z$.

2.77. If the independent random variables X_1 and X_2 are identically distributed with density function
$$f(t) = \begin{cases} te^{-t} & t \ge 0 \\ 0 & t < 0 \end{cases}$$
find the density function of $X_1 + X_2$.

Applications to geometric probability

2.78. Two points are to be chosen at random on a line segment whose length is $a > 0$. Find the probability that the three line segments thus formed will be the sides of a triangle.

2.79. It is known that a bus will arrive at random at a certain location sometime between 3:00 P.M. and 3:30 P.M. A man decides that he will go at random to this location between these two times and will wait at most 5 minutes for the bus. If he misses it, he will take the subway. What is the probability that he will take the subway?

2.80. Two line segments, AB and CD, have lengths 8 and 6 units, respectively. Two points P and Q are to be chosen at random on AB and CD, respectively. Show that the probability that the area of a triangle will have height AP and that the base CQ will be greater than 12 square units is equal to $(1 - \ln 2)/2$.

Miscellaneous problems

2.81. Suppose that $f(x) = c/3^x$, $x = 1, 2, \ldots$, is the probability function for a random variable X. (a) Determine c. (b) Find the distribution function. (c) Graph the probability function and the distribution function. (d) Find $P(2 \le X < 5)$. (e) Find $P(X \ge 3)$.

2.82. Suppose that
$$f(x) = \begin{cases} cxe^{-2x} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$
is the density function for a random variable X. (a) Determine c. (b) Find the distribution function. (c) Graph the density function and the distribution function. (d) Find $P(X \ge 1)$. (e) Find $P(2 \le X < 3)$.

2.83. The probability function of a random variable X is given by

$$f(x) = \begin{cases} 2p & x = 1 \\ p & x = 2 \\ 4p & x = 3 \\ 0 & \text{otherwise} \end{cases}$$

where p is a constant. Find (a) $P(0 \le X < 3)$, (b) $P(X > 1)$.

2.84. (a) Prove that for a suitable constant c,

$$F(x) = \begin{cases} 0 & x \le 0 \\ c(1 - e^{-x})^2 & x > 0 \end{cases}$$

is the distribution function for a random variable X, and find this c. (b) Determine $P(1 < X < 2)$.

2.85. A random variable X has density function

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the density function of the random variable $Y = X^2$ and check your answer.

2.86. Two independent random variables, X and Y, have respective density functions

$$f(x) = \begin{cases} c_1 e^{-2x} & x > 0 \\ 0 & x \le 0 \end{cases} \qquad g(y) = \begin{cases} c_2 y e^{-3y} & y > 0 \\ 0 & y \le 0 \end{cases}$$

Find (a) c_1 and c_2, (b) $P(X + Y > 1)$, (c) $P(1 < X < 2, Y \ge 1)$, (d) $P(1 < X < 2)$, (e) $P(Y \ge 1)$.

2.87. In Problem 2.86 what is the relationship between the answers to (c), (d), and (e)? Justify your answer.

2.88. Let X and Y be random variables having joint density function

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

Find (a) the constant c, (b) $P(X > \frac{1}{2}, Y < \frac{3}{2})$, (c) the (marginal) density function of X, (d) the (marginal) density function of Y.

2.89. In Problem 2.88 is $P(X > \frac{1}{2}, Y < \frac{3}{2}) = P(X > \frac{1}{2})P(Y < \frac{3}{2})$? Why?

2.90. In Problem 2.86 find the density function (a) of X^2, (b) of $X + Y$.

2.91. Let X and Y have joint density function

$$f(x, y) = \begin{cases} 1/y & 0 < x < y, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Determine whether X and Y are independent, (b) Find $P(X > \frac{1}{2})$. (c) Find $P(X < \frac{1}{2}, Y > \frac{1}{3})$. (d) Find $P(X + Y > \frac{1}{2})$.

2.92. Generalize (a) Problem 2.74 and (b) Problem 2.75 to three or more variables.

2.93. Let X and Y be identically distributed independent random variables having density function $f(u) = (2\pi)^{-1/2}e^{-u^2/2}$, $-\infty < u < \infty$. Find the density function of $Z = X^2 + Y^2$.

2.94. The joint probability function for the random variables X and Y is given in Table 2-9. (a) Find the marginal probability functions of X and Y. (b) Find $P(1 \le X < 3, Y \ge 1)$. (c) Determine whether X and Y are independent.

Table 2-9

X \ Y	0	1	2
0	1/18	1/9	1/6
1	1/9	1/18	1/9
2	1/6	1/6	1/18

2.95. Suppose that the joint probability function of random variables X and Y is given by

$$f(x, y) = \begin{cases} cxy & 0 \le x \le 2, 0 \le y \le x \\ 0 & \text{otherwise} \end{cases}$$

(a) Determine whether X and Y are independent. (b) Find $P(\frac{1}{2} < X < 1)$. (c) Find $P(Y \ge 1)$. (d) Find $P(\frac{1}{2} < X < 1, Y \ge 1)$.

2.96. Let X and Y be independent random variables each having density function

$$f(u) = \frac{\lambda^u e^{-\lambda}}{u} \qquad u = 0, 1, 2, \ldots$$

where $\lambda > 0$. Prove that the density function of $X + Y$ is

$$g(u) = \frac{(2\lambda)^u e^{-2\lambda}}{u!} \qquad u = 0, 1, 2, \ldots$$

2.97. A stick of length L is to be broken into two parts. What is the probability that one part will have a length of more than double the other? State clearly what assumptions would you have made. Discuss whether you believe these assumptions are realistic and how you might improve them if they are not.

2.98. A floor is made up of squares of side l. A needle of length $a < l$ is to be tossed onto the floor. Prove that the probability of the needle intersecting at least one side is equal to $a(4l - a)/\pi l^2$.

2.99. For a needle of given length, what should be the side of a square in Problem 2.98 so that the probability of intersection is a maximum? Explain your answer.

2.100. Let

$$f(x, y, z) = \begin{cases} 24xy^2z^3 & 0 < x < 1, 0 < y < 1, 0 < z < 1 \\ 0 & \text{otherwise} \end{cases}$$

be the joint density function of three random variables X, Y, and Z. Find (a) $P(X > \frac{1}{2}, Y < \frac{1}{2}, Z > \frac{1}{2})$, (b) $P(Z < X + Y)$.

2.101. A cylindrical stream of particles, of radius a, is directed toward a hemispherical target ABC with center at O as indicated in Fig. 2-26. Assume that the distribution of particles is given by

$$f(r) = \begin{cases} 1/a & 0 < r < a \\ 0 & \text{otherwise} \end{cases}$$

where r is the distance from the axis OB. Show that the distribution of particles along the target is given by

$$g(\theta) = \begin{cases} \cos\theta & 0 < \theta < \pi/2 \\ 0 & \text{otherwise} \end{cases}$$

where θ is the angle that line OP (from O to any point P on the target) makes with the axis.

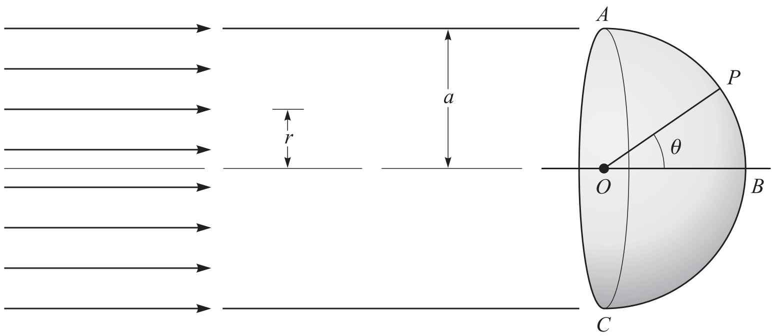


Fig. 2-26

2.102. In Problem 2.101 find the probability that a particle will hit the target between $\theta = 0$ and $\theta = \pi/4$.

2.103. Suppose that random variables X, Y, and Z have joint density function

$$f(x, y, z) = \begin{cases} 1 - \cos\pi x \cos\pi y \cos\pi z & 0 < x < 1, 0 < y < 1, 0 < z < 1 \\ 0 & \text{otherwise} \end{cases}$$

Show that although any two of these random variables are independent, i.e., their marginal density function factors, all three are not independent.

ANSWERS TO SUPPLEMENTARY PROBLEMS

2.38.

x	0	1	2	3
$f(x)$	1/8	3/8	3/8	1/8

2.39.

x	0	1	2
$f(x)$	3/28	15/28	5/14

2.40.

x	0	1	2
$f(x)$	9/64	15/32	25/64

2.42.

x	0	1	2	3	4
$f(x)$	$\frac{194{,}580}{270{,}725}$	$\frac{69{,}184}{270{,}725}$	$\frac{6768}{270{,}725}$	$\frac{192}{270{,}725}$	$\frac{1}{270{,}725}$

2.43.

x	0	1	2	3
$f(x)$	1/8	1/2	7/8	1

2.46. (a)

x	1	2	3	4
$f(x)$	1/8	1/4	3/8	1/4

(b) 3/4 (c) 7/8 (d) 3/8 (e) 7/8

2.47. (a) 3 (b) $e^{-3} - e^{-6}$ (c) e^{-9} (d) $1 - e^{-3}$ **2.48.** $F(x) = \begin{cases} 1 - e^{-3x} & x \geq 0 \\ 0 & x \leq 0 \end{cases}$

2.49. (a) 6/29 (b) 15/29 (c) 19/116 **2.50.** $F(x) = \begin{cases} 0 & x \leq 1 \\ (2x^3 - 2)/29 & 1 \leq x \leq 2 \\ (3x^2 + 2)/29 & 2 \leq x \leq 3 \\ 1 & x \geq 3 \end{cases}$

2.51. (a) 1/27 (b) $f(x) = \begin{cases} x^2/9 & 0 \leq x < 3 \\ 0 & \text{otherwise} \end{cases}$ (c) 26/27 (d) 7/27

2.53. (a) 1/2 (b) 1/2 (c) 3/4 (d) $F(x) = \begin{cases} 0 & x \leq 0 \\ x^2/4 & 0 \leq x \leq 2 \\ 1 & x \geq 2 \end{cases}$

2.54. (a) 1/36 (b) 1/6 (c) 1/4 (d) 5/6 (e) 1/6 (f) 1/6 (g) 1/2

2.55. (a) $f_1(x) = \begin{cases} x/6 & x = 1, 2, 3 \\ 0 & \text{other } x \end{cases}$ (b) $f_2(y) = \begin{cases} y/6 & y = 1, 2, 3 \\ 0 & \text{other } y \end{cases}$

2.56. (a) 3/2 (b) 1/4 (c) 29/64 (d) 5/16

2.57. (a) $F_1(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2}(x^3 + x) & 0 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$ (b) $F_2(y) = \begin{cases} 0 & y \leq 0 \\ \frac{1}{2}(y^3 + y) & 0 \leq y \leq 1 \\ 1 & y \geq 1 \end{cases}$

2.58. (a) $f(x \mid y) = f_1(x)$ for $y = 1, 2, 3$ (see Problem 2.55)

(b) $f(y \mid x) = f_2(y)$ for $x = 1, 2, 3$ (see Problem 2.55)

2.59. (a) $f(x|y) = \begin{cases} (x + y)/(y + \frac{1}{2}) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{other } x, 0 \leq y \leq 1 \end{cases}$

(b) $f(y|x) = \begin{cases} (x + y)/(x + \frac{1}{2}) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & 0 \leq x \leq 1, \text{other } y \end{cases}$

2.60. (a) $f(x|y) = \begin{cases} (x^2 + y^2)/(y^2 + \frac{1}{3}) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{other } x, 0 \leq y \leq 1 \end{cases}$

(b) $f(y|x) = \begin{cases} (x^2 + y^2)/(x^2 + \frac{1}{3}) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & 0 \leq x \leq 1, \text{other } y \end{cases}$

2.61. (a) $f(x|y) = \begin{cases} e^{-x} & x \geq 0, y \geq 0 \\ 0 & x < 0, y \geq 0 \end{cases}$ (b) $f(y|x) = \begin{cases} e^{-y} & x \geq 0, y \geq 0 \\ 0 & x \geq 0, y < 0 \end{cases}$

2.62. $e^{-\sqrt{y}}/2\sqrt{y}$ for $y > 0$; 0 otherwise **2.64.** $(2\pi)^{-1/2}y^{-1/2}e^{-y/2}$ for $y > 0$; 0 otherwise

2.66. $1/\pi$ for $-\pi/2 < y < \pi/2$; 0 otherwise

2.68. (a) $g(y) = \begin{cases} \frac{1}{6} & -5 < y < 1 \\ 0 & \text{otherwise} \end{cases}$ (b) $g(y) = \begin{cases} \frac{1}{6}(1 - y)^{-2/3} & 0 < y < 1 \\ \frac{1}{6}(y - 1)^{-2/3} & 1 < y < 2 \\ 0 & \text{otherwise} \end{cases}$

2.70. $ve^{-v}/(1 + u)^2$ for $u \geq 0$, $v \geq 0$; 0 otherwise

2.73. $g(z) = \begin{cases} -\ln z & 0 < z < 1 \\ 0 & \text{otherwise} \end{cases}$ **2.77.** $g(x) = \begin{cases} x^3e^{-x}/6 & x \geq 0 \\ 0 & x < 0 \end{cases}$

2.74. $g(u) = \begin{cases} u & 0 \leq u \leq 1 \\ 2 - u & 1 \leq u \leq 2 \\ 0 & \text{otherwise} \end{cases}$ **2.78.** $1/4$

2.75. $g(u) = \begin{cases} ue^{-u} & u \geq 0 \\ 0 & u < 0 \end{cases}$ **2.79.** $61/72$

2.81. (a) 2 (b) $F(x) = \begin{cases} 0 & x < 1 \\ 1 - 3^{-y} & y \leq x < y + 1;\ y = 1, 2, 3, \ldots \end{cases}$ (d) $26/81$ (e) $1/9$

2.82. (a) 4 (b) $F(x) = \begin{cases} 1 - e^{-2x}(2x + 1) & x \geq 0 \\ 0 & x < 0 \end{cases}$ (d) $3e^{-2}$ (e) $5e^{-4} - 7e^{-6}$

2.83. (a) $3/7$ (b) $5/7$ **2.84.** (a) $c = 1$ (b) $e^{-4} - 3e^{-2} + 2e^{-1}$

2.86. (a) $c_1 = 2, c_2 = 9$ (b) $9e^{-2} - 14e^{-3}$ (c) $4e^{-5} - 4e^{-7}$ (d) $e^{-2} - e^{-4}$ (e) $4e^{-3}$

2.88. (a) $1/4$ (b) $27/64$ (c) $f_1(x) = \begin{cases} x + \frac{1}{2} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$ (d) $f_2(y) = \begin{cases} \frac{1}{4}(y + 1) & 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$

2.90. (a) $\begin{cases} e^{-2y}/\sqrt{y} & y > 0 \\ 0 & \text{otherwise} \end{cases}$ (b) $\begin{cases} 18e^{-2u} & u > 0 \\ 0 & \text{otherwise} \end{cases}$

2.91. (b) $\frac{1}{2}(1 - \ln 2)$ (c) $\frac{1}{6} + \frac{1}{2}\ln 2$ (d) $\frac{1}{2}\ln 2$ **2.95.** (b) $15/256$ (c) $9/16$ (d) 0

2.93. $g(z) = \begin{cases} \frac{1}{2}e^{-z/2} & z \geq 0 \\ 0 & z < 0 \end{cases}$ **2.100.** (a) $45/512$ (b) $1/14$

2.94. (b) $7/18$ **2.102.** $\sqrt{2}/2$

CHAPTER 3

Mathematical Expectation

Definition of Mathematical Expectation

A very important concept in probability and statistics is that of the *mathematical expectation*, *expected value*, or briefly the *expectation*, of a random variable. For a discrete random variable X having the possible values $x_1, \ldots, x_n$, the expectation of X is defined as

$$E(X) = x_1P(X = x_1) + \cdots + x_nP(X = x_n) = \sum_{j=1}^{n} x_jP(X = x_j) \tag{1}$$

or equivalently, if $P(X = x_j) = f(x_j)$,

$$E(X) = x_1f(x_1) + \cdots + x_nf(x_n) = \sum_{j=1}^{n} x_jf(x_j) = \sum xf(x) \tag{2}$$

where the last summation is taken over all appropriate values of x. As a special case of (2), where the probabilities are all equal, we have

$$E(X) = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3}$$

which is called the *arithmetic mean*, or simply the *mean*, of $x_1, x_2, \ldots, x_n$.

If X takes on an infinite number of values $x_1, x_2, \ldots$, then $E(X) = \sum_{j=1}^{\infty} x_jf(x_j)$ provided that the infinite series converges absolutely.

For a continuous random variable X having density function $f(x)$, the expectation of X is defined as

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx \tag{4}$$

provided that the integral converges absolutely.

The expectation of X is very often called the *mean* of X and is denoted by μ_X, or simply μ, when the particular random variable is understood.

The mean, or expectation, of X gives a single value that acts as a representative or average of the values of X, and for this reason it is often called a *measure of central tendency*. Other measures are considered on page 83.

EXAMPLE 3.1 Suppose that a game is to be played with a single die assumed fair. In this game a player wins \$20 if a 2 turns up, \$40 if a 4 turns up; loses \$30 if a 6 turns up; while the player neither wins nor loses if any other face turns up. Find the expected sum of money to be won.

Let X be the random variable giving the amount of money won on any toss. The possible amounts won when the die turns up 1, 2, . . . , 6 are $x_1, x_2, \ldots, x_6$, respectively, while the probabilities of these are $f(x_1), f(x_2), \ldots, f(x_6)$. The probability function for X is displayed in Table 3-1. Therefore, the expected value or expectation is

$$E(X) = (0)\left(\frac{1}{6}\right) + (20)\left(\frac{1}{6}\right) + (0)\left(\frac{1}{6}\right) + (40)\left(\frac{1}{6}\right) + (0)\left(\frac{1}{6}\right) + (-30)\left(\frac{1}{6}\right) = 5$$

Table 3-1

x_j	0	+20	0	+40	0	−30
$f(x_j)$	1/6	1/6	1/6	1/6	1/6	1/6

It follows that the player can expect to win \$5. In a fair game, therefore, the player should be expected to pay \$5 in order to play the game.

EXAMPLE 3.2 The density function of a random variable X is given by

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

The expected value of X is then

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^2 x\left(\frac{1}{2}x\right)dx = \int_0^2 \frac{x^2}{2}\,dx = \frac{x^3}{6}\bigg|_0^2 = \frac{4}{3}$$

Functions of Random Variables

Let X be a discrete random variable with probability function $f(x)$. Then $Y = g(X)$ is also a discrete random variable, and the probability function of Y is

$$h(y) = P(Y = y) = \sum_{\{x|g(x)=y\}} P(X = x) = \sum_{\{x|g(x)=y\}} f(x)$$

If X takes on the values $x_1, x_2, \ldots, x_n$, and Y the values $y_1, y_2, \ldots, y_m$ $(m \le n)$, then $y_1h(y_1) + y_2h(y_2) + \cdots + y_mh(y_m) = g(x_1)f(x_1) + g(x_2)f(x_2) + \cdots + g(x_n)f(x_n)$. Therefore,

$$\begin{aligned} E[g(X)] &= g(x_1)f(x_1) + g(x_2)f(x_2) + \cdots + g(x_n)f(x_n) \\ &= \sum_{j=1}^{n} g(x_j)f(x_j) = \sum g(x)f(x) \end{aligned} \tag{5}$$

Similarly, if X is a continuous random variable having probability density $f(x)$, then it can be shown that

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx \tag{6}$$

Note that (5) and (6) do not involve, respectively, the probability function and the probability density function of $Y = g(X)$.

Generalizations are easily made to functions of two or more random variables. For example, if X and Y are two continuous random variables having joint density function $f(x, y)$, then the expectation of $g(X, Y)$ is given by

$$E[g(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f(x, y)\,dx\,dy \tag{7}$$

EXAMPLE 3.3 If X is the random variable of Example 3.2,

$$E(3X^2 - 2X) = \int_{-\infty}^{\infty} (3x^2 - 2x)f(x)\,dx = \int_0^2 (3x^2 - 2x)\left(\frac{1}{2}x\right)dx = \frac{10}{3}$$

Some Theorems on Expectation

Theorem 3-1 If c is any constant, then

$$E(cX) = cE(X) \tag{8}$$

Theorem 3-2 If X and Y are any random variables, then

$$E(X + Y) = E(X) + E(Y) \tag{9}$$

Theorem 3-3 If X and Y are independent random variables, then

$$E(XY) = E(X)E(Y) \tag{10}$$

Generalizations of these theorems are easily made.

The Variance and Standard Deviation

We have already noted on page 75 that the expectation of a random variable X is often called the *mean* and is denoted by μ. Another quantity of great importance in probability and statistics is called the *variance* and is defined by

$$\operatorname{Var}(X) = E[(X - \mu)^2] \tag{11}$$

The variance is a nonnegative number. The positive square root of the variance is called the *standard deviation* and is given by

$$\sigma_X = \sqrt{\operatorname{Var}(X)} = \sqrt{E[(X - \mu)^2]} \tag{12}$$

Where no confusion can result, the standard deviation is often denoted by σ instead of σ_X, and the variance in such case is σ^2.

If X is a discrete random variable taking the values $x_1, x_2, \ldots, x_n$ and having probability function $f(x)$, then the variance is given by

$$\sigma_X^2 = E[(X - \mu)^2] = \sum_{j=1}^{n}(x_j - \mu)^2 f(x_j) = \sum (x - \mu)^2 f(x) \tag{13}$$

In the special case of (13) where the probabilities are all equal, we have

$$\sigma^2 = [(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2]/n \tag{14}$$

which is the variance for a set of n numbers $x_1, \ldots, x_n$.

If X takes on an infinite number of values $x_1, x_2, \ldots$, then $\sigma_X^2 = \sum_{j=1}^{\infty}(x_j - \mu)^2 f(x_j)$, provided that the series converges.

If X is a continuous random variable having density function $f(x)$, then the variance is given by

$$\sigma_X^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \tag{15}$$

provided that the integral converges.

The variance (or the standard deviation) is a measure of the *dispersion*, or *scatter*, of the values of the random variable about the mean μ. If the values tend to be concentrated near the mean, the variance is small; while if the values tend to be distributed far from the mean, the variance is large. The situation is indicated graphically in Fig. 3-1 for the case of two continuous distributions having the same mean μ.

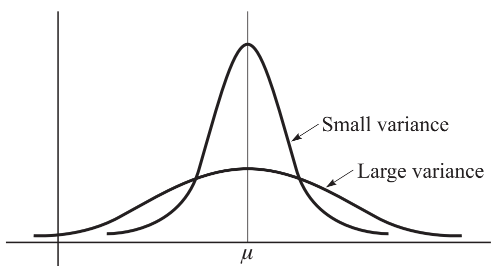


Fig. 3-1

EXAMPLE 3.4 Find the variance and standard deviation of the random variable of Example 3.2. As found in Example 3.2, the mean is $\mu = E(X) = 4/3$. Then the variance is given by

$$\sigma^2 = E\left[\left(X - \frac{4}{3}\right)^2\right] = \int_{-\infty}^{\infty}\left(x - \frac{4}{3}\right)^2 f(x)\,dx = \int_0^2\left(x - \frac{4}{3}\right)^2\left(\frac{1}{2}x\right)dx = \frac{2}{9}$$

and so the standard deviation is $\sigma = \sqrt{\frac{2}{9}} = \frac{\sqrt{2}}{3}$

Note that if X has certain *dimensions* or *units*, such as *centimeters* (cm), then the variance of X has units cm^2 while the standard deviation has the same unit as X, i.e., cm. It is for this reason that the standard deviation is often used.

Some Theorems on Variance

Theorem 3-4

$$\sigma^2 = E[(X - \mu)^2] = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2 \tag{16}$$

where $\mu = E(X)$.

Theorem 3-5 If c is any constant,

$$\text{Var}(cX) = c^2\,\text{Var}(X) \tag{17}$$

Theorem 3-6 The quantity $E[(X - a)^2]$ is a minimum when $a = \mu = E(X)$.

Theorem 3-7 If X and Y are independent random variables,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \quad \text{or} \quad \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y \tag{18}$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) \quad \text{or} \quad \sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y \tag{19}$$

Generalizations of Theorem 3-7 to more than two independent variables are easily made. In words, the variance of a sum of independent variables equals the sum of their variances.

Standardized Random Variables

Let X be a random variable with mean μ and standard deviation $\sigma\,(\sigma > 0)$. Then we can define an associated *standardized random variable* given by

$$X^* = \frac{X - \mu}{\sigma} \tag{20}$$

An important property of X^* is that it has a mean of zero and a variance of 1, which accounts for the name *standardized,* i.e.,

$$E(X^*) = 0, \qquad \text{Var}(X^*) = 1 \tag{21}$$

The values of a standardized variable are sometimes called *standard scores,* and X is then said to be expressed in *standard units* (i.e., σ is taken as the unit in measuring $X - \mu$).

Standardized variables are useful for comparing different distributions.

Moments

The *rth moment of a random variable X about the mean* μ, also called the *rth central moment*, is defined as

$$\mu_r = E[(X - \mu)^r] \tag{22}$$

where $r = 0, 1, 2, \ldots$. It follows that $\mu_0 = 1$, $\mu_1 = 0$, and $\mu_2 = \sigma^2$, i.e., the second central moment or second moment about the mean is the variance. We have, assuming absolute convergence,

$$\mu_r = \sum (x - \mu)^r f(x) \qquad \text{(discrete variable)} \tag{23}$$

$$\mu_r = \int_{-\infty}^{\infty} (x - \mu)^r f(x)\, dx \qquad \text{(continuous variable)} \tag{24}$$

The *rth moment of X about the origin*, also called the *rth raw moment*, is defined as

$$\mu'_r = E(X^r) \tag{25}$$

where $r = 0, 1, 2, \ldots$, and in this case there are formulas analogous to (23) and (24) in which $\mu = 0$.

The relationship between these moments is given by

$$\mu_r = \mu'_r - \binom{r}{1}\mu'_{r-1}\mu + \cdots + (-1)^j \binom{r}{j}\mu'_{r-j}\mu^j + \cdots + (-1)^r \mu'_0 \mu^r \tag{26}$$

As special cases we have, using $\mu'_1 = \mu$ and $\mu'_0 = 1$,

$$\begin{aligned} \mu_2 &= \mu'_2 - \mu^2 \\ \mu_3 &= \mu'_3 - 3\mu'_2\mu + 2\mu^3 \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu + 6\mu'_2\mu^2 - 3\mu^4 \end{aligned} \tag{27}$$

Moment Generating Functions

The *moment generating function* of X is defined by

$$M_X(t) = E(e^{tX}) \tag{28}$$

that is, assuming convergence,

$$M_X(t) = \sum e^{tx} f(x) \qquad \text{(discrete variable)} \tag{29}$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \qquad \text{(continuous variable)} \tag{30}$$

We can show that the Taylor series expansion is [Problem 3.15(a)]

$$M_X(t) = 1 + \mu t + \mu'_2 \frac{t^2}{2!} + \cdots + \mu'_r \frac{t^r}{r!} + \cdots \tag{31}$$

Since the coefficients in this expansion enable us to find the moments, the reason for the name *moment generating function* is apparent. From the expansion we can show that [Problem 3.15(b)]

$$\mu'_r = \left. \frac{d^r}{dt^r} M_X(t) \right|_{t=0} \tag{32}$$

i.e., μ'_r is the rth derivative of $M_X(t)$ evaluated at $t = 0$. Where no confusion can result, we often write $M(t)$ instead of $M_X(t)$.

Some Theorems on Moment Generating Functions

Theorem 3-8 If $M_X(t)$ is the moment generating function of the random variable X and a and b ($b \neq 0$) are constants, then the moment generating function of $(X + a)/b$ is

$$M_{(X+a)/b}(t) = e^{at/b} M_X\left(\frac{t}{b}\right) \tag{33}$$

Theorem 3-9 If X and Y are independent random variables having moment generating functions $M_X(t)$ and $M_Y(t)$, respectively, then

$$M_{X+Y}(t) = M_X(t)\, M_Y(t) \tag{34}$$

Generalizations of Theorem 3-9 to more than two independent random variables are easily made. In words, the moment generating function of a sum of independent random variables is equal to the product of their moment generating functions.

Theorem 3-10 **(Uniqueness Theorem)** Suppose that X and Y are random variables having moment generating functions $M_X(t)$ and $M_Y(t)$, respectively. Then X and Y have the same probability distribution if and only if $M_X(t) = M_Y(t)$ identically.

Characteristic Functions

If we let $t = i\omega$, where i is the imaginary unit, in the moment generating function we obtain an important function called the *characteristic function.* We denote this by

$$\phi_X(\omega) = M_X(i\omega) = E(e^{i\omega X}) \tag{35}$$

It follows that

$$\phi_X(\omega) = \sum e^{i\omega x} f(x) \qquad \text{(discrete variable)} \tag{36}$$

$$\phi_X(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} f(x)\, dx \qquad \text{(continuous variable)} \tag{37}$$

Since $|e^{i\omega x}| = 1$, the series and the integral always converge absolutely.

The corresponding results (31) and (32) become

$$\phi_X(\omega) = 1 + i\mu\omega - \mu_2' \frac{\omega^2}{2!} + \cdots + i^r \mu_r' \frac{\omega^r}{r!} + \cdots \tag{38}$$

where

$$\mu_r' = (-1)^r i^r \frac{d^r}{d\omega^r} \phi_X(\omega)\bigg|_{\omega=0} \tag{39}$$

When no confusion can result, we often write $\phi(\omega)$ instead of $\phi_X(\omega)$.

Theorems for characteristic functions corresponding to Theorems 3-8, 3-9, and 3-10 are as follows.

Theorem 3-11 If $\phi_X(\omega)$ is the characteristic function of the random variable X and a and b ($b \neq 0$) are constants, then the characteristic function of $(X + a)/b$ is

$$\phi_{(X+a)/b}(\omega) = e^{ai\omega/b} \phi_X\left(\frac{\omega}{b}\right) \tag{40}$$

Theorem 3-12 If X and Y are independent random variables having characteristic functions $\phi_X(\omega)$ and $\phi_Y(\omega)$, respectively, then

$$\phi_{X+Y}(\omega) = \phi_X(\omega)\, \phi_Y(\omega) \tag{41}$$

More generally, the characteristic function of a sum of independent random variables is equal to the product of their characteristic functions.

Theorem 3-13 **(Uniqueness Theorem)** Suppose that X and Y are random variables having characteristic functions $\phi_X(\omega)$ and $\phi_Y(\omega)$, respectively. Then X and Y have the same probability distribution if and only if $\phi_X(\omega) = \phi_Y(\omega)$ identically.

An important reason for introducing the characteristic function is that (37) represents the *Fourier transform* of the density function $f(x)$. From the theory of Fourier transforms, we can easily determine the density function from the characteristic function. In fact,

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega x}\phi_X(\omega)\,d\omega \tag{42}$$

which is often called an *inversion formula*, or *inverse Fourier transform.* In a similar manner we can show in the discrete case that the probability function $f(x)$ can be obtained from (36) by use of *Fourier series*, which is the analog of the Fourier integral for the discrete case. See Problem 3.39.

Another reason for using the characteristic function is that it always exists whereas the moment generating function may not exist.

Variance for Joint Distributions. Covariance

The results given above for one variable can be extended to two or more variables. For example, if X and Y are two continuous random variables having joint density function $f(x, y)$, the means, or expectations, of X and Y are

$$\mu_X = E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x, y)\,dx\,dy, \qquad \mu_Y = E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x, y)\,dx\,dy \tag{43}$$

and the variances are

$$\begin{aligned} \sigma_X^2 &= E[(X - \mu_X)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)^2 f(x, y)\,dx\,dy \\ \sigma_Y^2 &= E[(Y - \mu_Y)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y - \mu_Y)^2 f(x, y)\,dx\,dy \end{aligned} \tag{44}$$

Note that the marginal density functions of X and Y are not directly involved in (43) and (44).

Another quantity that arises in the case of two variables X and Y is the *covariance* defined by

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{45}$$

In terms of the joint density function $f(x, y)$, we have

$$\sigma_{XY} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x, y)\,dx\,dy \tag{46}$$

Similar remarks can be made for two discrete random variables. In such cases (43) and (46) are replaced by

$$\mu_X = \sum_x\sum_y xf(x, y) \qquad \mu_Y = \sum_x\sum_y yf(x, y) \tag{47}$$

$$\sigma_{XY} = \sum_x\sum_y (x - \mu_X)(y - \mu_Y)f(x, y) \tag{48}$$

where the sums are taken over all the discrete values of X and Y.

The following are some important theorems on covariance.

Theorem 3-14
$$\sigma_{XY} = E(XY) - E(X)E(Y) = E(XY) - \mu_X\mu_Y \tag{49}$$

Theorem 3-15 If X and Y are independent random variables, then

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = 0 \tag{50}$$

Theorem 3-16
$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\,\mathrm{Cov}(X, Y) \tag{51}$$

or

$$\sigma_{X\pm Y}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2\sigma_{XY} \tag{52}$$

Theorem 3-17
$$|\sigma_{XY}| \le \sigma_X\sigma_Y \tag{53}$$

The converse of Theorem 3-15 is not necessarily true. If X and Y are independent, Theorem 3-16 reduces to Theorem 3-7.

Correlation Coefficient

If X and Y are independent, then $\operatorname{Cov}(X, Y) = \sigma_{XY} = 0$. On the other hand, if X and Y are completely dependent, for example, when $X = Y$, then $\operatorname{Cov}(X, Y) = \sigma_{XY} = \sigma_X \sigma_Y$. From this we are led to a *measure of the dependence* of the variables X and Y given by

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{54}$$

We call ρ the *correlation coefficient*, or *coefficient of correlation.* From Theorem 3-17 we see that $-1 \le \rho \le 1$. In the case where $\rho = 0$ (i.e., the covariance is zero), we call the variables X and Y *uncorrelated.* In such cases, however, the variables may or may not be independent. Further discussion of correlation cases will be given in Chapter 8.

Conditional Expectation, Variance, and Moments

If X and Y have joint density function $f(x, y)$, then as we have seen in Chapter 2, the conditional density function of Y given X is $f(y \mid x) = f(x, y)/f_1(x)$ where $f_1(x)$ is the marginal density function of X. We can define the *conditional expectation*, or *conditional mean*, of Y given X by

$$E(Y \mid X = x) = \int_{-\infty}^{\infty} y f(y \mid x)\, dy \tag{55}$$

where "$X = x$" is to be interpreted as $x < X \le x + dx$ in the continuous case. Theorems 3-1 and 3-2 also hold for conditional expectation.

We note the following properties:

1. $E(Y \mid X = x) = E(Y)$ when X and Y are independent.
2. $E(Y) = \int_{-\infty}^{\infty} E(Y \mid X = x) f_1(x)\, dx$.

It is often convenient to calculate expectations by use of Property 2, rather than directly.

EXAMPLE 3.5 The average travel time to a distant city is c hours by car or b hours by bus. A woman cannot decide whether to drive or take the bus, so she tosses a coin. What is her expected travel time?

Here we are dealing with the joint distribution of the outcome of the toss, X, and the travel time, Y, where $Y = Y_{\text{car}}$ if $X = 0$ and $Y = Y_{\text{bus}}$ if $X = 1$. Presumably, both Y_{car} and Y_{bus} are independent of X, so that by Property 1 above

$$E(Y \mid X = 0) = E(Y_{\text{car}} \mid X = 0) = E(Y_{\text{car}}) = c$$

and

$$E(Y \mid X = 1) = E(Y_{\text{bus}} \mid X = 1) = E(Y_{\text{bus}}) = b$$

Then Property 2 (with the integral replaced by a sum) gives, for a fair coin,

$$E(Y) = E(Y \mid X = 0)P(X = 0) + E(Y \mid X = 1)P(X = 1) = \frac{c + b}{2}$$

In a similar manner we can define the *conditional variance* of Y given X as

$$E[(Y - \mu_2)^2 \mid X = x] = \int_{-\infty}^{\infty} (y - \mu_2)^2 f(y \mid x)\, dy \tag{56}$$

where $\mu_2 = E(Y \mid X = x)$. Also we can define the rth *conditional moment* of Y about any value a given X as

$$E[(Y - a)^r \mid X = x] = \int_{-\infty}^{\infty} (y - a)^r f(y \mid x)\, dy \tag{57}$$

The usual theorems for variance and moments extend to conditional variance and moments.

Chebyshev's Inequality

An important theorem in probability and statistics that reveals a general property of discrete or continuous random variables having finite mean and variance is known under the name of *Chebyshev's inequality.*

Theorem 3-18 **(Chebyshev's Inequality)** Suppose that X is a random variable (discrete or continuous) having mean μ and variance σ^2, which are finite. Then if ϵ is any positive number,

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \tag{58}$$

or, with $\epsilon = k\sigma$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{59}$$

EXAMPLE 3.6 Letting $k = 2$ in Chebyshev's inequality (59), we see that

$$P(|X - \mu| \geq 2\sigma) \leq 0.25 \qquad \text{or} \qquad P(|X - \mu| < 2\sigma) \geq 0.75$$

In words, the probability of X differing from its mean by more than 2 standard deviations is less than or equal to 0.25; equivalently, the probability that X will lie within 2 standard deviations of its mean is greater than or equal to 0.75. This is quite remarkable in view of the fact that we have not even specified the probability distribution of X.

Law of Large Numbers

The following theorem, called the *law of large numbers*, is an interesting consequence of Chebyshev's inequality.

Theorem 3-19 **(Law of Large Numbers):** Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables (discrete or continuous), each having finite mean μ and variance σ^2. Then if $S_n = X_1 + X_2 + \cdots + X_n (n = 1, 2, \ldots)$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0 \tag{60}$$

Since S_n/n is the arithmetic mean of $X_1, \ldots, X_n$, this theorem states that the probability of the arithmetic mean S_n/n differing from its expected value μ by more than ϵ approaches zero as $n \to \infty$. A stronger result, which we might expect to be true, is that $\lim_{n\to\infty} S_n/n = \mu$, but this is actually false. However, we can prove that $\lim_{n\to\infty} S_n/n = \mu$ *with probability one.* This result is often called the *strong law of large numbers*, and, by contrast, that of Theorem 3-19 is called the *weak law of large numbers.* When the "law of large numbers" is referred to without qualification, the weak law is implied.

Other Measures of Central Tendency

As we have already seen, the mean, or expectation, of a random variable X provides a measure of central tendency for the values of a distribution. Although the mean is used most, two other measures of central tendency are also employed. These are the *mode* and the *median.*

1. **MODE.** The *mode* of a discrete random variable is that value which occurs most often or, in other words, has the greatest probability of occurring. Sometimes we have two, three, or more values that have relatively large probabilities of occurrence. In such cases, we say that the distribution is *bimodal*, *trimodal*, or *multimodal*, respectively. The mode of a continuous random variable X is the value (or values) of X where the probability density function has a relative maximum.
2. **MEDIAN.** The *median* is that value x for which $P(X < x) \leq \frac{1}{2}$ and $P(X > x) \leq \frac{1}{2}$. In the case of a continuous distribution we have $P(X < x) = \frac{1}{2} = P(X > x)$, and the median separates the density curve into two parts having equal areas of $1/2$ each. In the case of a discrete distribution a unique median may not exist (see Problem 3.34).

Percentiles

It is often convenient to subdivide the area under a density curve by use of ordinates so that the area to the left of the ordinate is some percentage of the total unit area. The values corresponding to such areas are called *percentile values*, or briefly *percentiles.* Thus, for example, the area to the left of the ordinate at x_α in Fig. 3-2 is α. For instance, the area to the left of $x_{0.10}$ would be 0.10, or 10%, and $x_{0.10}$ would be called the *10th percentile* (also called the *first decile*). The median would be the *50th percentile* (or *fifth decile*).

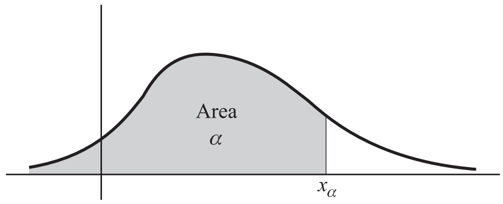


Fig. 3-2

Other Measures of Dispersion

Just as there are various measures of central tendency besides the mean, there are various measures of dispersion or scatter of a random variable besides the variance or standard deviation. Some of the most common are the following.

1. **SEMI-INTERQUARTILE RANGE.** If $x_{0.25}$ and $x_{0.75}$ represent the 25th and 75th percentile values, the difference $x_{0.75} - x_{0.25}$ is called the *interquartile range* and $\frac{1}{2}(x_{0.75} - x_{0.25})$ is the *semi-interquartile range.*
2. **MEAN DEVIATION.** The *mean deviation* (M.D.) of a random variable X is defined as the expectation of $|X - \mu|$, i.e., assuming convergence,

$$\text{M.D.}(X) = E[|X - \mu|] = \sum |x - \mu| f(x) \qquad \text{(discrete variable)} \tag{61}$$

$$\text{M.D.}(X) = E[|X - \mu|] = \int_{-\infty}^{\infty} |x - \mu| f(x)\, dx \qquad \text{(continuous variable)} \tag{62}$$

Skewness and Kurtosis

1. **SKEWNESS.** Often a distribution is not symmetric about any value but instead has one of its tails longer than the other. If the longer tail occurs to the right, as in Fig. 3-3, the distribution is said to be *skewed to the right*, while if the longer tail occurs to the left, as in Fig. 3-4, it is said to be *skewed to the left*. Measures describing this asymmetry are called *coefficients of skewness*, or briefly *skewness*. One such measure is given by

$$\alpha_3 = \frac{E[(X - \mu)^3]}{\sigma^3} = \frac{\mu_3}{\sigma^3} \tag{63}$$

The measure σ_3 will be positive or negative according to whether the distribution is skewed to the right or left, respectively. For a symmetric distribution, $\sigma_3 = 0$.

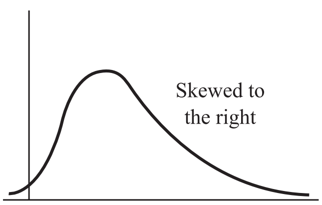


Fig. 3-3

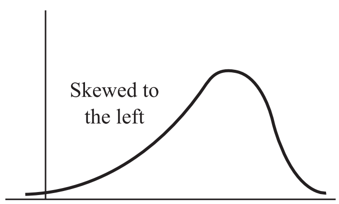


Fig. 3-4

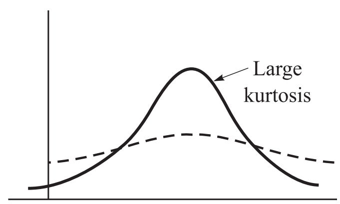


Fig. 3-5

2. **KURTOSIS.** In some cases a distribution may have its values concentrated near the mean so that the distribution has a large peak as indicated by the solid curve of Fig. 3-5. In other cases the distribution may be

relatively flat as in the dashed curve of Fig. 3-5. Measures of the degree of peakedness of a distribution are called *coefficients of kurtosis*, or briefly *kurtosis*. A measure often used is given by

$$\alpha_4 = \frac{E[(X - \mu)^4]}{\sigma^4} = \frac{\mu_4}{\sigma^4} \tag{64}$$

This is usually compared with the normal curve (see Chapter 4), which has a coefficient of kurtosis equal to 3. See also Problem 3.41.

SOLVED PROBLEMS

Expectation of random variables

3.1. In a lottery there are 200 prizes of \$5, 20 prizes of \$25, and 5 prizes of \$100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to pay for a ticket?

Let X be a random variable denoting the amount of money to be won on a ticket. The various values of X together with their probabilities are shown in Table 3-2. For example, the probability of getting one of the 20 tickets giving a \$25 prize is $20/10{,}000 = 0.002$. The expectation of X in dollars is thus

$$E(X) = (5)(0.02) + (25)(0.002) + (100)(0.0005) + (0)(0.9775) = 0.2$$

or 20 cents. Thus the fair price to pay for a ticket is 20 cents. However, since a lottery is usually designed to raise money, the price per ticket would be higher.

Table 3-2

x (dollars)	5	25	100	0
$P(X = x)$	0.02	0.002	0.0005	0.9775

3.2. Find the expectation of the sum of points in tossing a pair of fair dice.

Let X and Y be the points showing on the two dice. We have

$$E(X) = E(Y) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right) = \frac{7}{2}$$

Then, by Theorem 3-2,

$$E(X + Y) = E(X) + E(Y) = 7$$

3.3. Find the expectation of a discrete random variable X whose probability function is given by

$$f(x) = \left(\frac{1}{2}\right)^x \qquad (x = 1, 2, 3, \ldots)$$

We have

$$E(X) = \sum_{x=1}^{\infty} x\left(\frac{1}{2}\right)^x = \frac{1}{2} + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + \cdots$$

To find this sum, let $$S = \frac{1}{2} + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{16}\right) + \cdots$$

Then $$\frac{1}{2}S = \frac{1}{4} + 2\left(\frac{1}{8}\right) + 3\left(\frac{1}{16}\right) + \cdots$$

Subtracting, $$\frac{1}{2}S = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots = 1$$

Therefore, $S = 2$.

3.4. A continuous random variable X has probability density given by

$$f(x) = \begin{cases} 2e^{-2x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Find (a) $E(X)$, (b) $E(X^2)$.

(a)
$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{0}^{\infty} x(2e^{-2x})\,dx = 2\int_{0}^{\infty} xe^{-2x}\,dx$$
$$= 2\left[(x)\left(\frac{e^{-2x}}{-2}\right) - (1)\left(\frac{e^{-2x}}{4}\right)\right]\Bigg|_0^{\infty} = \frac{1}{2}$$

(b)
$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = 2\int_{0}^{\infty} x^2 e^{-2x}\,dx$$
$$= 2\left[(x^2)\left(\frac{e^{-2x}}{-2}\right) - (2x)\left(\frac{e^{-2x}}{4}\right) + (2)\left(\frac{e^{-2x}}{-8}\right)\right]\Bigg|_0^{\infty} = \frac{1}{2}$$

3.5. The joint density function of two random variables X and Y is given by

$$f(x, y) = \begin{cases} xy/96 & 0 < x < 4,\ 1 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $E(X)$, (b) $E(Y)$, (c) $E(XY)$, (d) $E(2X + 3Y)$.

(a)
$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x, y)\,dx\,dy = \int_{x=0}^{4}\int_{y=1}^{5} x\left(\frac{xy}{96}\right)dx\,dy = \frac{8}{3}$$

(b)
$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x, y)\,dx\,dy = \int_{x=0}^{4}\int_{y=1}^{5} y\left(\frac{xy}{96}\right)dx\,dy = \frac{31}{9}$$

(c)
$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy)f(x, y)\,dx\,dy = \int_{x=0}^{4}\int_{y=1}^{5} (xy)\left(\frac{xy}{96}\right)dx\,dy = \frac{248}{27}$$

(d)
$$E(2X + 3Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (2x + 3y)f(x, y)\,dx\,dy = \int_{x=0}^{4}\int_{y=1}^{5} (2x + 3y)\left(\frac{xy}{96}\right)dx\,dy = \frac{47}{3}$$

Another method

(c) Since X and Y are independent, we have, using parts (a) and (b),

$$E(XY) = E(X)E(Y) = \left(\frac{8}{3}\right)\left(\frac{31}{9}\right) = \frac{248}{27}$$

(d) By Theorems 3-1 and 3-2, pages 76–77, together with (a) and (b),

$$E(2X + 3Y) = 2E(X) + 3E(Y) = 2\left(\frac{8}{3}\right) + 3\left(\frac{31}{9}\right) = \frac{47}{3}$$

3.6. Prove Theorem 3-2, page 77.

Let $f(x, y)$ be the joint probability function of X and Y, assumed discrete. Then

$$\begin{aligned} E(X + Y) &= \sum_x \sum_y (x + y) f(x, y) \\ &= \sum_x \sum_y xf(x, y) + \sum_x \sum_y yf(x, y) \\ &= E(X) + E(Y) \end{aligned}$$

If either variable is continuous, the proof goes through as before, with the appropriate summations replaced by integrations. Note that the theorem is true whether or not X and Y are independent.

3.7. Prove Theorem 3-3, page 77.

Let $f(x, y)$ be the joint probability function of X and Y, assumed discrete. If the variables X and Y are independent, we have $f(x, y) = f_1(x) f_2(y)$. Therefore,

$$\begin{aligned} E(XY) &= \sum_x \sum_y xyf(x, y) = \sum_x \sum_y xyf_1(x) f_2(y) \\ &= \sum_x \left[xf_1(x) \sum_y yf_2(y) \right] \\ &= \sum_x [(xf_1(x)E(y)] \\ &= E(X)E(Y) \end{aligned}$$

If either variable is continuous, the proof goes through as before, with the appropriate summations replaced by integrations. Note that the validity of this theorem hinges on whether $f(x, y)$ can be expressed as a function of x multiplied by a function of y, for all x and y, i.e., on whether X and Y are independent. For dependent variables it is not true in general.

Variance and standard deviation

3.8. Find (a) the variance, (b) the standard deviation of the sum obtained in tossing a pair of fair dice.

(a) Referring to Problem 3.2, we have $E(X) = E(Y) = 1/2$. Moreover,

$$E(X^2) = E(Y^2) = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + \cdots + 6^2\left(\frac{1}{6}\right) = \frac{91}{6}$$

Then, by Theorem 3-4,

$$\mathrm{Var}(X) = \mathrm{Var}(Y) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

and, since X and Y are independent, Theorem 3-7 gives

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = \frac{35}{6}$$

(b)
$$\sigma_{X+Y} = \sqrt{\mathrm{Var}(X + Y)} = \sqrt{\frac{35}{6}}$$

3.9. Find (a) the variance, (b) the standard deviation for the random variable of Problem 3.4.

(a) As in Problem 3.4, the mean of X is $\mu = E(X) = \frac{1}{2}$. Then the variance is

$$\mathrm{Var}(X) = E[(X - \mu)^2] = E\left[\left(X - \frac{1}{2}\right)^2\right] = \int_{-\infty}^{\infty} \left(x - \frac{1}{2}\right)^2 f(x)\,dx$$

$$= \int_0^{\infty} \left(x - \frac{1}{2}\right)^2 (2e^{-2x})\,dx = \frac{1}{4}$$

Another method

By Theorem 3-4,

$$\mathrm{Var}(X) = E[(X - \mu)^2] = E(X^2) - [E(X)]^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

(b)
$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{4}} = \frac{1}{2}$$

3.10. Prove Theorem 3-4, page 78.

We have

$$\begin{aligned} E[(X-\mu)^2] &= E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

3.11. Prove Theorem 3-6, page 78.

$$\begin{aligned} E[(X-a)^2] &= E[\{(X-\mu) + (\mu - a)\}^2] \\ &= E[(X-\mu)^2 + 2(X-\mu)(\mu-a) + (\mu-a)^2] \\ &= E[(X-\mu)^2] + 2(\mu-a)E(X-\mu) + (\mu-a)^2 \\ &= E[(X-\mu)^2] + (\mu-a)^2 \end{aligned}$$

since $E(X-\mu) = E(X) - \mu = 0$. From this we see that the minimum value of $E[(X-a)^2]$ occurs when $(\mu-a)^2 = 0$, i.e., when $a = \mu$.

3.12. If $X^* = (X-\mu)/\sigma$ is a standardized random variable, prove that (a) $E(X^*) = 0$, (b) $\text{Var}(X^*) = 1$.

(a)
$$E(X^*) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}[E(X-\mu)] = \frac{1}{\sigma}[E(X) - \mu] = 0$$

since $E(X) = \mu$.

(b)
$$\text{Var}(X^*) = \text{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}E[(X-\mu)^2] = 1$$

using Theorem 3-5, page 78, and the fact that $E[(X-\mu)^2] = \sigma^2$.

3.13. Prove Theorem 3-7, page 78.

$$\begin{aligned} \text{Var}(X+Y) &= E[\{(X+Y) - (\mu_X + \mu_Y)\}^2] \\ &= E[\{(X-\mu_X) + (Y-\mu_Y)\}^2] \\ &= E[(X-\mu_X)^2 + 2(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)^2] \\ &= E[(X-\mu_X)^2] + 2E[(X-\mu_X)(Y-\mu_Y)] + E[(Y-\mu_Y)^2] \\ &= \text{Var}(X) + \text{Var}(Y) \end{aligned}$$

using the fact that

$$E[(X-\mu_X)(Y-\mu_Y)] = E(X-\mu_X)E(Y-\mu_Y) = 0$$

since X and Y, and therefore $X - \mu_X$ and $Y - \mu_Y$, are independent. The proof of (*19*), page 78, follows on replacing Y by $-Y$ and using Theorem 3-5.

Moments and moment generating functions

3.14. Prove the result (26), page 79.

$$\begin{aligned} \mu_r &= E[(X-\mu)^r] \\ &= E\left[X^r - \binom{r}{1}X^{r-1}\mu + \cdots + (-1)^j\binom{r}{j}X^{r-j}\mu^j \right. \\ &\qquad \left. + \cdots + (-1)^{r-1}\binom{r}{r-1}X\mu^{r-1} + (-1)^r\mu^r\right] \end{aligned}$$

$$= E(X^r) - \binom{r}{1}E(X^{r-1})\mu + \cdots + (-1)^j\binom{r}{j}E(X^{r-j})\mu^j$$

$$+ \cdots + (-1)^{r-1}\binom{r}{r-1}E(X)\mu^{r-1} + (-1)^r\mu^r$$

$$= \mu'_r - \binom{r}{1}\mu'_{r-1}\mu + \cdots + (-1)^j\binom{r}{j}\mu'_{r-j}\mu^j$$

$$+ \cdots + (-1)^{r-1}r\mu^r + (-1)^{-r}\mu^r$$

where the last two terms can be combined to give $(-1)^{r-1}(r-1)\mu^r$.

3.15. Prove (a) result (31), (b) result (32), page 79.

(a) Using the power series expansion for e^u (3., Appendix A), we have

$$M_X(t) = E(e^{tX}) = E\left(1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots\right)$$

$$= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots$$

$$= 1 + \mu t + \mu'_2\frac{t^2}{2!} + \mu'_3\frac{t^3}{3!} + \cdots$$

(b) This follows immediately from the fact known from calculus that if the Taylor series of $f(t)$ about $t = a$ is

$$f(t) = \sum_{n=0}^{\infty} c_n(t-a)^n$$

then

$$c_n = \frac{1}{n!}\frac{d^n}{dt^n}f(t)\bigg|_{t=a}$$

3.16. Prove Theorem 3-9, page 80.

Since X and Y are independent, any function of X and any function of Y are independent. Hence,

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t)$$

3.17. The random variable X can assume the values 1 and -1 with probability $\frac{1}{2}$ each. Find (a) the moment generating function, (b) the first four moments about the origin.

(a)
$$E(e^{tX}) = e^{t(1)}\left(\frac{1}{2}\right) + e^{t(-1)}\left(\frac{1}{2}\right) = \frac{1}{2}(e^t + e^{-t})$$

(b) We have
$$e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots$$

$$e^{-t} = 1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \frac{t^4}{4!} - \cdots$$

Then (1)
$$\frac{1}{2}(e^t + e^{-t}) = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots$$

But (2)
$$M_X(t) = 1 + \mu t + \mu'_2\frac{t^2}{2!} + \mu'_3\frac{t^3}{3!} + \mu'_4\frac{t^4}{4!} + \cdots$$

Then, comparing (1) and (2), we have

$$\mu = 0, \qquad \mu'_2 = 1, \qquad \mu'_3 = 0, \qquad \mu'_4 = 1, \ldots$$

The odd moments are all zero, and the even moments are all one.

3.18. A random variable X has density function given by

$$f(x) = \begin{cases} 2e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Find (a) the moment generating function, (b) the first four moments about the origin.

(a)
$$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx$$
$$= \int_0^{\infty} e^{tx}(2e^{-2x})\,dx = 2\int_0^{\infty} e^{(t-2)x}\,dx$$
$$= \frac{2e^{(t-2)x}}{t-2}\bigg|_0^{\infty} = \frac{2}{2-t}, \quad \text{assuming } t < 2$$

(b) If $|t| < 2$ we have

$$\frac{2}{2-t} = \frac{1}{1-t/2} = 1 + \frac{t}{2} + \frac{t^2}{4} + \frac{t^3}{8} + \frac{t^4}{16} + \cdots$$

But
$$M(t) = 1 + \mu t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \mu_4' \frac{t^4}{4!} + \cdots$$

Therefore, on comparing terms, $\mu = \frac{1}{2}, \mu_2' = \frac{1}{2}, \mu_3' = \frac{3}{4}, \mu_4' = \frac{3}{2}$.

3.19. Find the first four moments (a) about the origin, (b) about the mean, for a random variable X having density function

$$f(x) = \begin{cases} 4x(9 - x^2)/81 & 0 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

(a)
$$\mu_1' = E(X) = \frac{4}{81}\int_0^3 x^2(9 - x^2)\,dx = \frac{8}{5} = \mu$$
$$\mu_2' = E(X^2) = \frac{4}{81}\int_0^3 x^3(9 - x^2)\,dx = 3$$
$$\mu_3' = E(X^3) = \frac{4}{81}\int_0^3 x^4(9 - x^2)\,dx = \frac{216}{35}$$
$$\mu_4' = E(X^4) = \frac{4}{81}\int_0^3 x^5(9 - x^2)\,dx = \frac{27}{2}$$

(b) Using the result (27), page 79, we have

$$\mu_1 = 0$$
$$\mu_2 = 3 - \left(\frac{8}{5}\right)^2 = \frac{11}{25} = \sigma^2$$
$$\mu_3 = \frac{216}{35} - 3(3)\left(\frac{8}{5}\right) + 2\left(\frac{8}{5}\right)^3 = -\frac{32}{875}$$
$$\mu_4 = \frac{27}{2} - 4\left(\frac{216}{35}\right)\left(\frac{8}{5}\right) + 6(3)\left(\frac{8}{5}\right)^2 - 3\left(\frac{8}{5}\right)^4 = \frac{3693}{8750}$$

Characteristic functions

3.20. Find the characteristic function of the random variable X of Problem 3.17.

The characteristic function is given by

$$E(e^{i\omega X}) = e^{i\omega(1)}\left(\frac{1}{2}\right) + e^{i\omega(-1)}\left(\frac{1}{2}\right) = \frac{1}{2}(e^{i\omega} + e^{-i\omega}) = \cos\omega$$

using Euler's formulas,

$$e^{i\theta} = \cos\theta + i\sin\theta \qquad e^{-i\theta} = \cos\theta - i\sin\theta$$

with $\theta = \omega$. The result can also be obtained from Problem 3.17(a) on putting $t = i\omega$.

3.21. Find the characteristic function of the random variable X having density function given by

$$f(x) = \begin{cases} 1/2a & |x| < a \\ 0 & \text{otherwise} \end{cases}$$

The characteristic function is given by

$$E(e^{i\omega X}) = \int_{-\infty}^{\infty} e^{i\omega x} f(x)\,dx = \frac{1}{2a}\int_{-a}^{a} e^{i\omega x}\,dx$$
$$= \frac{1}{2a}\frac{e^{i\omega x}}{i\omega}\bigg|_{-a}^{a} = \frac{e^{ia\omega} - e^{-ia\omega}}{2ia\omega} = \frac{\sin a\omega}{a\omega}$$

using Euler's formulas (see Problem 3.20) with $\theta = a\omega$.

3.22. Find the characteristic function of the random variable X having density function $f(x) = ce^{-a|x|}$, $-\infty < x < \infty$, where $a > 0$, and c is a suitable constant.

Since $f(x)$ is a density function, we must have

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

so that

$$c\int_{-\infty}^{\infty} e^{-a|x|}\,dx = c\left[\int_{-\infty}^{0} e^{-a(-x)}\,dx + \int_{0}^{\infty} e^{-a(x)}\,dx\right]$$
$$= c\frac{e^{ax}}{a}\bigg|_{-\infty}^{0} + c\frac{e^{-ax}}{-a}\bigg|_{0}^{\infty} = \frac{2c}{a} = 1$$

Then $c = a/2$. The characteristic function is therefore given by

$$E(e^{i\omega X}) = \int_{-\infty}^{\infty} e^{i\omega x} f(x)\,dx$$
$$= \frac{a}{2}\left[\int_{-\infty}^{0} e^{i\omega x} e^{-a(-x)}\,dx + \int_{0}^{\infty} e^{i\omega x} e^{-a(x)}\,dx\right]$$
$$= \frac{a}{2}\left[\int_{-\infty}^{0} e^{(a+i\omega)x}\,dx + \int_{0}^{\infty} e^{-(a-i\omega)x}\,dx\right]$$
$$= \frac{a}{2}\frac{e^{(a+i\omega)x}}{a + i\omega}\bigg|_{-\infty}^{0} + a\frac{e^{-(a-i\omega)x}}{-(a - i\omega)}\bigg|_{0}^{\infty}$$
$$= \frac{a}{2(a + i\omega)} + \frac{a}{2(a - i\omega)} = \frac{a^2}{a^2 + \omega^2}$$

Covariance and correlation coefficient

3.23. Prove Theorem 3-14, page 81.

By definition the covariance of X and Y is

$$\begin{aligned}
\sigma_{XY} &= \mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \\
&= E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\
&= E(XY) - \mu_X E(Y) - \mu_Y E(X) + E(\mu_X\mu_Y) \\
&= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

3.24. Prove Theorem 3-15, page 81.

If X and Y are independent, then $E(XY) = E(X)E(Y)$. Therefore, by Problem 3.23,

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0$$

3.25. Find (a) $E(X)$, (b) $E(Y)$, (c) $E(XY)$, (d) $E(X^2)$, (e) $E(Y^2)$, (f) Var (X), (g) Var (Y), (h) Cov (X, Y), (i) ρ, if the random variables X and Y are defined as in Problem 2.8, pages 47–48.

(a)
$$E(X) = \sum_x \sum_y x f(x, y) = \sum_x x\left[\sum_y f(x, y)\right]$$
$$= (0)(6c) + (1)(14c) + (2)(22c) = 58c = \frac{58}{42} = \frac{29}{21}$$

(b)
$$E(Y) = \sum_x \sum_y y f(x, y) = \sum_y y\left[\sum_x f(x, y)\right]$$
$$= (0)(6c) + (1)(9c) + (2)(12c) + (3)(15c) = 78c = \frac{78}{42} = \frac{13}{7}$$

(c)
$$E(XY) = \sum_x \sum_y xy f(x, y)$$
$$\begin{aligned} &= (0)(0)(0) + (0)(1)(c) + (0)(2)(2c) + (0)(3)(3c) \\ &\quad + (1)(0)(2c) + (1)(1)(3c) + (1)(2)(4c) + (1)(3)(5c) \\ &\quad + (2)(0)(4c) + (2)(1)(5c) + (2)(2)(6c) + (2)(3)(7c) \\ &= 102c = \frac{102}{42} = \frac{17}{7} \end{aligned}$$

(d)
$$E(X^2) = \sum_x \sum_y x^2 f(x, y) = \sum_x x^2\left[\sum_y f(x, y)\right]$$
$$= (0)^2(6c) + (1)^2(14c) + (2)^2(22c) = 102c = \frac{102}{42} = \frac{17}{7}$$

(e)
$$E(Y^2) = \sum_x \sum_y y^2 f(x, y) = \sum_y y^2\left[\sum_x f(x, y)\right]$$
$$= (0)^2(6c) + (1)^2(9c) + (2)^2(12c) + (3)^2(15c) = 192c = \frac{192}{42} = \frac{32}{7}$$

(f)
$$\sigma_X^2 = \mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{17}{7} - \left(\frac{29}{21}\right)^2 = \frac{230}{441}$$

(g)
$$\sigma_Y^2 = \mathrm{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{32}{7} - \left(\frac{13}{7}\right)^2 = \frac{55}{49}$$

(h)
$$\sigma_{XY} = \mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{17}{7} - \left(\frac{29}{21}\right)\left(\frac{13}{7}\right) = -\frac{20}{147}$$

(i)
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-20/147}{\sqrt{230/441}\sqrt{55/49}} = \frac{-20}{\sqrt{230}\sqrt{55}} = -0.2103 \text{ approx.}$$

3.26. Work Problem 3.25 if the random variables X and Y are defined as in Problem 2.33, pages 61–63.

Using $c = 1/210$, we have:

(a)
$$E(X) = \frac{1}{210}\int_{x=2}^{6}\int_{y=0}^{5} (x)(2x + y)\,dx\,dy = \frac{268}{63}$$

(b)
$$E(Y) = \frac{1}{210}\int_{x=2}^{6}\int_{y=0}^{5} (y)(2x + y)\,dx\,dy = \frac{170}{63}$$

(c)
$$E(XY) = \frac{1}{210}\int_{x=2}^{6}\int_{y=0}^{5} (xy)(2x + y)\,dx\,dy = \frac{80}{7}$$

(d) $$E(X^2) = \frac{1}{210}\int_{x=2}^{6}\int_{y=0}^{5}(x^2)(2x + y)\,dx\,dy = \frac{1220}{63}$$

(e) $$E(Y^2) = \frac{1}{210}\int_{x=2}^{6}\int_{y=0}^{5}(y^2)(2x + y)\,dx\,dy = \frac{1175}{126}$$

(f) $$\sigma_X^2 = \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1220}{63} - \left(\frac{268}{63}\right)^2 = \frac{5036}{3969}$$

(g) $$\sigma_Y^2 = \text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{1175}{126} - \left(\frac{170}{63}\right)^2 = \frac{16{,}225}{7938}$$

(h) $$\sigma_{XY} = \text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{80}{7} - \left(\frac{268}{63}\right)\left(\frac{170}{63}\right) = -\frac{200}{3969}$$

(i) $$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{-200/3969}{\sqrt{5036/3969}\sqrt{16{,}225/7938}} = \frac{-200}{\sqrt{2518}\sqrt{16{,}225}} = -0.03129 \text{ approx.}$$

Conditional expectation, variance, and moments

3.27. Find the conditional expectation of Y given $X = 2$ in Problem 2.8, pages 47–48.

As in Problem 2.27, page 58, the conditional probability function of Y given $X = 2$ is

$$f(y\,|\,2) = \frac{4 + y}{22}$$

Then the conditional expectation of Y given $X = 2$ is

$$E(Y\,|\,X = 2) = \sum_y y\left(\frac{4 + y}{22}\right)$$

where the sum is taken over all y corresponding to $X = 2$. This is given by

$$E(Y\,|\,X = 2) = (0)\left(\frac{4}{22}\right) + 1\left(\frac{5}{22}\right) + 2\left(\frac{6}{22}\right) + 3\left(\frac{7}{22}\right) = \frac{19}{11}$$

3.28. Find the conditional expectation of (a) Y given X, (b) X given Y in Problem 2.29, pages 58–59.

(a) $$E(Y\,|\,X = x)\int_{-\infty}^{\infty} yf_2(y\,|\,x)\,dy = \int_0^x y\left(\frac{2y}{x^2}\right)dy = \frac{2x}{3}$$

(b) $$E(X\,|\,Y = y) = \int_{-\infty}^{\infty} xf_1(x\,|\,y)\,dx = \int_y^1 x\left(\frac{2x}{1 - y^2}\right)dx$$

$$= \frac{2(1 - y^3)}{3(1 - y^2)} = \frac{2(1 + y + y^2)}{3(1 + y)}$$

3.29. Find the conditional variance of Y given X for Problem 2.29, pages 58–59.

The required variance (second moment about the mean) is given by

$$E[(Y - \mu_2)^2\,|\,X = x] = \int_{-\infty}^{\infty}(y - \mu_2)^2 f_2(y\,|\,x)\,dy = \int_0^x\left(y - \frac{2x}{3}\right)^2\left(\frac{2y}{x^2}\right)dy = \frac{x^2}{18}$$

where we have used the fact that $\mu_2 = E(Y\,|\,X = x) = 2x/3$ from Problem 3.28(a).

Chebyshev's inequality

3.30. Prove Chebyshev's inequality.

We shall present the proof for continuous random variables. A proof for discrete variables is similar if integrals are replaced by sums. If $f(x)$ is the density function of X, then

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty}(x - \mu)^2 f(x)\,dx$$

Since the integrand is nonnegative, the value of the integral can only decrease when the range of integration is diminished. Therefore,

$$\sigma^2 \geq \int_{|x-\mu| \geq \epsilon} (x - \mu)^2 f(x)\,dx \geq \int_{|x-\mu| \geq \epsilon} \epsilon^2 f(x)\,dx = \epsilon^2 \int_{|x-\mu| \geq \epsilon} f(x)\,dx$$

But the last integral is equal to $P(|X - \mu| \geq \epsilon)$. Hence,

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

3.31. For the random variable of Problem 3.18, (a) find $P(|X - \mu| > 1)$. (b) Use Chebyshev's inequality to obtain an upper bound on $P(|X - \mu| > 1)$ and compare with the result in (a).

(a) From Problem 3.18, $\mu = 1/2$. Then

$$P(|X - \mu| < 1) = P\left(\left|X - \frac{1}{2}\right| < 1\right) = P\left(-\frac{1}{2} < X < \frac{3}{2}\right)$$

$$= \int_0^{3/2} 2e^{-2x}\,dx = 1 - e^{-3}$$

Therefore $$P\left(\left|X - \frac{1}{2}\right| \geq 1\right) = 1 - (1 - e^{-3}) = e^{-3} = 0.04979$$

(b) From Problem 3.18, $\sigma^2 = \mu_2' - \mu^2 = 1/4$. Chebyshev's inequality with $\epsilon = 1$ then gives

$$P(|X - \mu| \geq 1) \leq \sigma^2 = 0.25$$

Comparing with (a), we see that the bound furnished by Chebyshev's inequality is here quite crude. In practice, Chebyshev's inequality is used to provide estimates when it is inconvenient or impossible to obtain exact values.

Law of large numbers

3.32. Prove the law of large numbers stated in Theorem 3-19, page 83.

We have $$E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$$

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \cdots = \mathrm{Var}(X_n) = \sigma^2$$

Then $$E\left(\frac{S_n}{n}\right) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}[E(X_1) + \cdots + E(X_n)] = \frac{1}{n}(n\mu) = \mu$$

$$\mathrm{Var}(S_n) = \mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\sigma^2$$

so that $$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{\sigma^2}{n}$$

where we have used Theorem 3-5 and an extension of Theorem 3-7.

Therefore, by Chebyshev's inequality with $X = S_n/n$, we have

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

Taking the limit as $n \to \infty$, this becomes, as required,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

Other measures of central tendency

3.33. The density function of a continuous random variable X is

$$f(x) = \begin{cases} 4x(9 - x^2)/81 & 0 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the mode. (b) Find the median. (c) Compare mode, median, and mean.

(a) The mode is obtained by finding where the density $f(x)$ has a relative maximum. The relative maxima of $f(x)$ occur where the derivative is zero, i.e.,

$$\frac{d}{dx}\left[\frac{4x(9-x^2)}{81}\right] = \frac{36-12x^2}{81} = 0$$

Then $x = \sqrt{3} = 1.73$ approx., which is the required mode. Note that this does give the maximum since the second derivative, $-24x/81$, is negative for $x = \sqrt{3}$.

(b) The median is that value a for which $P(X \le a) = 1/2$. Now, for $0 < a < 3$,

$$P(X \le a) = \frac{4}{81}\int_0^a x(9-x^2)\,dx = \frac{4}{81}\left(\frac{9a^2}{2} - \frac{a^4}{4}\right)$$

Setting this equal to $1/2$, we find that

$$2a^4 - 36a^2 + 81 = 0$$

from which

$$a^2 = \frac{36 \pm \sqrt{(36)^2 - 4(2)(81)}}{2(2)} = \frac{36 \pm \sqrt{648}}{4} = 9 \pm \frac{9}{2}\sqrt{2}$$

Therefore, the required median, which must lie between 0 and 3, is given by

$$a^2 = 9 - \frac{9}{2}\sqrt{2}$$

from which $a = 1.62$ approx.

(c)
$$E(X) = \frac{4}{81}\int_0^3 x^2(9-x^2)\,dx = \frac{4}{81}\left(3x^3 - \frac{x^5}{5}\right)\Bigg|_0^3 = 1.60$$

which is practically equal to the median. The mode, median, and mean are shown in Fig. 3-6.

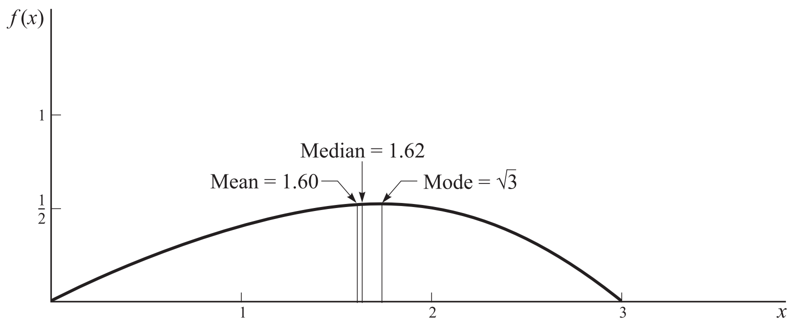


Fig. 3-6

3.34. A discrete random variable has probability function $f(x) = 1/2^x$ where $x = 1, 2, \ldots$. Find (a) the mode, (b) the median, and (c) compare them with the mean.

(a) The mode is the value x having largest associated probability. In this case it is $x = 1$, for which the probability is $1/2$.

(b) If x is any value between 1 and 2, $P(X < x) = \frac{1}{2}$ and $P(X > x) = \frac{1}{2}$. Therefore, *any number* between 1 and 2 could represent the median. For convenience, we choose the midpoint of the interval, i.e., $3/2$.

(c) As found in Problem 3.3, $\mu = 2$. Therefore, the ordering of the three measures is just the reverse of that in Problem 3.33.

Percentiles

3.35. Determine the (a) 10th, (b) 25th, (c) 75th percentile values for the distribution of Problem 3.33.

From Problem 3.33(b) we have

$$P(X \leq a) = \frac{4}{81}\left(\frac{9a^2}{2} - \frac{a^4}{4}\right) = \frac{18a^2 - a^4}{81}$$

(a) The 10th percentile is the value of a for which $P(X \leq a) = 0.10$, i.e., the solution of $(18a^2 - a^4)/81 = 0.10$. Using the method of Problem 3.33, we find $a = 0.68$ approx.

(b) The 25th percentile is the value of a such that $(18a^2 - a^4)/81 = 0.25$, and we find $a = 1.098$ approx.

(c) The 75th percentile is the value of a such that $(18a^2 - a^4)/81 = 0.75$, and we find $a = 2.121$ approx.

Other measures of dispersion

3.36. Determine, (a) the semi-interquartile range, (b) the mean deviation for the distribution of Problem 3.33.

(a) By Problem 3.35 the 25th and 75th percentile values are 1.098 and 2.121, respectively. Therefore,

$$\text{Semi-interquartile range} = \frac{2.121 - 1.098}{2} = 0.51 \text{ approx.}$$

(b) From Problem 3.33 the mean is $\mu = 1.60 = 8/5$. Then

$$\begin{aligned}
\text{Mean deviation} = \text{M.D.} = E(|X - \mu|) &= \int_{-\infty}^{\infty} |x - \mu| f(x)\,dx \\
&= \int_0^3 \left|x - \frac{8}{5}\right| \left[\frac{4x}{81}(9 - x^2)\right] dx \\
&= \int_0^{8/5} \left(\frac{8}{5} - x\right)\left[\frac{4x}{81}(9 - x^2)\right] dx + \int_{8/5}^{3} \left(x - \frac{8}{5}\right)\left[\frac{4x}{81}(9 - x^2)\right] dx \\
&= 0.555 \text{ approx.}
\end{aligned}$$

Skewness and kurtosis

3.37. Find the coefficient of (a) skewness, (b) kurtosis for the distribution of Problem 3.19.

From Problem 3.19(b) we have

$$\sigma^2 = \frac{11}{25} \qquad \mu_3 = -\frac{32}{875} \qquad \mu_4 = \frac{3693}{8750}$$

(a) Coefficient of skewness $= \alpha_3 = \dfrac{\mu_3}{\sigma^3} = -0.1253$

(b) Coefficient of kurtosis $= \alpha_4 = \dfrac{\mu_4}{\sigma^4} = 2.172$

It follows that there is a moderate skewness to the left, as is indicated in Fig. 3-6. Also the distribution is somewhat less peaked than the normal distribution, which has a kurtosis of 3.

Miscellaneous problems

3.38. If $M(t)$ is the moment generating function for a random variable X, prove that the mean is $\mu = M'(0)$ and the variance is $\sigma^2 = M''(0) - [M'(0)]^2$.

From (32), page 79, we have on letting $r = 1$ and $r = 2$,

$$\mu_1' = M'(0) \qquad \mu_2' = M''(0)$$

Then from (27)

$$\mu = M'(0) \qquad \mu_2 = \sigma^2 = M''(0) - [M'(0)]^2$$

3.39. Let X be a random variable that takes on the values $x_k = k$ with probabilities p_k where $k = \pm 1, \ldots, \pm n$. (a) Find the characteristic function $\phi(\omega)$ of X, (b) obtain p_k in terms of $\phi(\omega)$.

(a) The characteristic function is

$$\phi(\omega) = E(e^{i\omega X}) = \sum_{k=-n}^{n} e^{i\omega x_k} p_k = \sum_{k=-n}^{n} p_k e^{ik\omega}$$

(b) Multiply both sides of the expression in (a) by $e^{-ij\omega}$ and integrate with respect to ω from 0 to 2π. Then

$$\int_{\omega=0}^{2\pi} e^{-ij\omega}\phi(\omega)\,d\omega = \sum_{k=-n}^{n} p_k \int_{\omega=0}^{2\pi} e^{i(k-j)\omega}\,d\omega = 2\pi p_j$$

since

$$\int_{\omega=0}^{2\pi} e^{i(k-j)\omega}\,d\omega = \begin{cases} \left.\dfrac{e^{i(k-j)\omega}}{i(k-j)}\right|_0^{2\pi} = 0 & k \neq j \\ 2\pi & k = j \end{cases}$$

Therefore,

$$p_j = \frac{1}{2\pi}\int_{\omega=0}^{2\pi} e^{-ij\omega}\phi(\omega)\,d\omega$$

or, replacing j by k,

$$p_k = \frac{1}{2\pi}\int_{\omega=0}^{2\pi} e^{-ik\omega}\phi(\omega)\,d\omega$$

We often call $\sum_{k=-n}^{n} p_k e^{ik\omega}$ (where n can theoretically be infinite) the *Fourier series* of $\phi(\omega)$ and p_k the *Fourier coefficients*. For a continuous random variable, the Fourier series is replaced by the Fourier integral (see page 81).

3.40. Use Problem 3.39 to obtain the probability distribution of a random variable X whose characteristic function is $\phi(\omega) = \cos\omega$.

From Problem 3.39

$$\begin{aligned} p_k &= \frac{1}{2\pi}\int_{\omega=0}^{2\pi} e^{-ik\omega}\cos\omega\,d\omega \\ &= \frac{1}{2\pi}\int_{\omega=0}^{2\pi} e^{-ik\omega}\left[\frac{e^{i\omega}+e^{-i\omega}}{2}\right]d\omega \\ &= \frac{1}{4\pi}\int_{\omega=0}^{2\pi} e^{i(1-k)\omega}\,d\omega + \frac{1}{4\pi}\int_{\omega=0}^{2\pi} e^{-i(1+k)\omega}\,d\omega \end{aligned}$$

If $k = 1$, we find $p_1 = \frac{1}{2}$; if $k = -1$, we find $p_{-1} = \frac{1}{2}$. For all other values of k, we have $p_k = 0$. Therefore, the random variable is given by

$$X = \begin{cases} 1 & \text{probability } 1/2 \\ -1 & \text{probability } 1/2 \end{cases}$$

As a check, see Problem 3.20.

3.41. Find the coefficient of (a) skewness, (b) kurtosis of the distribution defined by the *normal curve*, having density

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \qquad -\infty < x < \infty$$

(a) The distribution has the appearance of Fig. 3-7. By symmetry, $\mu_1' = \mu = 0$ and $\mu_3' = 0$. Therefore the coefficient of skewness is zero.

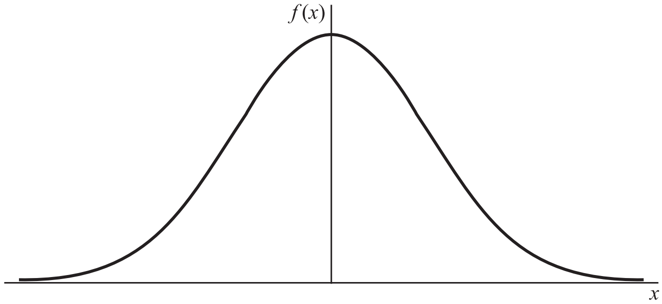


Fig. 3-7

(b) We have

$$\mu_2' = E(X^2) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} x^2 e^{-x^2/2}\,dx$$
$$= \frac{2}{\sqrt{\pi}}\int_0^{\infty} v^{1/2} e^{-v}\,dv$$
$$= \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = \frac{2}{\sqrt{\pi}}\cdot\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = 1$$

where we have made the transformation $x^2/2 = v$ and used properties of the gamma function given in (2) and (5) of Appendix A. Similarly we obtain

$$\mu_4' = E(X^4) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4 e^{-x^2/2}\,dx = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} x^4 e^{-x^2/2}\,dx$$
$$= \frac{4}{\sqrt{\pi}}\int_0^{\infty} v^{3/2} e^{-v}\,dv$$
$$= \frac{4}{\sqrt{\pi}}\Gamma\left(\frac{5}{2}\right) = \frac{4}{\sqrt{\pi}}\cdot\frac{3}{2}\cdot\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = 3$$

Now

$$\sigma^2 = E[(X - \mu)^2] = E(X)^2 = \mu_2' = 1$$
$$\mu_4 = E[(X - \mu)^4] = E(X^4) = \mu_4' = 3$$

Thus the coefficient of kurtosis is

$$\frac{\mu_4}{\sigma^4} = 3$$

3.42. Prove that $-1 \le \rho \le 1$ (see page 82).

For any real constant c, we have

$$E[\{Y - \mu_Y - c(X - \mu)\}^2] \ge 0$$

Now the left side can be written

$$E[(Y - \mu_Y)^2] + c^2E[(X - \mu_X)^2] - 2cE[(X - \mu_X)(Y - \mu_Y)] = \sigma_Y^2 + c^2\sigma_X^2 - 2c\sigma_{XY}$$
$$= \sigma_Y^2 + \sigma_X^2\left(c^2 - \frac{2c\sigma_{XY}}{\sigma_X^2}\right)$$
$$= \sigma_Y^2 + \sigma_X^2\left(c^2 - \frac{\sigma_{XY}}{\sigma_X^2}\right)^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$$
$$= \frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2} + \sigma_X^2\left(c - \frac{\sigma_{XY}}{\sigma_X^2}\right)^2$$

In order for this last quantity to be greater than or equal to zero for every value of c, we must have

$$\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2 \geq 0 \quad \text{or} \quad \frac{\sigma_{XY}^2}{\sigma_X^2\sigma_Y^2} \leq 1$$

which is equivalent to $\rho^2 \leq 1$ or $-1 \leq \rho \leq 1$.

SUPPLEMENTARY PROBLEMS

Expectation of random variables

3.43. A random variable X is defined by $X = \begin{cases} -2 & \text{prob. } 1/3 \\ 3 & \text{prob. } 1/2. \\ 1 & \text{prob. } 1/6 \end{cases}$ Find (a) $E(X)$, (b) $E(2X + 5)$, (c) $E(X^2)$.

3.44. Let X be a random variable defined by the density function $f(x) = \begin{cases} 3x^2 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$.

Find (a) $E(X)$, (b) $E(3X - 2)$, (c) $E(X^2)$.

3.45. The density function of a random variable X is $f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$.

Find (a) $E(X)$, (b) $E(X^2)$, (c) $E[(X - 1)^2]$.

3.46. What is the expected number of points that will come up in 3 successive tosses of a fair die? Does your answer seem reasonable? Explain.

3.47. A random variable X has the density function $f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$. Find $E(e^{2X/3})$.

3.48. Let X and Y be independent random variables each having density function

$$f(u) = \begin{cases} 2e^{-2u} & u \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $E(X + Y)$, (b) $E(X^2 + Y^2)$, (c) $E(XY)$.

3.49. Does (a) $E(X + Y) = E(X) + E(Y)$, (b) $E(XY) = E(X)E(Y)$, in Problem 3.48? Explain.

3.50. Let X and Y be random variables having joint density function

$$f(x, y) = \begin{cases} \frac{3}{5}x(x + y) & 0 \leq x \leq 1, 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $E(X)$, (b) $E(Y)$, (c) $E(X + Y)$, (d) $E(XY)$.

3.51. Does (a) $E(X + Y) = E(X) + E(Y)$, (b) $E(XY) = E(X)E(Y)$, in Problem 3.50? Explain.

3.52. Let X and Y be random variables having joint density

$$f(x, y) = \begin{cases} 4xy & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $E(X)$, (b) $E(Y)$, (c) $E(X + Y)$, (d) $E(XY)$.

3.53. Does (a) $E(X + Y) = E(X) + E(Y)$, (b) $E(XY) = E(X)E(Y)$, in Problem 3.52? Explain.

3.54. Let $f(x, y) = \begin{cases} \frac{1}{4}(2x + y) & 0 \le x \le 1, 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases}$. Find (a) $E(X)$, (b) $E(Y)$, (c) $E(X^2)$, (d) $E(Y^2)$, (e) $E(X + Y)$, (f) $E(XY)$.

3.55. Let X and Y be independent random variables such that

$$X = \begin{cases} 1 & \text{prob. } 1/3 \\ 0 & \text{prob. } 2/3 \end{cases} \qquad Y = \begin{cases} 2 & \text{prob. } 3/4 \\ -3 & \text{prob. } 1/4 \end{cases}$$

Find (a) $E(3X + 2Y)$, (b) $E(2X^2 - Y^2)$, (c) $E(XY)$, (d) $E(X^2Y)$.

3.56. Let $X_1, X_2, \ldots, X_n$ be n random variables which are identically distributed such that

$$X_k = \begin{cases} 1 & \text{prob. } 1/2 \\ 2 & \text{prob. } 1/3 \\ -1 & \text{prob. } 1/6 \end{cases}$$

Find (a) $E(X_1 + X_2 + \cdots + X_n)$, (b) $E(X_1^2 + X_2^2 + \cdots + X_n^2)$.

Variance and standard deviation

3.57. Find (a) the variance, (b) the standard deviation of the number of points that will come up on a single toss of a fair die.

3.58. Let X be a random variable having density function

$$f(x) = \begin{cases} 1/4 & -2 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $\mathrm{Var}(X)$, (b) σ_X.

3.59. Let X be a random variable having density function

$$f(x) = \begin{cases} e^{-x} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $\mathrm{Var}(X)$, (b) σ_X.

3.60. Find the variance and standard deviation for the random variable X of (a) Problem 3.43, (b) Problem 3.44.

3.61. A random variable X has $E(X) = 2$, $E(X^2) = 8$. Find (a) $\mathrm{Var}(X)$, (b) σ_X.

3.62. If a random variable X is such that $E[(X - 1)^2] = 10$, $E[(X - 2)^2] = 6$ find (a) $E(X)$, (b) $\mathrm{Var}(X)$, (c) σ_X.

Moments and moment generating functions

3.63. Find (a) the moment generating function of the random variable

$$X = \begin{cases} 1/2 & \text{prob. } 1/2 \\ -1/2 & \text{prob. } 1/2 \end{cases}$$

and (b) the first four moments about the origin.

3.64. (a) Find the moment generating function of a random variable X having density function

$$f(x) = \begin{cases} x/2 & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

(b) Use the generating function of (a) to find the first four moments about the origin.

3.65. Find the first four moments about the mean in (a) Problem 3.43, (b) Problem 3.44.

3.66. (a) Find the moment generating function of a random variable having density function

$$f(x) = \begin{cases} e^{-x} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

and (b) determine the first four moments about the origin.

3.67. In Problem 3.66 find the first four moments about the mean.

3.68. Let X have density function $f(x) = \begin{cases} 1/(b-a) & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$. Find the kth moment about (a) the origin, (b) the mean.

3.69. If $M(t)$ is the moment generating function of the random variable X, prove that the 3rd and 4th moments about the mean are given by

$$\mu_3 = M'''(0) - 3M''(0)M'(0) + 2[M'(0)]^3$$
$$\mu_4 = M^{(\text{iv})}(0) - 4M'''(0)M'(0) + 6M''(0)[M'(0)]^2 - 3[M'(0)]^4$$

Characteristic functions

3.70. Find the characteristic function of the random variable $X = \begin{cases} a & \text{prob. } p \\ b & \text{prob. } q = 1 - p \end{cases}$.

3.71. Find the characteristic function of a random variable X that has density function

$$f(x) = \begin{cases} 1/2a & |x| \le a \\ 0 & \text{otherwise} \end{cases}$$

3.72. Find the characteristic function of a random variable with density function

$$f(x) = \begin{cases} x/2 & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

3.73. Let $X_k = \begin{cases} 1 & \text{prob. } 1/2 \\ -1 & \text{prob. } 1/2 \end{cases}$ be independent random variables ($k = 1, 2, \ldots, n$). Prove that the characteristic function of the random variable

$$\frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}$$

is $[\cos(\omega/\sqrt{n})]^n$.

3.74. Prove that as $n \to \infty$ the characteristic function of Problem 3.73 approaches $e^{-\omega^2/2}$. (*Hint*: Take the logarithm of the characteristic function and use L'Hospital's rule.)

Covariance and correlation coefficient

3.75. Let X and Y be random variables having joint density function

$$f(x, y) = \begin{cases} x + y & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $\mathrm{Var}(X)$, (b) $\mathrm{Var}(Y)$, (c) σ_X, (d) σ_Y, (e) σ_{XY}, (f) ρ.

3.76. Work Problem 3.75 if the joint density function is $f(x, y) = \begin{cases} e^{-(x+y)} & x \ge 0,\ y \ge 0 \\ 0 & \text{otherwise} \end{cases}$.

3.77. Find (a) $\mathrm{Var}(X)$, (b) $\mathrm{Var}(Y)$, (c) σ_X, (d) σ_Y, (e) σ_{XY}, (f) ρ, for the random variables of Problem 2.56.

3.78. Work Problem 3.77 for the random variables of Problem 2.94.

3.79. Find (a) the covariance, (b) the correlation coefficient of two random variables X and Y if $E(X) = 2$, $E(Y) = 3$, $E(XY) = 10$, $E(X^2) = 9$, $E(Y^2) = 16$.

3.80. The correlation coefficient of two random variables X and Y is $-\frac{1}{4}$ while their variances are 3 and 5. Find the covariance.

Conditional expectation, variance, and moments

3.81. Let X and Y have joint density function

$$f(x, y) = \begin{cases} x + y & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the conditional expectation of (a) Y given X, (b) X given Y.

3.82. Work Problem 3.81 if $f(x, y) = \begin{cases} 2e^{-(x+2y)} & x \ge 0,\ y \ge 0 \\ 0 & \text{otherwise} \end{cases}$

3.83. Let X and Y have the joint probability function given in Table 2-9, page 71. Find the conditional expectation of (a) Y given X, (b) X given Y.

3.84. Find the conditional variance of (a) Y given X, (b) X given Y for the distribution of Problem 3.81.

3.85. Work Problem 3.84 for the distribution of Problem 3.82.

3.86. Work Problem 3.84 for the distribution of Problem 2.94.

Chebyshev's inequality

3.87. A random variable X has mean 3 and variance 2. Use Chebyshev's inequality to obtain an upper bound for (a) $P(|X - 3| \ge 2)$, (b) $P(|X - 3| \ge 1)$.

3.88. Prove Chebyshev's inequality for a discrete variable X. *(Hint*: See Problem 3.30.)

3.89. A random variable X has the density function $f(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$. (a) Find $P(|X - \mu| > 2)$. (b) Use Chebyshev's inequality to obtain an upper bound on $P(|X - \mu| > 2)$ and compare with the result in (a).

Law of large numbers

3.90. Show that the (weak) law of large numbers can be stated as

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1$$

and interpret.

3.91. Let X_k ($k = 1, \ldots, n$) be n independent random variables such that

$$X_k = \begin{cases} 1 & \text{prob. } p \\ 0 & \text{prob. } q = 1 - p \end{cases}$$

(a) If we interpret X_k to be the number of heads on the kth toss of a coin, what interpretation can be given to $S_n = X_1 + \cdots + X_n$?

(b) Show that the law of large numbers in this case reduces to

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) = 0$$

and interpret this result.

Other measures of central tendency

3.92. Find (a) the mode, (b) the median of a random variable X having density function

$$f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and (c) compare with the mean.

3.93. Work Problem 3.100 if the density function is

$$f(x) = \begin{cases} 4x(1 - x^2) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

3.94. Find (a) the median, (b) the mode for a random variable X defined by

$$X = \begin{cases} 2 & \text{prob. } 1/3 \\ -1 & \text{prob. } 2/3 \end{cases}$$

and (c) compare with the mean.

3.95. Find (a) the median, (b) the mode of the set of numbers 1, 3, 2, 1, 5, 6, 3, 3, and (c) compare with the mean.

Percentiles

3.96. Find the (a) 25th, (b) 75th percentile values for the random variable having density function

$$f(x) = \begin{cases} 2(1 - x) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

3.97. Find the (a) 10th, (b) 25th, (c) 75th, (d) 90th percentile values for the random variable having density function

$$f(x) = \begin{cases} c(x - x^3) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where c is an appropriate constant.

Other measures of dispersion

3.98. Find (a) the semi-interquartile range, (b) the mean deviation for the random variable of Problem 3.96.

3.99. Work Problem 3.98 for the random variable of Problem 3.97.

3.100. Find the mean deviation of the random variable X in each of the following cases.

(a) $f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$ (b) $f(x) = \dfrac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty.$

3.101. Obtain the probability that the random variable X differs from its mean by more than the semi-interquartile range in the case of (a) Problem 3.96, (b) Problem 3.100(a).

Skewness and kurtosis

3.102. Find the coefficient of (a) skewness, (b) kurtosis for the distribution of Problem 3.100(a).

3.103. If

$$f(x) = \begin{cases} c\left(1 - \dfrac{|x|}{a}\right) & |x| \leq a \\ 0 & |x| > a \end{cases}$$

where c is an appropriate constant, is the density function of X, find the coefficient of (a) skewness, (b) kurtosis.

3.104. Find the coefficient of (a) skewness, (b) kurtosis, for the distribution with density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Miscellaneous problems

3.105. Let X be a random variable that can take on the values 2, 1, and 3 with respective probabilities $1/3$, $1/6$, and $1/2$. Find (a) the mean, (b) the variance, (c) the moment generating function, (d) the characteristic function, (e) the third moment about the mean.

3.106. Work Problem 3.105 if X has density function

$$f(x) = \begin{cases} c(1 - x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where c is an appropriate constant.

3.107. Three dice, assumed fair, are tossed successively. Find (a) the mean, (b) the variance of the sum.

3.108. Let X be a random variable having density function

$$f(x) = \begin{cases} cx & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

where c is an appropriate constant. Find (a) the mean, (b) the variance, (c) the moment generating function, (d) the characteristic function, (e) the coefficient of skewness, (f) the coefficient of kurtosis.

3.109. Let X and Y have joint density function

$$f(x, y) = \begin{cases} cxy & 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $E(X^2 + Y^2)$, (b) $E(\sqrt{X^2 + Y^2})$.

3.110. Work Problem 3.109 if X and Y are independent identically distributed random variables having density function $f(u) = (2\pi)^{-1/2}e^{-u^2/2}$, $-\infty < u < \infty$.

3.111. Let X be a random variable having density function

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

and let $Y = X^2$. Find (a) $E(X)$, (b) $E(Y)$, (c) $E(XY)$.

ANSWERS TO SUPPLEMENTARY PROBLEMS

3.43. (a) 1 (b) 7 (c) 6 **3.44.** (a) 3/4 (b) 1/4 (c) 3/5

3.45. (a) 1 (b) 2 (c) 1 **3.46.** 10.5 **3.47.** 3

3.48. (a) 1 (b) 1 (c) 1/4

3.50. (a) 7/10 (b) 6/5 (c) 19/10 (d) 5/6

3.52. (a) 2/3 (b) 2/3 (c) 4/3 (d) 4/9

3.54. (a) 7/12 (b) 7/6 (c) 5/12 (d) 5/3 (e) 7/4 (f) 2/3

3.55. (a) 5/2 (b) –55/12 (c) 1/4 (d) 1/4

3.56. (a) n (b) $2n$ **3.57.** (a) 35/12 (b) $\sqrt{35/12}$

3.58. (a) 4/3 (b) $\sqrt{4/3}$ **3.59.** (a) 1 (b) 1

3.60. (a) $\text{Var}(X) = 5$, $\sigma_X = \sqrt{5}$ (b) $\text{Var}(X) = 3/80$, $\sigma_X = \sqrt{15}/20$

3.61. (a) 4 (b) 2 **3.62.** (a) 7/2 (b) 15/4 (c) $\sqrt{15}/2$

3.63. (a) $\frac{1}{2}(e^{t/2} + e^{-t/2}) = \cosh(t/2)$ (b) $\mu = 0, \mu'_2 = 1, \mu'_3 = 0, \mu'_4 = 1$

3.64. (a) $(1 + 2te^{2t} - e^{2t})/2t^2$ (b) $\mu = 4/3, \mu'_2 = 2, \mu'_3 = 16/5, \mu'_4 = 16/3$

3.65. (a) $\mu_1 = 0, \mu_2 = 5, \mu_3 = -5, \mu_4 = 35$ (b) $\mu_1 = 0, \mu_2 = 3/80, \mu_3 = -121/160, \mu_4 = 2307/8960$

3.66. (a) $1/(1 - t)$, $|t| < 1$ (b) $\mu = 1, \mu'_2 = 2, \mu'_3 = 6, \mu'_4 = 24$

3.67. $\mu_1 = 0, \mu_2 = 1, \mu_3 = 2, \mu_4 = 33$

3.68. (a) $(b^{k+1} - a^{k+1})/(k + 1)(b - a)$ (b) $[1 + (-1)^k](b - a)^k/2^{k+1}(k + 1)$

3.70. $pe^{i\omega a} + qe^{i\omega b}$ **3.71.** $(\sin a\omega)/a\omega$ **3.72.** $(e^{2i\omega} - 2i\omega e^{2i\omega} - 1)/2\omega^2$

3.75. (a) $11/144$ (b) $11/144$ (c) $\sqrt{11}/12$ (d) $\sqrt{11}/12$ (e) $-1/144$ (f) $-1/11$

3.76. (a) 1 (b) 1 (c) 1 (d) 1 (e) 0 (f) 0

3.77. (a) $73/960$ (b) $73/960$ (c) $\sqrt{73/960}$ (d) $\sqrt{73/960}$ (e) $-1/64$ (f) $-15/73$

3.78. (a) $233/324$ (b) $233/324$ (c) $\sqrt{233}/18$ (d) $\sqrt{233}/18$ (e) $-91/324$ (f) $-91/233$

3.79. (a) 4 (b) $4/\sqrt{35}$ **3.80.** $-\sqrt{15}/4$

3.81. (a) $(3x + 2)/(6x + 3)$ for $0 \le x \le 1$ (b) $(3y + 2)/(6y + 3)$ for $0 \le y \le 1$

3.82. (a) $1/2$ for $x \ge 0$ (b) 1 for $y \ge 0$

3.83. (a)

X	0	1	2
$E(Y \mid X)$	$4/3$	1	$5/7$

(b)

Y	0	1	2
$E(X \mid Y)$	$4/3$	$7/6$	$1/2$

3.84. (a) $\dfrac{6x^2 + 6x + 1}{18(2x + 1)^2}$ for $0 \le x \le 1$ (b) $\dfrac{6y^2 + 6y + 1}{18(2y + 1)^2}$ for $0 \le y \le 1$

3.85. (a) $1/9$ (b) 1

3.86. (a)

X	0	1	2
$\mathrm{Var}(Y \mid X)$	$5/9$	$4/5$	$24/49$

(b)

Y	0	1	2
$\mathrm{Var}(X \mid Y)$	$5/9$	$29/36$	$7/12$

3.87. (a) $1/2$ (b) 2 (useless) **3.89.** (a) e^{-2} (b) 0.5

3.92. (a) $+0$ (b) ln 2 (c) 1 **3.93.** (a) $1/\sqrt{3}$ (b) $\sqrt{1 - (1/\sqrt{2})}$ (c) $8/15$

3.94. (a) does not exist (b) -1 (c) 0 **3.95.** (a) 3 (b) 3 (c) 3

3.96. (a) $1 - \frac{1}{2}\sqrt{3}$ (b) $1/2$

3.97. (a) $\sqrt{1 - (3/\sqrt{10})}$ (b) $\sqrt{1 - (\sqrt{3}/2)}$ (c) $\sqrt{1/2}$ (d) $\sqrt{1 - (1/\sqrt{10})}$

3.98. (a) 1 (b) $(\sqrt{3} - 1)/4$ (c) $16/81$

3.99. (a) 1 (b) 0.17 (c) 0.051 **3.100.** (a) $1 - 2e^{-1}$ (b) does not exist

3.101. (a) $(5 - 2\sqrt{3})/3$ (b) $(3 - 2e^{-1}\sqrt{3})/3$

3.102. (a) 2 (b) 9 **3.103.** (a) 0 (b) $24/5a$ **3.104.** (a) 2 (b) 9

3.105. (a) 7/3 (b) 5/9 (c) $(e^t + 2e^{2t} + 3e^{3t})/6$ (d) $(e^{i\omega} + 2e^{2i\omega} + 3e^{3i\omega})/6$ (e) $-7/27$

3.106. (a) 1/3 (b) 1/18 (c) $2(e^t - 1 - t)/t^2$ (d) $-2(e^{i\omega} - 1 - i\omega)/\omega^2$ (*e*) 1/135

3.107. (a) 21/2 (b) 35/4

3.108. (a) 4/3 (b) 2/9 (c) $(1 + 2te^{2t} - e^{2t})/2t^2$ (d) $-(1 + 2i\omega e^{2i\omega} - e^{2i\omega})/2\omega^2$
(e) $-2\sqrt{18}/15$ (f) 12/5

3.109. (a) 1 (b) $8(2\sqrt{2} - 1)/15$

3.110. (a) 2 (b) $\sqrt{2\pi}/2$

3.111. (a) 0 (b) 1/3 (c) 0

Unit – 3 & 4
Test of Hypothesis

1. **Define Sample.**
   Solution: A Sample is a part of the statistical population (i.e) it is a subset which is collected to draw an inference about the population.
2. **Define Sample size.**
   Solution: The number of individuals in a sample is called the sample size
3. **Define Null hypotheses and Alternative hypothesis.**
   Solution: For applying the test of significance, we first set up of a hypothesis, a definite statement about the population parameter, such a hypothesis is usually called as null hypothesis and it is denoted by H_0.
   Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis and it is denoted byH_1.
4. **A random sample of 200 tins of coconut oil gave an average weight of 4.95 kgs with SD of 0.21 kg. Do we accept the hypothesis of net weight 5 kgs per tin at 1% level ? Explain. (L6)**
   Solution:
   Sample size n=200
   Sample mean$\bar{x}$=4.95kg
   Sample SD s=0.21kg
   Population mean μ=5kg.
   The sample is a large sample and so apply z-test.
   H_0 : μ=5kg
   H_1 : μ≠5kg
   The test statistic is z=$\frac{\bar{x}-\mu}{s/\sqrt{n}}$
   $$=\frac{4.95-5}{0.21/200}=\frac{-.05\times\sqrt{200}}{0.21}=-3.37$$
   ∴ $|z|$=3.37
   $$At\ 1\%\ level\ of\ significance\ the\ tabulated\ value\ of\ z\ is\ 2.58.$$
   Conclusion: H_0 is rejected at 1% $level$ since calculated value of $|z|$ is greater than the table value of z.Therefore the net weight tin is not equal to 5 kg.
5. **A sample of 900 items has mean 3.4 and SD 2.61. Test weather the sample be regarded as drawn from a population with mean 3.25 at 5% level of significance? (L4)**
   Solution:
   Sample size n=900

Sample mean $\bar{x}$=3.4

Sample SD s=2.61

Population mean μ=3.25

The sample is a large sample and so apply z-test.

H_0 : μ=3.25

H_1 : μ≠3.25

The test statistic is z=$\frac{\bar{x}-\mu}{s/\sqrt{n}}$

$$=\frac{3.4-3.25}{2.61/\sqrt{900}}=\frac{0.15}{2.61/30}=1.72$$

∴ $|z|$=1.72

$At\ 1\%\ level\ of\ significance\ the\ tabulated\ value\ of\ z\ is\ 2.58.$

Conclusion: H_0 is accepted at 1% *level* since calculated value of $|z|$ is less than the table value of z.Therefore H_0 is accepted.

6. **A Sample of 400 male students is found to have a mean height of 171.38 cms. Can it be reasonable regarded as a sample from a large population with mean height 171.17 cms and standard deviation 3.30 cms? Justify? (L6)**

Solution:

Sample size n=400

Sample mean $\bar{x}$=171.38cm

Population SD σ=3.30cm

Population mean μ=171.17cm

The sample is a large sample and so apply z-test.

H_0 : μ=171.17cm

H_1 : μ≠171.17cm

The test statistic is z=$\frac{\bar{x}-\mu}{s/\sqrt{n}}$

$$=\frac{171.38-171.17}{3.30/\sqrt{400}}=\frac{0.21\times 20}{3.30}=1.27$$

∴ $|z|$=1.72

$At\ 5\%\ level\ of\ significance\ the\ tabulated\ value\ of\ z\ is\ 1.96.$

Conclusion: H_0 is accepted at 5% *level* since calculated value of $|z|$ is less than the table value of z.Therefore H_0 is accepted andμ=171.17cm.

7. **The mean of two samples of 1000 and 2000 numbers are respectively 67.5 and 68 inches. Can they be regarded as draws from the same population with SD 2.5 inches? Justify? (L6)**

Solution:

$\bar{x}_1$=67.5, $\bar{x}_2$=68

n_1=1000, n_2=2000

Population SDσ=2.5

The two given samples are large samples.

$H_0 : \mu_1=\mu_2$

$H_1 : \mu_1 \neq \mu_2$

The test statistic is z=$\frac{\bar{x}_1-\bar{x}_2}{\sigma/\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ = $\frac{67.5-68}{2.5/\sqrt{\frac{1}{1000}+\frac{1}{2000}}}$ =-6.25

$\therefore |z|$=6.25

At 1% *level of significance the tabulated value of* z *is* 2.58.

Conclusion: H_0 is rejected at 1% *level* since calculated value of $|z|$ is greater than the table value of z.

$\therefore H_0$ is rejected at 1% *level* of significance and so the two samples cannot be regarded as belonging to the same population.

8. **The random samples of sizes 400 and 500 have mean 10.9 and 11.5 respectively. Can the samples be regarded as drawn from the same population with variance 25? Justify? (L6)**

Solution:

$\bar{x}_1$=10.9, $\bar{x}_2$=11.5

n_1=400, n_2=500

σ^2=25

The two given samples are large samples.

$H_0 : \mu_1=\mu_2$

$H_1 : \mu_1 \neq \mu_2$

The test statistic is z=$\frac{\bar{x}_1-\bar{x}_2}{\sigma/\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ = $\frac{10.9-11.5}{5/\sqrt{\frac{1}{400}+\frac{1}{500}}}$ =-2.38

$\therefore |z|$=2.38

At 1% *level of significance the tabulated value of* z *is* 2.58.

Conclusion: H_0 is accepted at 1% *level* since calculated value of $|z|$ is less than the table value of z.

Therefore the samples come from the population with variance 25.

9. **A sample of 26 bulbs given a mean life of 990 hours with a SD of 20 hours. The manufactures claims that the mean life of bulbs is 1000 hours. Is the sample not upto the standard? Justify? (L6)**

Solution:

Sample size n=26< 30(*small sample*)

Sample mean$\bar{x}$=990

Sample SD s=20

Population mean μ=1000

Degrees of freedom=n-1=26-1=25

Here we know $\bar{x}$,μ,SD and n.Therefore, we use student's 't' test.

H_0 : The sample is upto the standard.

H_1 : The sample is not upto the standard.

The test statistic is t=$\frac{\bar{x}-\mu}{s/\sqrt{n-1}}$

$$=\frac{990-1000}{20/\sqrt{25}}=-2.5$$

∴ $|t|$=2.5 (i.e) Calculated t=2.5

$At\ 5\%\ level\ of\ significance\ the\ tabulated\ value\ of\ z\ at\ 25d.f\ is\ 2.06$

Conclusion:$H_0$is rejected as calculated value is greater than the tabulated value. ∴The sample is not upto the standard.

10.In one sample of 8 observations the sum of the squares of deviations of the sample values from the sample mean was 84.4 and it the other sample of 10 observations it was 102.6. Test whether this difference is significant at 5% level? (L4)

Solution:

n_1=8, n_2=10

$S_1^2=\sum\frac{(x-\bar{x})^2}{n_1-1}=\frac{84.4}{7}$=12.057

$S_2^2=\sum\frac{(y-\bar{y})^2}{n_2-1}=\frac{102.6}{9}$ =11.4

H_0 :$S_1^2=S_2^2$

Now F=$\frac{S_1^2}{S_2^2}=\frac{12.057}{11.4}$ = 1.057

(i.e) calculated F=1.057

Tabulated value of F for(7,9) degrees of freedom is 3.29.

Calculated value F<Tabulated value F

∴We accept the null hypothesis.

11.A sample of size 13 gave an estimates population variance of 3.0, while another sample of size 15 gave an estimate of 2.5. Could both samples be from populations with the same variance. Justify? (L6)

Solution:

n_1=13, n_2=15

$$S_1^2=\sum\frac{(x-\bar{x})^2}{n_1-1}$$

$$S_2^2=\sum\frac{(y-\bar{y})^2}{n_2-1}$$

$H_0 : S_1^2 = S_2^2$.The two samples have come from populations with same variance.

∴The test statistic is

$F=\frac{S_1^2}{S_2^2} = \frac{(Greater\ variance)}{(Smaller\ variance)} = \frac{3.0}{2.5}=1.2$

(i.e) calculated F=1.2

Tabulated value of F for(12,14) degrees of freedom is 2.53

Calculated value F<Tabulated value F

∴We accept the null hypothesis H_0

$(i.e)$ Both samples have come from the populations with the same variance.

12.Write the test procedure of Chi-square test? (L5)

Solution:

(i) Write down the null hypothesis

(ii) Write down the alternative hypothesis.

(iii) Calculate the theoretical frequencies for the contingency.

(iv) Calculate $\aleph^2=\sum \frac{(O-E)^2}{E}$

(v) Write down the number of degress of freedom.

(vi) Write the conclusion on the hypothesis by comparing the calculated values of $\aleph^2$with table value of $\aleph^2$

13.Write the uses of $\aleph^2$ – test? (L1)

Solution:

(i) It is used to test the goodness of a distribution.

(ii) It is used to test the significance of the difference between the observed frequencies in a sample and the expected frequencies,obtained from the theoretical distribution.

(iii)It is also used to test the independence of attributes.

(iv)In case of small samples(where the population standard deviation is not known) $\aleph^2$ statistic is used to test whether a specified value can be the population variance σ^2.

14.A machine is designed to produce insulation washers for electrical devices of average thickness of 0.025cm. A random sample of 10 washers was found to have a thickness of 0.024cm with a S.D of 0.002 cm. Test the significance of the deviation value of t for 9 degrees of freedom at 5% $level$ is 2.262. (L4)

Solution:

Sample size n=10$< 30(small\ sample)$

Sample mean$\bar{x}$=0.024cm

Sample SD s=0.002cm

Population mean μ=0.025cm

Degrees of freedom=n-1=10-1=9

Here we know $\bar{x}$,μ,SD and n.Therefore, we use student's 't' test.

H_0 : The difference between $\bar{x}$ and μ is not significant

The test statistic is t=$\frac{\bar{x}-\mu}{s/\sqrt{n-1}} = -1.5$

∴ $|t|$=1.5 (i.e) Calculated t=1.5

$At\ 5\%\ level\ of\ significance\ the\ tabulated\ value\ of\ z\ at\ 9d.f\ is\ 2.06$

Conclusion:H_0 is accepted as calculated value is less than the tabulated value.

PART-B

1. **Find student's t, for the following variate value in a sample of eight -4, -2,-2,0,2,2,3,3 taking the mean of the universe to be zero.(L1)**

Solution:

Number of samples=8

∴n=8

Mean of universe is zero

∴ μ=0

$\bar{X}$=average value of X

$=\frac{(-4)+(-2)+(-2)+0+2+2+3+3}{8}$

=0.25

To calculate S,we have the formula

$S = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$

Hypothesis: There is no significant difference between sample mean and population mean

X	$X-\bar{X}$	$(X-\bar{X})^2$
-4	-4.25	18.06
-2	-2.25	5.06
-2	-2.25	5.06
0	-0.25	0.06
2	1.75	3.06
2	1.75	3.06
3	2.75	7.56
3	2.75	7.56

$$\sum(X-\bar{X})^2 = 49.98$$

S = $\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}} = \sqrt{\frac{49,48}{8-1}} = \sqrt{5.497} = 2.658$

t= $\frac{\bar{X}-\mu}{S/\sqrt{n-1}} = \frac{0.25-0}{2.658/\sqrt{7}}$ =0.248

Table value=2.26

∴calculated value <tabulated value

∴ Hypothesis is accepted and so there is no significant difference between sample mean and population mean.

2. **Ten students are selected at random in a university and their heights are measured in inches as 64,65,65,67,67,69,69,70,72 and 72.Using these data, Discuss the suggestion that the mean height of the students in the university is 66.(At 5% level of significance the value of t for 9 d.f is 2.262).(L2)**

Solution:

t= $\frac{\bar{X}-\mu}{S/\sqrt{n}}$

S = $\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$

$\bar{X}$=average value of X

= $\frac{64+65+65+67+67+69+69+70+72+72}{10}$ =68

Hypothesis: There is no significant difference between sample height and population height.

X	$X-\bar{X}$	$(X-\bar{X})^2$
64	-4	16
65	-3	9
65	-3	9
67	-1	1
67	-1	1
69	1	1
69	1	1
70	2	4
72	4	16
72	4	16

$$\sum(X-\bar{X})^2 = 74$$

S=$\sqrt{\frac{74}{10-1}}$=$\sqrt{\frac{74}{9}}$ =2.867

Here $\bar{X}$=68, $\mu = 66, n = 10$

$t = \frac{68-66}{2.867/\sqrt{10}}$=2.205

Table value=2.26

∴calculated value <tabulated value,therefore Hypothesis is accepted and the height of population group can be taken as 66.

3. **A fertilizer mixing machine is set to give 12kg of nitrate for every quintal bag of fertilizer.Ten 100kg bags are examined.The percentages of nitrate are as follows 11,14,13,12,13,12,13,14,11,12.Is there reason to belive that the machines is defective? (value of t for 9 d.f is 2.262). Justify? (L6)**

Solution:

Hypothesis: There is no significant difference between sample percentage and population percentage.

Here n=10

$$\mu = 12$$

$\bar{X}$=average value of X

=$\frac{11+14+13+12+13+12+13+14+11+12}{10}$ =12.5

X	$X-\bar{X}$	$(X-\bar{X})^2$
11	-1.5	2.25
14	1.5	2.25
13	0.5	0.25
12	-0.5	0.25
13	0.5	0.25
12	-0.5	0.25
13	0.5	0.25
14	1.5	2.25
11	-1.5	2.25
12	-0.5	0.25

$$\sum (X-\bar{X})^2 = 10.5$$

To calculate S,we have the formula

S = $\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$=$\sqrt{\frac{10.5}{10-1}}$

S=1.08

$$t=\frac{\bar{X}-\mu}{S/\sqrt{n}}=\frac{12.5-12}{1.08}\times 3$$

t=1.389

Table value=2.26

∴calculated value <tabulated value

∴ Hypothesis is accepted and the machine cannot be believed to be defective.

4. Two random samples drawn from two normal populations are given below.Test whether the two populations have the same variances (L4)

Samples I	**20**	**16**	**26**	**27**	**23**	**22**	**18**	**24**	**25**	**19**		
Samples II	**17**	**23**	**32**	**25**	**22**	**24**	**28**	**6**	**31**	**20**	**33**	**27**

Solution:

Hypothesis: There is no significant difference between variances of the two samples.

By Formula

$F=\frac{S_1^2}{S_2^2}$ if $S_1^2 > S_2^2$

$=\frac{S_2^2}{S_1^2}$ if $S_2^2 > S_1^2$

$$where\ S_1^2 = \sum \frac{(X_1-\overline{X_1})^2}{n_1-1}$$

$$S_2^2 = \sum \frac{(X_2-\overline{X_2})^2}{n_2-1}$$

Here $n_1 = 10, n_2$=12

Calculating the averages of two samples we get,

$$\overline{X_1} = 22, \overline{X_2} = 24$$

X_1	$X_1-\overline{X_1}$	$(X_1-\overline{X_1})^2$	X_2	$X_2-\overline{X_2}$	$(X_2-\overline{X_2})^2$
20	-2	4	17	-7	49
16	-6	36	23	-1	1
26	4	16	32	8	64
27	5	25	25	1	1
23	1	1	22	-2	4
22	0	0	24	0	0
18	-4	16	28	4	16
24	2	4	6	-18	324
25	3	9	31	7	49
19	-3	9	33	9	81
			20	-4	16

			27	3	9

$\sum(X_1 - \overline{X_1})^2 = 120 \sum(X_2 - \overline{X_2})^2$ =614

$$S_1^2 = \sum \frac{(X_1 - \overline{X_1})^2}{n_1 - 1} = \frac{120}{9} = 13.33$$

$$S_2^2 = \sum \frac{(X_2 - \overline{X_2})^2}{n_2 - 1} = \frac{614}{11} = 55.81 \qquad \Rightarrow S_2^2 > S_2^2$$

$\therefore$F=$\frac{S_2^2}{S_1^2}$=$\frac{55.81}{13.33}$=4.18

Degrees of freedom $\gamma_1 = 12 - 1 = 11, \quad \gamma_2 = 10 - 1 = 9$

Table value=3.10

Calculated value $>$ Table value

$\therefore$ Hypothesis is rejected.

$\therefore$ There is significant difference between the variance.

5. In two groups of ten children each increases in weight due to two different diets in the same period were in pounds.

8	**5**	**7**	**8**	**3**	**2**	**7**	**6**	**5**	**7**
3	**7**	**5**	**6**	**5**	**4**	**4**	**5**	**3**	**6**

Find whether the variance are significantly different . (L1)

Solution :

Ho : there is no significant Diffenence between the variance of the two samples

F = $\frac{S_1^2}{S_2^2}$ If $S_1^2 > S_2^2$= $\frac{S_2^2}{S_1^2}$ $if \quad S_2^2 > S_1^2$

Where S_1^2 =$\frac{\sum(X_1 - \overline{X_1})^2}{n_1 - 1}$ Here n_1=10 n_2= 10

S_2^2 =$\frac{\sum(X_2 - \overline{X_2})^2}{n_2 - 1}$ $\overleftarrow{X_1}$ =5.8$\overleftarrow{X_1}$=4.8

X_1	X_1- $\overleftarrow{X_1}$	$\left(X_1 - \overleftarrow{X_1}\right)^2$	X_2	X_2- $\overleftarrow{X_2}$	$\left(X_2 - \overleftarrow{X_2}\right)^2$
3	2.2	4.84	3	-1.8	3.24
5	-0.8	0.64	7	2.2	4.84
7	1.2	1.44	5	0.2	0.04
8	2.2	4.84	6	1.2	1.44
3	-2.8	7.84	5	0.2	0.04
2	-3.8	14.44	4	-0.8	0.64
7	1.2	1.44	4	-0.8	0.64
6	0.2	0.04	5	0.2	0.04
5	-0.8	1.64	3	-1.8	3.24
7	1.2	1.44	6	1.2	1.44

		$\sum\left(X_1 - \overleftarrow{X_1}\right)^2$ =37.6			$\sum\left(X_2 - \overleftarrow{X_2}\right)^2$=15.6

$S_1^2 = \frac{\sum(X_1 - \overline{X_1})^2}{n_1 - 1} = \frac{37.6}{9}$=4.18

$S_2^2 = \frac{\sum(X_2 - \overline{X_2})^2}{n_2 - 1} = \frac{15.6}{9}$=1.73

$$Here \quad S_1^2 > S_2^2$$

F = $\frac{S_1^2}{S_2^2} = \frac{4.18}{1.73}$ =2.42

$$Here$$

v_1=10-1=9 $\quad$ v_2=10-1=9

Degrees of freedom =9.9

Table value for the Degrees of freedom 9.9 at 5% level =3.23

Calculated value =2.42 $< Table\ value$

$\therefore H_0$ =Accepted

There is no significant deffeerence between the variance .

6. The nicotine contents in milligrams in two samples of tobacco were found to be as follows.

Samples A	**24**	**27**	**26**	**21**	**25**	
Samples B	**27**	**30**	**28**	**31**	**22**	**36**

Can it be said that the two samples have same variance.Justify?(L6)

Solution :

H_0 = There is no significant deffeerence between the variance of the two samples

X	X- $\bar{X}$	$(X - \bar{X})^2$	Y	Y- $\bar{Y}$	$(Y - \bar{Y})^2$
24	0.6	0.36	27	-2	4
27	2.4	5.76	30	1	1
26	1.4	1.96	28	-1	1
21	3.6	12.96	31	2	4
25	0.4	0.16	22	-7	49
			36	7	49
123		21.2	174		108

$\bar{X} = \frac{\sum X}{n} = \frac{123}{5}$ =24.6

$\bar{Y} = \frac{\sum Y}{n} = \frac{174}{6}$ =29

$S_1^2 = \frac{\sum(X - \bar{X})^2}{n_1 - 1} = \frac{21.2}{4}$=5.3

$S_2^2 = \frac{\sum(Y - \bar{Y})^2}{n_2 - 1} = \frac{108}{5}$=21.6

F= $\frac{S_2^2}{S_1^2}$=$\frac{21.6}{5.3}$ =4.07

Calculated value =4.07

Table value of F for $(5,4)$ d,f at 5% level is 6.26

$\therefore$ calculated value calculated value $<$ Table value.

$\therefore$ We accept H_0 ie ; The variance are equal .

7. Two random samples were drawn from two normal populations and their values are

A	**66**	**67**	**75**	**76**	**82**	**84**	**88**	**90**	**92**		
B	**64**	**66**	**74**	**78**	**82**	**85**	**87**	**92**	**93**	**95**	**97**

Test whether the two populations have the same variance at 5% level of Significance. (L4)

Solution :

There is no significant difference between the variance of the sample .

X	X- $\bar{X}$	$(X-\overline{X})^2$	Y	Y- $\bar{Y}$	$(Y-\overline{Y})^2$
66	-14	196	64	-19	361
67	-13	169	66	-17	289
75	-5	25	74	-9	81
76	-4	16	78	-5	25
82	2	4	82	-1	1
84	4	16	85	2	4
88	8	64	87	4	16
90	10	100	92	9	81
92	12	144	93	10	100
			95	12	144
			97	14	196
720	0	734	913	0	1298

S_1^2 =$\frac{\sum(X-\bar{X})^2}{n_1-1}$ =$\frac{734}{8}$=91. S_2^2 =$\frac{\sum(Y-\bar{Y})^2}{n_2-1}$ $=$ $\frac{1298}{10}$=129.8

$S_2^2 > S_1^2$

F= $\frac{S_2^2}{S_1^2}$=$\frac{129.8}{91.75}$ =1.41

Degree of freedom is $(10,8)$

Table value of F =3.34 AT 5% LEVEL

$\therefore$ calculated value $<$ Table value.

$\therefore$ WE Accepted H_0.

There is no significant difference between the variance of the two population .

8. **Do the following data give evidence of the effectiveness of inoculation?Justify?(L6)**

	Attacked	**Not attacked**
Inoculated	**20**	**300**
Not inoculated	**80**	**600**

Solution :

H_0 : There is no effect inoculation .

Table of observed frequencies is formed from the given data .

			TOTAL
	20 80	300 600	320 680
TOTAL	100	900	1000

Table of expected frequencies

			Total
	$\frac{100\times320}{1000}$=32	$\frac{900\times320}{1000}$=288	320
	$\frac{100\times6800}{1000}$=68	$\frac{900\times680}{1000}$=612	680
Total	100	900	1000

$chi\ square\ Table.$

O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
20 300 80 600	32 288 68 612	-12 12 12 -12	144 144 144 144	4.5 0.50 2.12 0.24
				$\sum\frac{(O-E)^2}{E} = 7.36$

Degrees of freedom =$(r-1)(c-1)$ =$(2-1)(2-1)$ =1

Table value of $\aleph^2$ for 1 d.f at 5% Level Is 3.841

c.v= 7.36 T.V =3.841 C.V>T.V

∴ Hypothesis is rejected ..There is effect of inoculation.

9. **The following data are collected on two characters**

	Smokers	**Non smokers**
Literates	**83**	**57**

Illiterates	45	68

Based on this ,can you say there is no relation between smoking and literacy. Justify? (L6)

<u>Solution :</u>

H_0 : There is no relation between smoking and literacy .

Table of observed frequencies

			Total
	83 45	57 68	140 113
Total	128	125	253

Table of expected frequencies

			Total
	$\frac{128\times140}{253}$=70.83	$\frac{125\times140}{253}$=69.17	140
	$\frac{100\times6800}{1000}$=57.17	$\frac{900\times680}{1000}$=55.83	113
Total	128	125	253

$chi\ square\ Table.$

O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
83 57 45 68	70.83 69.17 57.17 55.83	12.17 -12.17 -12.17 12.17	148.11 148.11 148.11 148.11	2.09 2.14 2.59 2.65
				$\sum\frac{(O-E)^2}{E} = 9..47$

Degrees of freedom =$(r-1)(c-1)$ =$(2-1)(2-1)$ =1

Table value of $\aleph^2$ for 1 d.f at 5% Level Is 3.841

c.v= 7.36 T.V =3.841 C.V>T.V

∴ Hypothesis is rejected ..There is a relation between smoking and literacy.

10.The following table gives the number of good and bad parts produced by each of three shifts in a factory.

Shifts	Good	Bad
Day	**900**	**130**

Evening	**700**	**170**
Night	**400**	**200**

Test if there is any association between shifts and quality. (L4)

Solution :

H_0 : There is no sognificant association between shifts and literacy quality.

Table of observed frequencies

			Total
	900 700 400	130 170 200	1030 870 600
Total	2000	500	2500

Table of expected frequencies.

			Total
	$\frac{2000\times1030}{2500}$ =824	$\frac{500\times1030}{2500}$ =206	1030
	$\frac{2000\times870}{2500}$ =696	$\frac{500\times870}{2500}$ =174	870
	$\frac{2000\times600}{2500}$ =480	$\frac{500\times600}{2500}$ =120	600
Total	2000	500	2500

$chi\ square\ Table.$

O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
900 130 700 170 400 200	824 206 696 174 480 120	76 -76 4 -4 -80 -80	5776 5776 16 16 6400 6400	7.01 28.04 0.02 0.09 13.33 53.33
				$\sum\frac{(O-E)^2}{E} = 9..47$

Degrees of freedom =$(r-1)(c-1)$ =$(3-1)(2-1)$ =2

Table value of $\aleph^2$ for 2 d.f at 5% Level Is 5.99

c.v= 101.83 T.V =5.99 C.V>T.V

∴ Hypothesis is rejected ..There is a association between shifts and quality.

11.The number of students in each category is given following table.

	Ability in Mathematics

Such in Medical school		**Low**	**Average**	**High**
	Low	**14**	**8**	**5**
	Average	**12**	**51**	**11**
	High	**7**	**24**	**18**

On the basis of contingency table,should we conclude that success in medical school is related to ability in Mathematics? Test at 0.05 level of significant. (L4)

<u>Solution :</u>

H_0 : There is no sognificant relation between success and abilityTable of observed frequencies

				Total
	14	8	5	27
	12	51	11	74
	7	24	18	49
Total	33	83	34	150

Table ofexpected frequencies.

				Total
	$\frac{33\times27}{150}$ =5.94	$\frac{83\times27}{150}$ =14.9	$\frac{34\times27}{150}$ =6.12	27
	$\frac{33\times74}{150}$ =16.2	$\frac{83\times74}{150}$ =40.9	$\frac{34\times74}{150}$ =16.7	74
	$\frac{33\times49}{150}$ =10.7	$\frac{83\times49}{150}$ =27.1	$\frac{34\times49}{150}$ =11.1	49
Total	33	83	34	150

$chi\ square\ Table.$

O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
14	5.94	8.06	64.96	10.94
8	14.9	-6.90	47.61	3.20
5	6.12	-1.12	1.25	0.20
12	16.2	-4.20	17.64	1.09
51	40.9	10.10	102.01	2.49
11	16.7	-5.7	32.49	1.95
7	10.7	-3.7	13.69	1.28
24	27.1	-3.10	9.61	0.35
18	11.1	6.90	47.61.	4.29

				$\sum \frac{(O-E)^2}{E} = 9..47$

Degrees of freedom =$(r-1)(c-1)$ =$(3-1)(3-1)$ =4

Table value of $\aleph^2$ for 4 d.f at 5% Level is 9.488

c.v= 25.79 T.V =9.488 C.V>T.V

∴ Hypothesis is rejected. There is a relation between success and ability.

12.A sample analysis of examination results of 500 students was made.It was found that 220 students had failed,170 had secured a third class,90 were placed in second class and 20 got a first class.Do these figures commensurate with the general examination result which is in the ratio of 4:3:2:1 for the various categories respectively.Explain? (L6)

<u>Solution:</u>

Null hypothesis H_0: The observed results commensurate with the general examination results.

Expected frequencies are in the ratio of 4: 3: 2: 1

Total frequency=500

If we divide the total frequency 500 in the ratio 4: 3: 2: 1 we get the expected frequencies as 200,150,100,50

$chi\ square\ Table.$

class	Observed frequency (O)	Expected frequencies (E)	O−E	$\frac{(O-E)^2}{E}$
Failed Third Second first	220 170 90 20	200 150 100 50	20 20 −10 −30	2.00 2.667 1.000 18.000
Total	500	500		23.667

Calculated $\aleph^2 = \sum \frac{(O-E)^2}{E} = 23.667$

Degrees of freedom = 4-1

(i.e) $\gamma = 3$

∴table value of $\aleph^2$ at 5% level for 3 d.f=7.81

∴calculated value>table value

∴We reject the null hypothesis (i.e) The observed results are not commensurate with the general examination results.

13.On the basis of information given below about the treatment of 200 patients suffering from a disease,state whether the new treatment is comparatively superior to the conventional treatment. (L1)

	Favourable	Not favourable	Total
New	**60**	**30**	**90**
Conventional	**40**	**70**	**110**

Solution:

Null hypothesis H_0:No difference between new and conventional treatment (or) New and conventional treatment are independent.

The no. of d.f is (2−1)(2−1)=1

Expected Frequency table:

			Total
	$\frac{90\times100}{100}$ =45	$\frac{90\times100}{200}$ =45	90
	$\frac{100\times110}{100}$ =55	$\frac{100\times110}{200}$ =55	110
Total	100	100	200

chi square Table.

Observed frequency (O)	Expected frequencies (E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
60	45	225	2.00
30	45	225	2.667
40	55	225	1.000
70	55	225	18.000
			18.18

Calculated $\aleph^2 = \sum\frac{(O-E)^2}{E} = 18.18$

∴Table value of $\aleph^2$ at 5% level for 1 d.f=3.841

∴calculated value>table value and so we reject the null hypothesis.

$(i.e)$New and conventional treatment are not independent.

14.Give the table for hair colour and eye colour.Find the value of ψ^2.Is there good association between the two. (L1)

	Hair colour				
Eye colour		**Fair**	**Brown**	**Black**	**Total**
	Blue	**15**	**5**	**20**	**40**

	Grey	**20**	**10**	**20**	**50**
	Brown	**25**	**15**	**20**	**60**
	Total	**60**	**30**	**60**	**150**

<u>Solution:</u>

Null hypothesis H_0:The two attributes Hair colour and Eye colour are independent.

<u>Expected Frequency table:</u>

				Total
	$\frac{60\times40}{150}$ =16	$\frac{30\times40}{150}$ =8	$\frac{60\times40}{150}$ =16	40
	$\frac{60\times50}{150}$ =20	$\frac{30\times50}{150}$ =10	$\frac{60\times50}{150}$ =20	50
	$\frac{60\times60}{150}$=24	$\frac{30\times60}{150}$ =12	$\frac{60\times60}{150}$ =24	60
Total	60	30	60	150

$chi\ square\ Table.$

Observed frequency (O)	Expected frequencies (E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
15	16	1	0.0625
5	8	9	1.125
20	16	16	1
20	20	0	0
10	10	0	0
20	20	0	0
25	24	1	0.042
15	12	9	0.75
20	24	16	0.666

Calculated $\aleph^2 = \sum\frac{(O-E)^2}{E} = 3.6458$

∴Table value of $\aleph^2$ at 5% level for 4 d.f=9.488

∴calculated value$<$table value and so we accept the null hypothesis.

$(i.e)$ The two attributes Hair colour and Eye colour are independent.

CHAPTER 6

Statistical Quality Control

Before studying this chapter you should know or, if necessary, review

1. Quality as a competitive priority, Chapter 2, page 00.
2. Total quality management (TQM) concepts, Chapter 5, pages 00–00.

LEARNING OBJECTIVES

After studying this chapter you should be able to

1 Describe categories of statistical quality control (SQC).
2 Explain the use of descriptive statistics in measuring quality characteristics.
3 Identify and describe causes of variation.
4 Describe the use of control charts.
5 Identify the differences between x-bar, R-, p-, and c-charts.
6 Explain the meaning of process capability and the process capability index.
7 Explain the term Six Sigma.
8 Explain the process of acceptance sampling and describe the use of operating characteristic (OC) curves.
9 Describe the challenges inherent in measuring quality in service organizations.

CHAPTER OUTLINE

We have all had the experience of purchasing a product only to discover that it is defective in some way or does not function the way it was designed to. This could be a new backpack with a broken zipper or an "out of the box" malfunctioning computer printer. Many of us have struggled to assemble a product the manufacturer has indicated would need only "minor" assembly, only to find that a piece of the product is missing or defective. As consumers, we expect the products we purchase to function as intended. However, producers of products know that it is not always possible to inspect every product and every aspect of the production process at all times. The challenge is to design ways to maximize the ability to monitor the quality of products being produced and eliminate defects.

One way to ensure a quality product is to build quality into the process. Consider Steinway & Sons, the premier maker of pianos used in concert halls all over the world. Steinway has been making pianos since the 1880s. Since that time the company's manufacturing process has not changed significantly. It takes the company nine months to a year to produce a piano by fashioning some 12,000-hand crafted parts, carefully measuring and monitoring every part of the process. While many of Steinway's competitors have moved to mass production, where pianos can be assembled in 20 days, Steinway has maintained a strategy of quality defined by skill and craftsmanship. Steinway's production process is focused on meticulous process precision and extremely high product consistency. This has contributed to making its name synonymous with top quality.

WHAT IS STATISTICAL QUALITY CONTROL?

Marketing, Management, Engineering

In Chapter 5 we learned that total quality management (TQM) addresses organizational quality from managerial and philosophical viewpoints. TQM focuses on customer-driven quality standards, managerial leadership, continuous improvement, quality built into product and process design, quality identified problems at the source, and quality made everyone's responsibility. However, talking about solving quality problems is not enough. We need specific tools that can help us make the right quality decisions. These tools come from the area of statistics and are used to help identify quality problems in the production process as well as in the product itself. Statistical quality control is the subject of this chapter.

▶ **Statistica1 quality control (SQC)**
The general category of statistical tools used to evaluate organizational quality.

Statistica1 quality control (SQC) is the term used to describe the set of statistical tools used by quality professionals. Statistical quality control can be divided into three broad categories:

▶ **Descriptive statistics**
Statistics used to describe quality characteristics and relationships.

1. **Descriptive statistics** are used to describe quality characteristics and relationships. Included are statistics such as the mean, standard deviation, the range, and a measure of the distribution of data.

2. **Statistical process control (SPC)** involves inspecting a random sample of the output from a process and deciding whether the process is producing products with characteristics that fall within a predetermined range. SPC answers the question of whether the process is functioning properly or not.
3. **Acceptance sampling** is the process of randomly inspecting a sample of goods and deciding whether to accept the entire lot based on the results. Acceptance sampling determines whether a batch of goods should be accepted or rejected.

► **Statistical process control (SPC)**
A statistical tool that involves inspecting a random sample of the output from a process and deciding whether the process is producing products with characteristics that fall within a predetermined range.

► **Acceptance sampling**
The process of randomly inspecting a sample of goods and deciding whether to accept the entire lot based on the results.

The tools in each of these categories provide different types of information for use in analyzing quality. Descriptive statistics are used to describe certain quality characteristics, such as the central tendency and variability of observed data. Although descriptions of certain characteristics are helpful, they are not enough to help us evaluate whether there is a problem with quality. Acceptance sampling can help us do this. Acceptance sampling helps us decide whether desirable quality has been achieved for a batch of products, and whether to accept or reject the items produced. Although this information is helpful in making the quality acceptance decision *after* the product has been produced, it does not help us identify and catch a quality problem *during* the production process. For this we need tools in the statistical process control (SPC) category.

All three of these statistical quality control categories are helpful in measuring and evaluating the quality of products or services. However, statistical process control (SPC) tools are used most frequently because they identify quality problems during the production process. For this reason, we will devote most of the chapter to this category of tools. The quality control tools we will be learning about do not only measure the value of a quality characteristic. They also help us identify a *change* or variation in some quality characteristic of the product or process. We will first see what types of variation we can observe when measuring quality. Then we will be able to identify specific tools used for measuring this variation.

Variation in the production process leads to quality defects and lack of product consistency. The Intel Corporation, the world's largest and most profitable manufacturer of microprocessors, understands this. Therefore, Intel has implemented a program it calls "copy-exactly" at all its manufacturing facilities. The idea is that regardless of whether the chips are made in Arizona, New Mexico, Ireland, or any of its other plants, they are made in exactly the same way. This means using the same equipment, the same exact materials, and workers performing the same tasks in the exact same order. The level of detail to which the "copy-exactly" concept goes is meticulous. For example, when a chipmaking machine was found to be a few feet longer at one facility than another, Intel made them match. When water quality was found to be different at one facility, Intel instituted a purification system to eliminate any differences. Even when a worker was found polishing equipment in one direction, he was asked to do it in the approved circular pattern. Why such attention to exactness of detail? The reason is to minimize all variation. Now let's look at the different types of variation that exist.

LINKS TO PRACTICE

Intel Corporation
www.intel.com

SOURCES OF VARIATION: COMMON AND ASSIGNABLE CAUSES

If you look at bottles of a soft drink in a grocery store, you will notice that no two bottles are filled to exactly the same level. Some are filled slightly higher and some slightly lower. Similarly, if you look at blueberry muffins in a bakery, you will notice that some are slightly larger than others and some have more blueberries than others. These types of differences are completely normal. No two products are exactly alike because of slight differences in materials, workers, machines, tools, and other factors. These are called **common, or random, causes of variation.** Common causes of variation are based on random causes that we cannot identify. These types of variation are unavoidable and are due to slight differences in processing.

► **Common causes of variation**
Random causes that cannot be identified.

An important task in quality control is to find out the range of natural random variation in a process. For example, if the average bottle of a soft drink called Cocoa Fizz contains 16 ounces of liquid, we may determine that the amount of natural variation is between 15.8 and 16.2 ounces. If this were the case, we would monitor the production process to make sure that the amount stays within this range. If production goes out of this range—bottles are found to contain on average 15.6 ounces—this would lead us to believe that there is a problem with the process because the variation is greater than the natural random variation.

The second type of variation that can be observed involves variations where the causes can be precisely identified and eliminated. These are called **assignable causes of variation.** Examples of this type of variation are poor quality in raw materials, an employee who needs more training, or a machine in need of repair. In each of these examples the problem can be identified and corrected. Also, if the problem is allowed to persist, it will continue to create a problem in the quality of the product. In the example of the soft drink bottling operation, bottles filled with 15.6 ounces of liquid would signal a problem. The machine may need to be readjusted. This would be an assignable cause of variation. We can assign the variation to a particular cause (machine needs to be readjusted) and we can correct the problem (readjust the machine).

► **Assignable causes of variation**
Causes that can be identified and eliminated.

DESCRIPTIVE STATISTICS

Descriptive statistics can be helpful in describing certain characteristics of a product and a process. The most important descriptive statistics are measures of central tendency such as the mean, measures of variability such as the standard deviation and range, and measures of the distribution of data. We first review these descriptive statistics and then see how we can measure their changes.

The Mean

In the soft drink bottling example, we stated that the average bottle is filled with 16 ounces of liquid. The arithmetic average, or the **mean,** is a statistic that measures the central tendency of a set of data. Knowing the central point of a set of data is highly important. Just think how important that number is when you receive test scores!

► **Mean (average)**
A statistic that measures the central tendency of a set of data.

To compute the mean we simply sum all the observations and divide by the total number of observations. The equation for computing the mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $\bar{x}$ = the mean
x_i = observation i, $i = 1, \ldots, n$
n = number of observations

The Range and Standard Deviation

In the bottling example we also stated that the amount of natural variation in the bottling process is between 15.8 and 16.2 ounces. This information provides us with the amount of variability of the data. It tells us how spread out the data is around the mean. There are two measures that can be used to determine the amount of variation in the data. The first measure is the **range,** which is the difference between the largest and smallest observations. In our example, the range for natural variation is 0.4 ounces.

Range
The difference between the largest and smallest observations in a set of data.

Another measure of variation is the **standard deviation.** The equation for computing the standard deviation is

Standard deviation
A statistic that measures the amount of data dispersion around the mean.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

where σ = standard deviation of a sample
$\bar{x}$ = the mean
x_i = observation i, $i = 1, \ldots, n$
n = the number of observations in the sample

Small values of the range and standard deviation mean that the observations are closely clustered around the mean. Large values of the range and standard deviation mean that the observations are spread out around the mean. Figure 6-1 illustrates the differences between a small and a large standard deviation for our bottling operation. You can see that the figure shows two distributions, both with a mean of 16 ounces. However, in the first distribution the standard deviation is large and the data are spread out far around the mean. In the second distribution the standard deviation is small and the data are clustered close to the mean.

FIGURE 6-1 Normal distributions with varying standard deviations

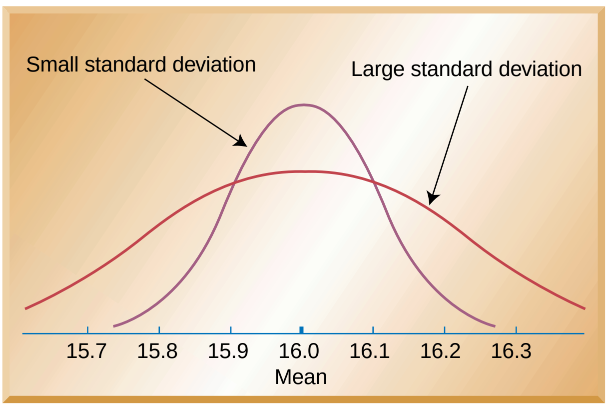
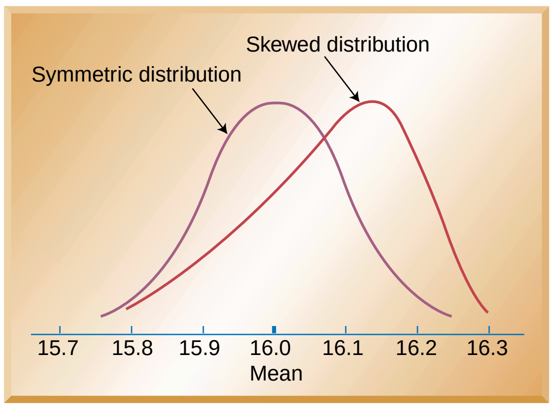


FIGURE 6-2 Differences between symmetric and skewed distributions



Distribution of Data

A third descriptive statistic used to measure quality characteristics is the shape of the distribution of the observed data. When a distribution is symmetric, there are the same number of observations below and above the mean. This is what we commonly find when only normal variation is present in the data. When a disproportionate number of observations are either above or below the mean, we say that the data has a *skewed distribution*. Figure 6-2 shows symmetric and skewed distributions for the bottling operation.

STATISTICAL PROCESS CONTROL METHODS

Statistical process control methods extend the use of descriptive statistics to monitor the quality of the product and process. As we have learned so far, there are common and assignable causes of variation in the production of every product. Using statistical process control we want to determine the amount of variation that is common or normal. Then we monitor the production process to make sure production stays within this normal range. That is, we want to make sure the process is in a *state of control*. The most commonly used tool for monitoring the production process is a control chart. Different types of control charts are used to monitor different aspects of the production process. In this section we will learn how to develop and use control charts.

Developing Control Charts

▶ **Control chart**
A graph that shows whether a sample of data falls within the common or normal range of variation.

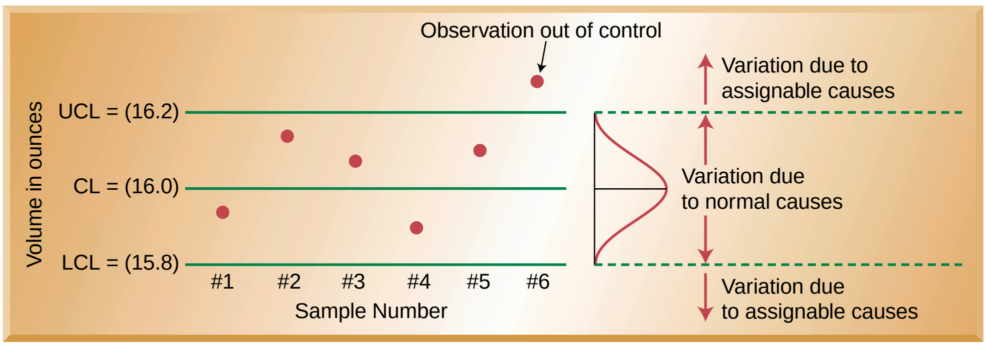
▶ **Out of control**
The situation in which a plot of data falls outside preset control limits.

A **control chart** (also called process chart or quality control chart) is a graph that shows whether a sample of data falls within the common or normal range of variation. A control chart has upper and lower control limits that separate common from assignable causes of variation. The common range of variation is defined by the use of control chart limits. We say that a process is **out of control** when a plot of data reveals that one or more samples fall outside the control limits.

Figure 6-3 shows a control chart for the Cocoa Fizz bottling operation. The *x* axis represents samples (#1, #2, #3, etc.) taken from the process over time. The *y* axis represents the quality characteristic that is being monitored (ounces of liquid). The center line (CL) of the control chart is the mean, or average, of the quality characteristic that is being measured. In Figure 6-3 the mean is 16 ounces. The upper control limit (UCL) is the maximum acceptable variation from the mean for a process that is in a state of control. Similarly, the lower control limit (LCL) is the minimum acceptable variation from the mean for a process that is in a state of control. In our example, the

FIGURE 6-3

Quality control chart for Cocoa Fizz



upper and lower control limits are 16.2 and 15.8 ounces, respectively. You can see that if a sample of observations falls outside the control limits we need to look for assignable causes.

The upper and lower control limits on a control chart are usually set at ± 3 standard deviations from the mean. If we assume that the data exhibit a normal distribution, these control limits will capture 99.74 percent of the normal variation. Control limits can be set at ± 2 standard deviations from the mean. In that case, control limits would capture 95.44 percent of the values. Figure 6-4 shows the percentage of values that fall within a particular range of standard deviation.

Looking at Figure 6-4, we can conclude that observations that fall outside the set range represent assignable causes of variation. However, there is a small probability that a value that falls outside the limits is still due to normal variation. This is called Type I error, with the error being the chance of concluding that there are assignable causes of variation when only normal variation exists. Another name for this is alpha risk (α), where alpha refers to the sum of the probabilities in both tails of the distribution that falls outside the confidence limits. The chance of this happening is given by the percentage or probability represented by the shaded areas of Figure 6-5. For limits of ± 3 standard deviations from the mean, the probability of a Type I error is .26% (100% − 99.74%), whereas for limits of ± 2 standard deviations it is 4.56% (100% − 95.44%).

Types of Control Charts

Control charts are one of the most commonly used tools in statistical process control. They can be used to measure any characteristic of a product, such as the weight of a cereal box, the number of chocolates in a box, or the volume of bottled water. The different characteristics that can be measured by control charts can be divided into two groups: **variables** and **attributes.** A *control chart for variables* is used to monitor characteristics that can be measured and have a continuum of values, such as height, weight, or volume. A soft drink bottling operation is an example of a variable measure, since the amount of liquid in the bottles is measured and can take on a number of different values. Other examples are the weight of a bag of sugar, the temperature of a baking oven, or the diameter of plastic tubing.

▶ **Variable**
A product characteristic that can be measured and has a continuum of values (e.g., height, weight, or volume).

▶ **Attribute**
A product characteristic that has a discrete value and can be counted.

FIGURE 6-4 Percentage of values captured by different ranges of standard deviation

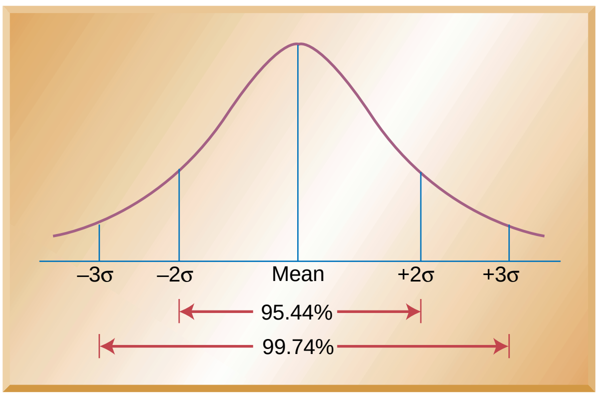
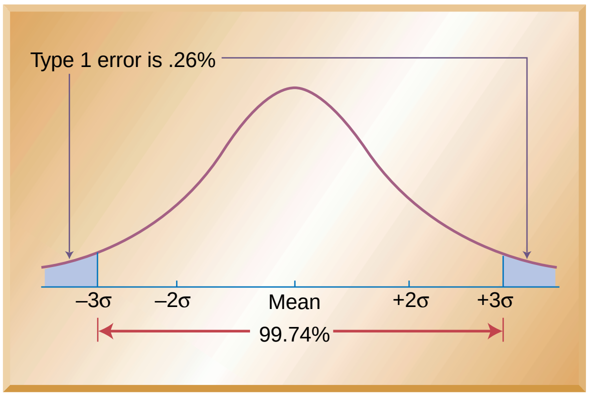


FIGURE 6-5 Chance of Type I error for $\pm 3\sigma$ (sigma-standard deviations)



A *control chart for attributes*, on the other hand, is used to monitor characteristics that have discrete values and can be counted. Often they can be evaluated with a simple yes or no decision. Examples include color, taste, or smell. The monitoring of attributes usually takes less time than that of variables because a variable needs to be measured (e.g., the bottle of soft drink contains 15.9 ounces of liquid). An attribute requires only a single decision, such as yes or no, good or bad, acceptable or unacceptable (e.g., the apple is good or rotten, the meat is good or stale, the shoes have a defect or do not have a defect, the lightbulb works or it does not work) or counting the number of defects (e.g., the number of broken cookies in the box, the number of dents in the car, the number of barnacles on the bottom of a boat).

Statistical process control is used to monitor many different types of variables and attributes. In the next two sections we look at how to develop control charts for variables and control charts for attributes.

CONTROL CHARTS FOR VARIABLES

Control charts for variables monitor characteristics that can be measured and have a continuous scale, such as height, weight, volume, or width. When an item is inspected, the variable being monitored is measured and recorded. For example, if we were producing candles, height might be an important variable. We could take samples of candles and measure their heights. Two of the most commonly used control charts for variables monitor both the central tendency of the data (the mean) and the variability of the data (either the standard deviation or the range). Note that each chart monitors a different type of information. When observed values go outside the control limits, the process is assumed not to be in control. Production is stopped, and employees attempt to identify the cause of the problem and correct it. Next we look at how these charts are developed.

Mean (x-Bar) Charts

▶ **x-bar chart**
A control chart used to monitor changes in the mean value of a process.

A mean control chart is often referred to as an ***x-bar chart.*** It is used to monitor changes in the mean of a process. To construct a mean chart we first need to construct the center line of the chart. To do this we take multiple samples and compute their means. Usually these samples are small, with about four or five observations. Each sample has its own mean, $\bar{x}$. The center line of the chart is then computed as the mean of all $\mathcal{K}$ sample means, where $\mathcal{K}$ is the number of samples:

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots \bar{x}_{\mathcal{K}}}{\mathcal{K}}$$

To construct the upper and lower control limits of the chart, we use the following formulas:

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + z\sigma_{\bar{x}}$$

$$\text{Lower control limit (LCL)} = \bar{\bar{x}} - z\sigma_{\bar{x}}$$

where $\bar{\bar{x}}$ = the average of the sample means
z = standard normal variable (2 for 95.44% confidence, 3 for 99.74% confidence)
$\sigma_{\bar{x}}$ = standard deviation of the distribution of sample means, computed as $\sigma/\sqrt{n}$
σ = population (process) standard deviation
n = sample size (number of observations per sample)

Example 6.1 shows the construction of a mean (x-bar) chart.

EXAMPLE 6.1

Constructing a Mean (x-Bar) Chart

A quality control inspector at the Cocoa Fizz soft drink company has taken twenty-five samples with four observations each of the volume of bottles filled. The data and the computed means are shown in the table. If the standard deviation of the bottling operation is 0.14 ounces, use this information to develop control limits of three standard deviations for the bottling operation.

Sample Number	Observations (bottle volume in ounces) 1	2	3	4	Average $\bar{x}$	Range R
1	15.85	16.02	15.83	15.93	15.91	0.19
2	16.12	16.00	15.85	16.01	15.99	0.27
3	16.00	15.91	15.94	15.83	15.92	0.17
4	16.20	15.85	15.74	15.93	15.93	0.46
5	15.74	15.86	16.21	16.10	15.98	0.47
6	15.94	16.01	16.14	16.03	16.03	0.20
7	15.75	16.21	16.01	15.86	15.96	0.46
8	15.82	15.94	16.02	15.94	15.93	0.20
9	16.04	15.98	15.83	15.98	15.96	0.21
10	15.64	15.86	15.94	15.89	15.83	0.30
11	16.11	16.00	16.01	15.82	15.99	0.29
12	15.72	15.85	16.12	16.15	15.96	0.43
13	15.85	15.76	15.74	15.98	15.83	0.24
14	15.73	15.84	15.96	16.10	15.91	0.37
15	16.20	16.01	16.10	15.89	16.05	0.31
16	16.12	16.08	15.83	15.94	15.99	0.29
17	16.01	15.93	15.81	15.68	15.86	0.33
18	15.78	16.04	16.11	16.12	16.01	0.34
19	15.84	15.92	16.05	16.12	15.98	0.28
20	15.92	16.09	16.12	15.93	16.02	0.20
21	16.11	16.02	16.00	15.88	16.00	0.23
22	15.98	15.82	15.89	15.89	15.90	0.16
23	16.05	15.73	15.73	15.93	15.86	0.32
24	16.01	16.01	15.89	15.86	15.94	0.15
25	16.08	15.78	15.92	15.98	15.94	0.30
Total					398.75	7.17

• Solution

The center line of the control data is the average of the samples:

$$\bar{\bar{x}} = \frac{398.75}{25}$$

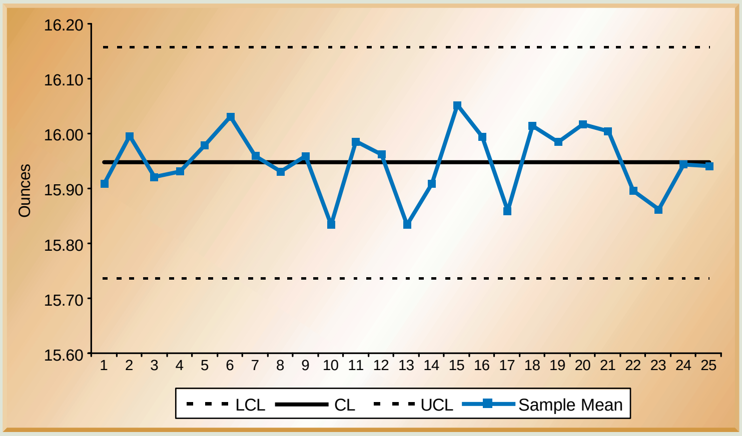
$$\bar{\bar{x}} = 15.95$$

The control limits are

$$UCL = \bar{\bar{x}} + z\sigma_{\bar{x}} = 15.95 + 3\left(\frac{.14}{\sqrt{4}}\right) = 16.16$$

$$LCL = \bar{\bar{x}} - z\sigma_{\bar{x}} = 15.95 - 3\left(\frac{.14}{\sqrt{4}}\right) = 15.74$$

The resulting control chart is:



This can also be computed using a spreadsheet as shown.

	A	B	C	D	E	F	G
1							
2	**X-Bar Chart: Cocoa Fizz**						
3					G7: =MAX(B7:E7)-MIN(B7:E7)		
4		F7: =AVERAGE(B7:E7)					
5		**Bottle Volume in Ounces**					
6	Sample Num	Obs 1	Obs 2	Obs 3	Obs 4	Average	Range
7	1	15.85	16.02	15.83	15.93	15.91	0.19
8	2	16.12	16.00	15.85	16.01	16.00	0.27
9	3	16.00	15.91	15.94	15.83	15.92	0.17
10	4	16.20	15.85	15.74	15.93	15.93	0.46
11	5	15.74	15.86	16.21	16.10	15.98	0.47
12	6	15.94	16.01	16.14	16.03	16.03	0.20
13	7	15.75	16.21	16.01	15.86	15.96	0.46
14	8	15.82	15.94	16.02	15.94	15.93	0.20
15	9	16.04	15.98	15.83	15.98	15.96	0.21
16	10	15.64	15.86	15.94	15.89	15.83	0.30
17	11	16.11	16.00	16.01	15.82	15.99	0.29
18	12	15.72	15.85	16.12	16.15	15.96	0.43
19	13	15.85	15.76	15.74	15.98	15.83	0.24
20	14	15.73	15.84	15.96	16.10	15.91	0.37
21	15	16.20	16.01	16.10	15.89	16.05	0.31
22	16	16.12	16.08	15.83	15.94	15.99	0.29
23	17	16.01	15.93	15.81	15.68	15.86	0.33
24	18	15.78	16.04	16.11	16.12	16.01	0.34
25	19	15.84	15.92	16.05	16.12	15.98	0.28
26	20	15.92	16.09	16.12	15.93	16.02	0.20
27	21	16.11	16.02	16.00	15.88	16.00	0.23
28	22	15.98	15.82	15.89	15.89	15.90	0.16
29	23	16.05	15.73	15.73	15.93	15.86	0.32
30	24	16.01	16.01	15.89	15.86	15.94	0.15
31	25	16.08	15.78	15.92	15.98	15.94	0.30
32						15.95	0.29
33		Number of Samples		25		Xbar-bar	R-bar
34	Number of Observations per Sample			4			
35				F32: =AVERAGE(F7:F31)		G32: =AVERAGE(G7:G31)	
36							

	A	B	C	D	E	F	G
39	**Computations for X-Bar Chart**				D40: =F32		
40			Overall Mean (Xbar-bar) =	15.95			
41			Sigma for Process =	0.14	ounces	D42: =D41/SQRT(D34)	
42			Standard Error of the Mean =	0.07			
43			Z-value for control charts =	3			
44					D45: =D40		
45			CL: Center Line =	**15.95**	D46: =D40-D43*D42		
46			LCL: Lower Control Limit =	**15.74**	D47: =D40+D43*D42		
47			UCL: Upper Control Limit =	**16.16**			

Another way to construct the control limits is to use the sample range as an estimate of the variability of the process. Remember that the range is simply the difference between the largest and smallest values in the sample. The spread of the range can tell us about the variability of the data. In this case control limits would be constructed as follows:

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + A_2\bar{R}$$
$$\text{Lower control limit (LCL)} = \bar{\bar{x}} - A_2\bar{R}$$

where $\bar{\bar{x}}$ = average of the sample means
$\bar{R}$ = average range of the samples
A_2 = factor obtained from Table 6-1.

Notice that A_2 is a factor that includes three standard deviations of ranges and is dependent on the sample size being considered.

EXAMPLE 6.2

Constructing a Mean (x-Bar) Chart from the Sample Range

A quality control inspector at Cocoa Fizz is using the data from Example 6.1 to develop control limits. If the average range ($\bar{R}$) for the twenty-five samples is .29 ounces (computed as $\frac{7.17}{25}$) and the average mean ($\bar{\bar{x}}$) of the observations is 15.95 ounces, develop three-sigma control limits for the bottling operation.

• Solution

$$\bar{\bar{x}} = 15.95 \text{ ounces} \qquad \bar{R} = .29$$

The value of A_2 is obtained from Table 6.1. For $n = 4$, $A_2 = .73$. This leads to the following limits:

$$\text{The center of the control chart} = \text{CL} = 15.95 \text{ ounces}$$
$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 15.95 + (.73)(.29) = 16.16$$
$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 15.95 - (.73)(.29) = 15.74$$

TABLE 6-1

Factors for three-sigma control limits of $\bar{x}$ and R-charts

Source: Factors adapted from the *ASTM Manual on Quality Control of Materials.*

Sample Size n	Factor for $\bar{x}$-Chart A_2	Factors for R-Chart D_3	D_4
2	1.88	0	3.27
3	1.02	0	2.57
4	0.73	0	2.28
5	0.58	0	2.11
6	0.48	0	2.00
7	0.42	0.08	1.92
8	0.37	0.14	1.86
9	0.34	0.18	1.82
10	0.31	0.22	1.78
11	0.29	0.26	1.74
12	0.27	0.28	1.72
13	0.25	0.31	1.69
14	0.24	0.33	1.67
15	0.22	0.35	1.65
16	0.21	0.36	1.64
17	0.20	0.38	1.62
18	0.19	0.39	1.61
19	0.19	0.40	1.60
20	0.18	0.41	1.59
21	0.17	0.43	1.58
22	0.17	0.43	1.57
23	0.16	0.44	1.56
24	0.16	0.45	1.55
25	0.15	0.46	1.54

Range (R) Charts

▶ **Range (R) chart**
A control chart that monitors changes in the dispersion or variability of process.

Range (R) charts are another type of control chart for variables. Whereas x-bar charts measure shift in the central tendency of the process, range charts monitor the dispersion or variability of the process. The method for developing and using R-charts is the same as that for x-bar charts. The center line of the control chart is the average range, and the upper and lower control limits are computed as follows:

$$\mathrm{CL} = \bar{R}$$

$$\mathrm{UCL} = D_4 \bar{R}$$

$$\mathrm{LCL} = D_3 \bar{R}$$

where values for D_4 and D_3 are obtained from Table 6-1.

EXAMPLE 6.3

Constructing a Range (R) Chart

The quality control inspector at Cocoa Fizz would like to develop a range (R) chart in order to monitor volume dispersion in the bottling process. Use the data from Example 6.1 to develop control limits for the sample range.

• Solution

From the data in Example 6.1 you can see that the average sample range is:

$$\bar{R} = \frac{7.17}{25}$$

$$\bar{R} = 0.29$$

$$n = 4$$

From Table 6-1 for $n = 4$:

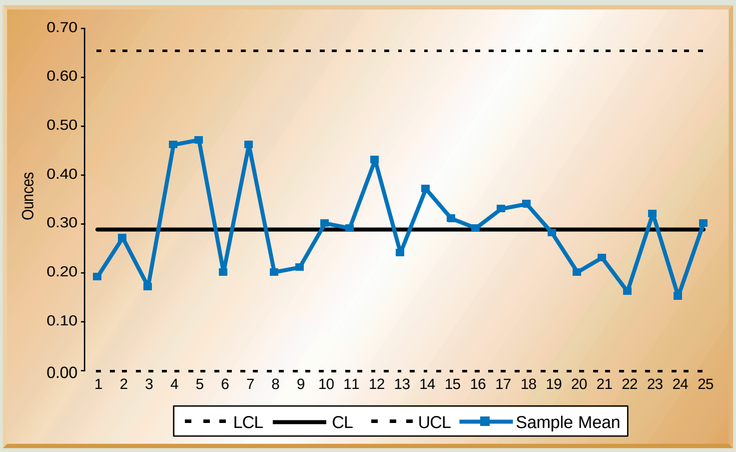
$$D_4 = 2.28$$

$$D_3 = 0$$

$$\text{UCL} = D_4\,\bar{R} = 2.28\,(0.29) = 0.6612$$

$$\text{LCL} = D_3\,\bar{R} = 0\,(0.29) = 0$$

The resulting control chart is:

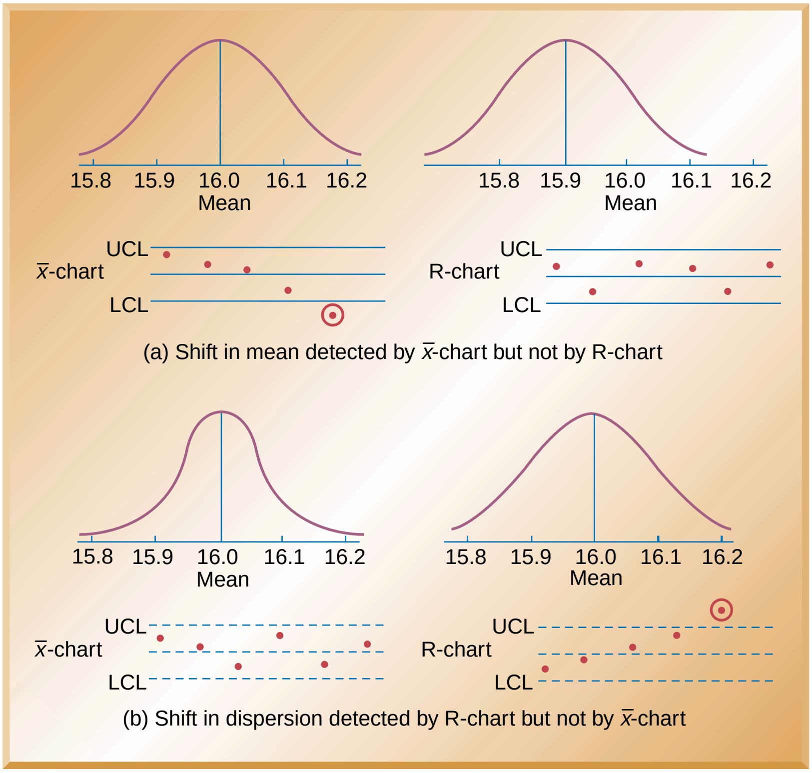


Using Mean and Range Charts Together

You can see that mean and range charts are used to monitor different variables. The mean or x-bar chart measures the central tendency of the process, whereas the range chart measures the dispersion or variance of the process. Since both variables are important, it makes sense to monitor a process using both mean and

FIGURE 6-6

Process shifts captured by $\bar{x}$-charts and R-charts



range charts. It is possible to have a shift in the mean of the product but not a change in the dispersion. For example, at the Cocoa Fizz bottling plant the machine setting can shift so that the average bottle filled contains not 16.0 ounces, but 15.9 ounces of liquid. The dispersion could be the same, and this shift would be detected by an x-bar chart but not by a range chart. This is shown in part (a) of Figure 6-6. On the other hand, there could be a shift in the dispersion of the product without a change in the mean. Cocoa Fizz may still be producing bottles with an average fill of 16.0 ounces. However, the dispersion of the product may have increased, as shown in part (b) of Figure 6-6. This condition would be detected by a range chart but not by an x-bar chart. Because a shift in either the mean or the range means that the process is out of control, it is important to use both charts to monitor the process.

CONTROL CHARTS FOR ATTRIBUTES

Control charts for attributes are used to measure quality characteristics that are counted rather than measured. Attributes are discrete in nature and entail simple yes-or-no decisions. For example, this could be the number of nonfunctioning lightbulbs, the proportion of broken eggs in a carton, the number of rotten apples, the number of scratches on a tile, or the number of complaints issued. Two of the most common types of control charts for attributes are p-charts and c-charts.

P-charts are used to measure the proportion of items in a sample that are defective. Examples are the proportion of broken cookies in a batch and the proportion of cars produced with a misaligned fender. P-charts are appropriate when both the number of defectives measured and the size of the total sample can be counted. A proportion can then be computed and used as the statistic of measurement.

C-charts count the actual number of defects. For example, we can count the number of complaints from customers in a month, the number of bacteria on a petri dish, or the number of barnacles on the bottom of a boat. However, we *cannot* compute the proportion of complaints from customers, the proportion of bacteria on a petri dish, or the proportion of barnacles on the bottom of a boat.

Problem-Solving Tip: The primary difference between using a p-chart and a c-chart is as follows. A p-chart is used when both the total sample size and the number of defects can be computed. A c-chart is used when we can compute *only* the number of defects but cannot compute the proportion that is defective.

P-Charts

P-charts are used to measure the proportion that is defective in a sample. The computation of the center line as well as the upper and lower control limits is similar to the computation for the other kinds of control charts. The center line is computed as the average proportion defective in the population, $\bar{p}$. This is obtained by taking a number of samples of observations at random and computing the average value of p across all samples.

P-chart
A control chart that monitors the *proportion* of defects in a sample.

To construct the upper and lower control limits for a p-chart, we use the following formulas:

$$\text{UCL} = \bar{p} + z\sigma_p$$

$$\text{LCL} = \bar{p} - z\sigma_p$$

where z = standard normal variable
$\bar{p}$ = the sample proportion defective
σ_p = the standard deviation of the average proportion defective

As with the other charts, z is selected to be either 2 or 3 standard deviations, depending on the amount of data we wish to capture in our control limits. Usually, however, they are set at 3.

The sample standard deviation is computed as follows:

$$\sigma_p = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

where n is the sample size.

EXAMPLE 6.4

Constructing a p-Chart

A production manager at a tire manufacturing plant has inspected the number of defective tires in twenty random samples with twenty observations each. Following are the number of defective tires found in each sample:

Sample Number	Number of Defective Tires	Number of Observations Sampled	Fraction Defective
1	3	20	.15
2	2	20	.10
3	1	20	.05
4	2	20	.10
5	1	20	.05
6	3	20	.15
7	3	20	.15
8	2	20	.10
9	1	20	.05
10	2	20	.10
11	3	20	.15
12	2	20	.10
13	2	20	.10
14	1	20	.05
15	1	20	.05
16	2	20	.10
17	4	20	.20
18	3	20	.15
19	1	20	.05
20	1	20	.05
Total	40	400	

Construct a three-sigma control chart ($z = 3$) with this information.

• Solution

The center line of the chart is

$$\text{CL} = \bar{p} = \frac{\text{total number of defective tires}}{\text{total number of observations}} = \frac{40}{400} = .10$$

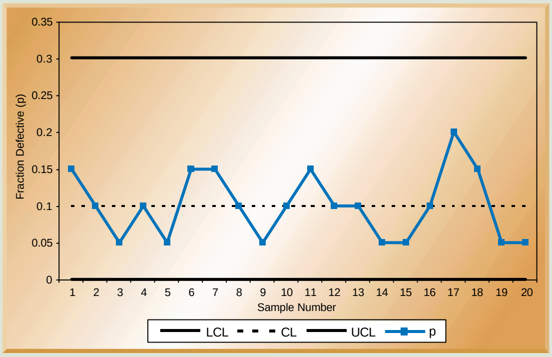
$$\sigma_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{(.10)(.90)}{20}} = .067$$

$$\text{UCL} = \bar{p} + z(\sigma_p) = .10 + 3(.067) = .301$$

$$\text{LCL} = \bar{p} - z(\sigma_p) = .10 - 3(.067) = -.101 \longrightarrow 0$$

In this example the lower control limit is negative, which sometimes occurs because the computation is an approximation of the binomial distribution. When this occurs, the LCL is rounded up to zero because we cannot have a negative control limit.

The resulting control chart is as follows:



This can also be computed using a spreadsheet as shown below.

	A	B	C	D
1				
2	**Constructing a p-Chart**			
3				
4		Size of Each Sample	20	
5		Number Samples	20	
6				
7	**Sample #**	**# Defective Tires**	**Fraction Defective**	C8: =B8/C$4
8	1	3	0.15	
9	2	2	0.10	
10	3	1	0.05	
11	4	2	0.10	
12	5	1	0.05	
13	6	3	0.15	
14	7	3	0.15	
15	8	2	0.10	
16	9	1	0.05	
17	10	2	0.10	
18	11	3	0.15	
19	12	2	0.10	
20	13	2	0.10	
21	14	1	0.05	
22	15	1	0.05	
23	16	2	0.10	
24	17	4	0.20	
25	18	3	0.15	
26	19	1	0.05	
27	20	1	0.05	

	A	B	C	D	E	F
29	**Computations for p-Chart**					
30		p bar =	0.100	C29: =SUM(B8:B27)/(C4*C5)		
31		Sigma_p =	0.067	C30: =SQRT((C29*(1-C29))/C4)		
32	Z-value for control charts =		3			
33						
34	CL: Center Line =		**0.100**	C33: =C29		
35	LCL: Lower Control Limit =		**0.000**	C34: =MAX(C$29-C$31*C$30,0)		
36	UCL: Upper Control Limit =		**0.301**	C35: =C$29+C$31*C$30		

C-CHARTS

▶ **C-chart**
A control chart used to monitor the *number* of defects per unit.

C-charts are used to monitor the number of defects per unit. Examples are the number of returned meals in a restaurant, the number of trucks that exceed their weight limit in a month, the number of discolorations on a square foot of carpet, and the number of bacteria in a milliliter of water. Note that the types of units of measurement we are considering are a period of time, a surface area, or a volume of liquid.

The average number of defects, $\bar{c}$, is the center line of the control chart. The upper and lower control limits are computed as follows:

$$\text{UCL} = \bar{c} + z\sqrt{\bar{c}}$$
$$\text{LCL} = \bar{c} - z\sqrt{\bar{c}}$$

EXAMPLE 6.5

Computing a C-Chart

The number of weekly customer complaints are monitored at a large hotel using a c-chart. Complaints have been recorded over the past twenty weeks. Develop three-sigma control limits using the following data:

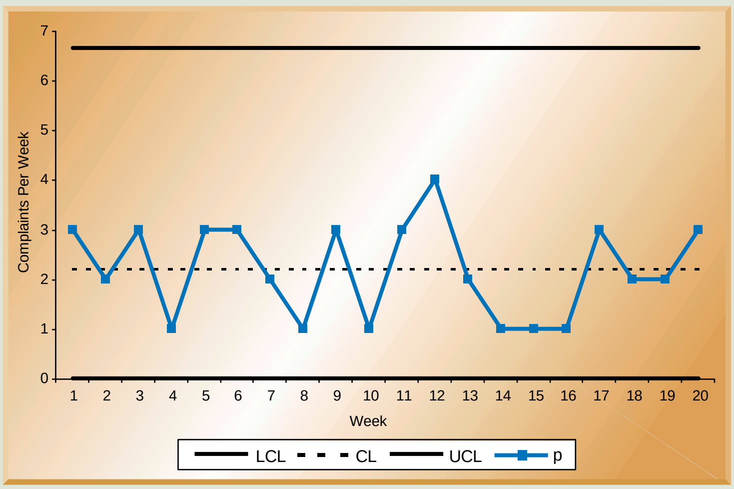
																					Tota
Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
No. of Complaints	3	2	3	1	3	3	2	1	3	1	3	4	2	1	1	1	3	2	2	3	**44**

• Solution

The average number of complaints per week is $\frac{44}{20} = 2.2$. Therefore, $\bar{c} = 2.2$.

$$\text{UCL} = \bar{c} + z\sqrt{\bar{c}} = 2.2 + 3\sqrt{2.2} = 6.65$$
$$\text{LCL} = \bar{c} - z\sqrt{\bar{c}} = 2.2 - 3\sqrt{2.2} = -2.25 \longrightarrow 0$$

As in the previous example, the LCL is negative and should be rounded up to zero. Following is the control chart for this example:



This can also be computed using a spreadsheet as shown below.

	A	B
1		
2	**Computing a C-Chart**	
3		
4	**Week**	**Number of Complaints**
5	1	3
6	2	2
7	3	3
8	4	1
9	5	3
10	6	3
11	7	2
12	8	1
13	9	3
14	10	1
15	11	3
16	12	4
17	13	2
18	14	1
19	15	1
20	16	1
21	17	3
22	18	2
23	19	2
24	20	3

	A	B	C	D	E	F	G
26	**Computations for a C-Chart**			C27: =AVERAGE(B5:B24)			
27		c bar =	2.2				
28		Z-value for control charts =	3	C30: =SQRT(C27)			
29							
30		Sigma_c =	1.4832397	C31: =C26			
31				C32: =MAX(C$26-C$27*C$29,0)			
32		CL: Center Line =	**2.20**	C33: =C$26+C$27*C$29			
33		LCL: Lower Control Limit =	**0.00**				
34		UCL: Upper Control Limit =	**6.65**				

Before You Go On

We have discussed several types of statistical quality control (SQC) techniques. One category of SQC techniques consists of descriptive statistics tools such as the mean, range, and standard deviation. These tools are used to describe quality characteristics and relationships. Another category of SQC techniques consists of statistical process control (SPC) methods that are used to monitor changes in the production process. To understand SPC methods you must understand the differences between common and assignable causes of variation. Common causes of variation are based on random causes that cannot be identified. A certain amount of common or normal variation occurs in every process due to differences in materials, workers, machines, and other factors. Assignable causes of variation, on the other hand, are variations that can be identified and eliminated. An important part of statistical process control (SPC) is monitoring the production process to make sure that the only variations in the process are those due to common or normal causes. Under these conditions we say that a production process is in a *state of control*.

You should also understand the different types of quality control charts that are used to monitor the production process: x-bar charts, R-range charts, p-charts, and c-charts.

PROCESS CAPABILITY

► **Process capability**
The ability of a production process to meet or exceed preset specifications.

► **Product specifications**
Preset ranges of acceptable quality characteristics.

So far we have discussed ways of monitoring the production process to ensure that it is in a *state of control* and that there are no assignable causes of variation. A critical aspect of statistical quality control is evaluating the ability of a production process to meet or exceed preset specifications. This is called **process capability.** To understand exactly what this means, let's look more closely at the term *specification*. **Product specifications,** often called *tolerances*, are preset ranges of acceptable quality characteristics, such as product dimensions. For a product to be considered acceptable, its characteristics must fall within this preset range. Otherwise, the product is not acceptable. Product specifications, or tolerance limits, are usually established by design engineers or product design specialists.

For example, the specifications for the width of a machine part may be specified as 15 inches $\pm.3$. This means that the width of the part should be 15 inches, though it is acceptable if it falls within the limits of 14.7 inches and 15.3 inches. Similarly, for Cocoa Fizz, the average bottle fill may be 16 ounces with tolerances of $\pm.2$ ounces. Although the bottles should be filled with 16 ounces of liquid, the amount can be as low as 15.8 or as high as 16.2 ounces.

Specifications for a product are preset on the basis of how the product is going to be used or what customer expectations are. As we have learned, any production process has a certain amount of natural variation associated with it. To be capable of producing an acceptable product, the process variation cannot exceed the preset specifications. Process capability thus involves evaluating process variability relative to preset product specifications in order to determine whether the process is capable of producing an acceptable product. In this section we will learn how to measure process capability.

Measuring Process Capability

Simply setting up control charts to monitor whether a process is in control does not guarantee process capability. To produce an acceptable product, the process must be *capable* and *in control* before production begins. Let's look at three examples of process variation relative to design specifications for the Cocoa Fizz soft drink company. Let's say that the specification for the acceptable volume of liquid is preset at 16 ounces $\pm.2$ ounces, which is 15.8 and 16.2 ounces. In part (a) of Figure 6-7 the process produces 99.74 percent (three sigma) of the product with volumes between 15.8 and 16.2 ounces. You can see that the process variability closely matches the preset specifications. Almost all the output falls within the preset specification range.

In part (b) of Figure 6-7, however, the process produces 99.74 percent (three sigma) of the product with volumes between 15.7 and 16.3 ounces. The process variability is outside the preset specifications. A large percentage of the product will fall outside the specified limits. This means that the process is *not capable* of producing the product within the preset specifications.

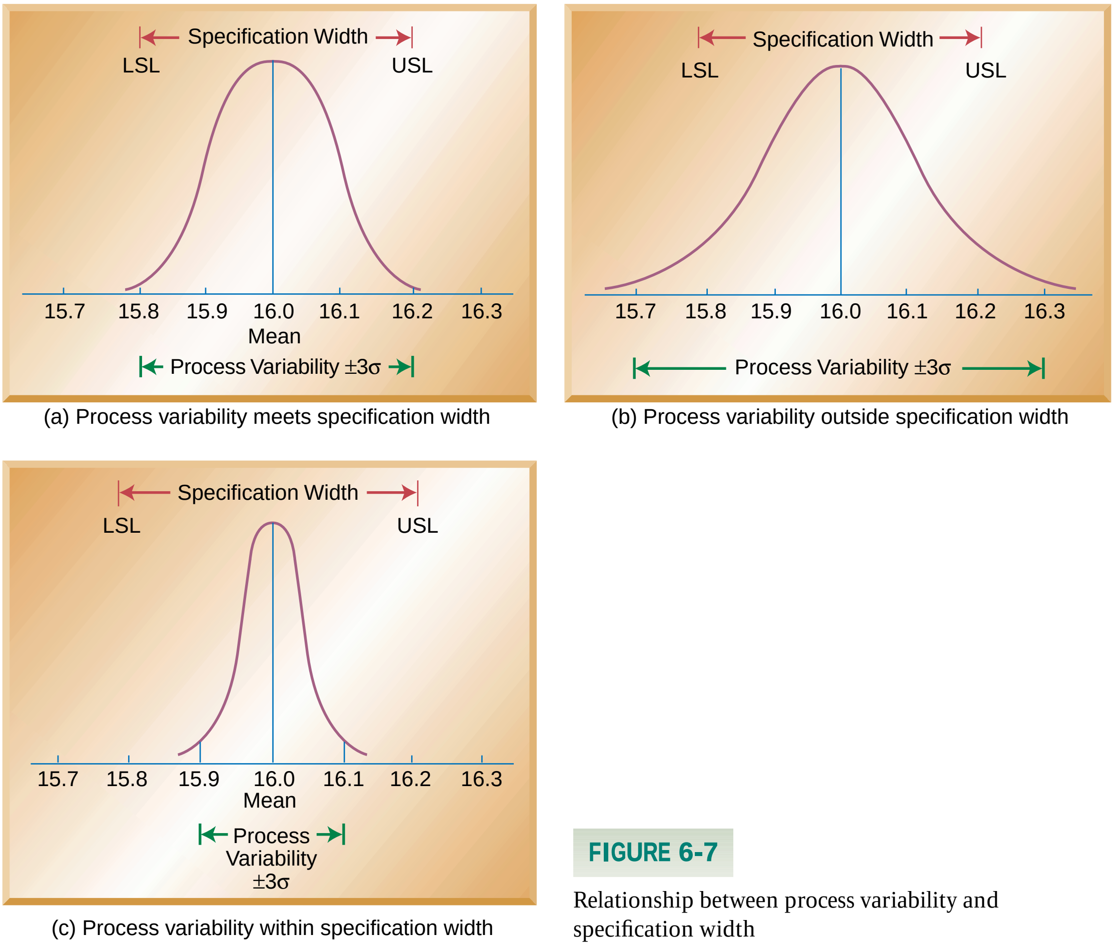
Part (c) of Figure 6-7 shows that the production process produces 99.74 percent (three sigma) of the product with volumes between 15.9 and 16.1 ounces. In this case the process variability is within specifications and the process exceeds the minimum capability.

Process capability is measured by the **process capability index,** C_p, which is computed as the ratio of the specification width to the width of the process variability:

Process capability index An index used to measure process capability.

$$C_p = \frac{\text{specification width}}{\text{process width}} = \frac{\text{USL} - \text{LSL}}{6\sigma}$$

where the specification width is the difference between the upper specification limit (USL) and the lower specification limit (LSL) of the process. The process width is



(a) Process variability meets specification width

(b) Process variability outside specification width

(c) Process variability within specification width

FIGURE 6-7

Relationship between process variability and specification width computed as 6 standard deviations (6σ) of the process being monitored. The reason we use 6σ is that most of the process measurement (99.74 percent) falls within ± 3 standard deviations, which is a total of 6 standard deviations.

There are three possible ranges of values for C_p that also help us interpret its value:

$C_p = 1$: A value of C_p equal to 1 means that the process variability just meets specifications, as in Figure 6-7(a). We would then say that the process is minimally capable.

$C_p \leq 1$: A value of C_p below 1 means that the process variability is outside the range of specification, as in Figure 6-7(b). This means that the process is not capable of producing within specification and the process must be improved.

$C_p \geq 1$: A value of C_p above 1 means that the process variability is tighter than specifications and the process exceeds minimal capability, as in Figure 6-7(c).

A C_p value of 1 means that 99.74 percent of the products produced will fall within the specification limits. This also means that .26 percent (100% − 99.74%) of the products will not be acceptable. Although this percentage sounds very small, when we think of it in terms of parts per million (ppm) we can see that it can still result in a lot of defects. The number .26 percent corresponds to 2600 parts per million (ppm) defective ($0.0026 \times 1{,}000{,}000$). That number can seem very high if we think of it in terms of 2600 wrong prescriptions out of a million, or 2600 incorrect medical procedures out of a million, or even 2600 malfunctioning aircraft out of a million. You can see that this number of defects is still high. The way to reduce the ppm defective is to increase process capability.

EXAMPLE 6.6

Computing the C_p Value at Cocoa Fizz

Three bottling machines at Cocoa Fizz are being evaluated for their capability:

Bottling Machine	Standard Deviation
A	.05
B	.1
C	.2

If specifications are set between 15.8 and 16.2 ounces, determine which of the machines are capable of producing within specifications.

• Solution

To determine the capability of each machine we need to divide the specification width ($USL - LSL = 16.2 - 15.8 = .4$) by 6σ for each machine:

Bottling Machine	σ	USL−LSL	6σ	$C_p = \dfrac{USL - LSL}{6\sigma}$
A	.05	.4	.3	1.33
B	.1	.4	.6	.67
C	.2	.4	1.2	.33

Looking at the C_p values, only machine A is capable of filling bottles within specifications, because it is the only machine that has a C_p value at or above 1.

C_p is valuable in measuring process capability. However, it has one shortcoming: it assumes that process variability is centered on the specification range. Unfortunately, this is not always the case. Figure 6-8 shows data from the Cocoa Fizz example. In the figure the specification limits are set between 15.8 and 16.2 ounces, with a mean of 16.0 ounces. However, the process variation is not centered; it has a mean of 15.9 ounces. Because of this, a certain proportion of products will fall outside the specification range.

The problem illustrated in Figure 6-8 is not uncommon, but it can lead to mistakes in the computation of the C_p measure. Because of this, another measure for process capability is used more frequently:

$$C_{pk} = \min\left(\frac{\text{USL} - \mu}{3\sigma}, \frac{\mu - \text{LSL}}{3\sigma}\right)$$

where μ = the mean of the process
σ = the standard deviation of the process

This measure of process capability helps us address a possible lack of centering of the process over the specification range. To use this measure, the process capability of each half of the normal distribution is computed and the minimum of the two is used.

Looking at Figure 6-8, we can see that the computed C_p is 1:

Process mean: $\mu = 15.9$

Process standard deviation $\sigma = 0.067$

$\text{LSL} = 15.8$

$\text{USL} = 16.2$

$$C_p = \frac{0.4}{6(0.067)} = 1$$

The C_p value of 1.00 leads us to conclude that the process is capable. However, from the graph you can see that the process is *not* centered on the specification range

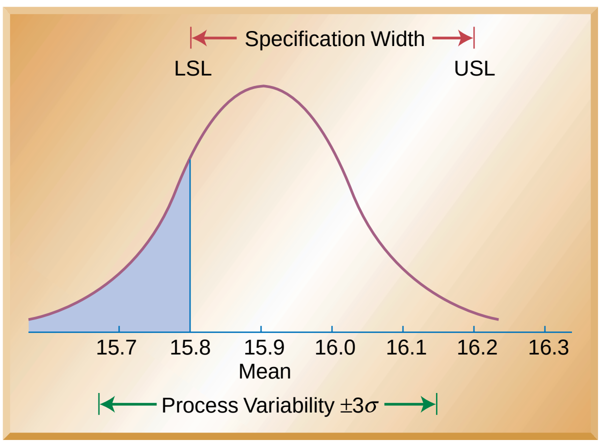


FIGURE 6-8

Process variability not centered across specification width and is producing out-of-spec products. Using only the C_p measure would lead to an incorrect conclusion in this case. Computing C_{pk} gives us a different answer and leads us to a different conclusion:

$$C_{pk} = \min\left(\frac{\text{USL} - \mu}{3\sigma}, \frac{\mu - \text{LSL}}{3\sigma}\right)$$

$$C_{pk} = \min\left(\frac{16.2 - 15.9}{3(.1)}, \frac{15.9 - 15.8}{3(.1)}\right)$$

$$C_{pk} = \min\,(1.00, 0.33)$$

$$C_{pk} = \frac{.1}{.3} = .33$$

The computed C_{pk} value is less than 1, revealing that the process is not capable.

EXAMPLE 6.7

Computing the C_{pk} Value

Compute the C_{pk} measure of process capability for the following machine and interpret the findings. What value would you have obtained with the C_p measure?

$$\begin{aligned} \text{Machine Data: USL} &= 110 \\ \text{LSL} &= 50 \\ \text{Process } \sigma &= 10 \\ \text{Process } \mu &= 70 \end{aligned}$$

• Solution

To compute the C_{pk} measure of process capability:

$$\begin{aligned} C_{pk} &= \min\left(\frac{\text{USL} - \mu}{3\sigma}, \frac{\mu - \text{LSL}}{3\sigma}\right) \\ &= \min\left(\frac{110 - 60}{3(10)}, \frac{60 - 50}{3(10)}\right) \\ &= \min\,(1.67, 0.33) \\ &= 0.33 \end{aligned}$$

This means that the process is not capable. The C_p measure of process capability gives us the following measure,

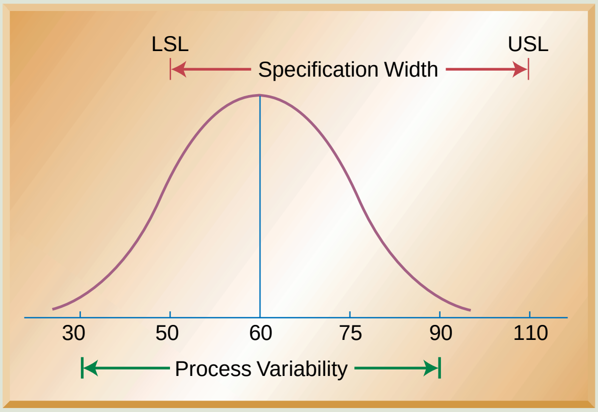
$$C_p = \frac{60}{6(10)} = 1$$

leading us to believe that the process is capable. The reason for the difference in the measures is that the process is not centered on the specification range, as shown in Figure 6-9.

Process capability of machines is a critical element of statistical process control.

FIGURE 6-9

Process variability not centered across specification width for Example 6.7



Six Sigma Quality

The term **Six Sigma®** was coined by the Motorola Corporation in the 1980s to describe the high level of quality the company was striving to achieve. Sigma (σ) stands for the number of standard deviations of the process. Recall that ± 3 sigma (σ) means that 2600 ppm are defective. The level of defects associated with Six Sigma is approximately 3.4 ppm. Figure 6-10 shows a process distribution with quality levels of ± 3 sigma (σ) and ± 6 sigma (σ). You can see the difference in the number of defects produced.

► **Six sigma quality**
A high level of quality associated with approximately 3.4 defective parts per million.

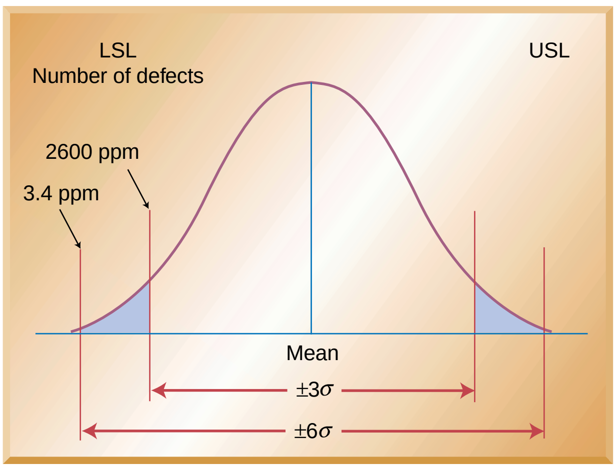


FIGURE 6-10

PPM defective for $\pm 3\sigma$ versus $\pm 6\sigma$ quality *(not to scale)*

LINKS TO PRACTICE

Motorola, Inc.
www.motorola.com

To achieve the goal of Six Sigma, Motorola has instituted a quality focus in every aspect of its organization. Before a product is designed, marketing ensures that product characteristics are exactly what customers want. Operations ensures that exact product characteristics can be achieved through product design, the manufacturing process, and the materials used. The Six Sigma concept is an integral part of other functions as well. It is used in the finance and accounting departments to reduce costing errors and the time required to close the books at the end of the month. Numerous other companies, such as General Electric and Texas Instruments, have followed Motorola's leadership and have also instituted the Six Sigma concept. In fact, the Six Sigma quality standard has become a benchmark in many industries.

There are two aspects to implementing the Six Sigma concept. The first is the use of technical tools to identify and eliminate causes of quality problems. These technical tools include the statistical quality control tools discussed in this chapter. They also include the problem-solving tools discussed in Chapter 5, such as cause-and-effect diagrams, flow charts, and Pareto analysis. In Six Sigma programs the use of these technical tools is integrated throughout the entire organizational system.

The second aspect of Six Sigma implementation is people involvement. In Six Sigma all employees have the training to use technical tools and are responsible for rooting out quality problems. Employees are given martial arts titles that reflect their skills in the Six Sigma process. *Black belts* and *master black belts* are individuals who have extensive training in the use of technical tools and are responsible for carrying out the implementation of Six Sigma. They are experienced individuals who oversee the measuring, analyzing, process controlling, and improving. They achieve this by acting as coaches, team leaders, and facilitators of the process of continuous improvement. *Green belts* are individuals who have sufficient training in technical tools to serve on teams or on small individual projects.

Successful Six Sigma implementation requires commitment from top company leaders. These individuals must promote the process, eliminate barriers to implementation, and ensure that proper resources are available. A key individual is a *champion* of Six Sigma. This is a person who comes from the top ranks of the organization and is responsible for providing direction and overseeing all aspects of the process.

ACCEPTANCE SAMPLING

Acceptance sampling, the third branch of statistical quality control, refers to the process of randomly inspecting a certain number of items from a lot or batch in order to decide whether to accept or reject the entire batch. What makes acceptance sampling different from statistical process control is that acceptance sampling is performed either *before* or *after* the process, rather than during the process. Acceptance sampling *before* the process involves sampling materials received from a supplier, such as randomly inspecting crates of fruit that will be used in a restaurant, boxes of glass dishes that will be sold in a department store, or metal castings that will be used in a machine shop. Sampling *after* the process involves sampling finished items that are to be shipped either to a customer or to a distribution center. Examples include randomly testing a certain number of computers from a batch to make sure they meet operational requirements, and randomly inspecting snowboards to make sure that they are not defective.

Sampling involves randomly inspecting items from a lot.

You may be wondering why we would only inspect some items in the lot and not the entire lot. Acceptance sampling is used when inspecting every item is not physically possible or would be overly expensive, or when inspecting a large number of items would lead to errors due to worker fatigue. This last concern is especially important when a large number of items are processed in a short period of time. Another example of when acceptance sampling would be used is in destructive testing, such as testing eggs for salmonella or vehicles for crash testing. Obviously, in these cases it would not be helpful to test every item! However, 100 percent inspection does make sense if the cost of inspecting an item is less than the cost of passing on a defective item.

As you will see in this section, the goal of acceptance sampling is to determine the criteria for acceptance or rejection based on the size of the lot, the size of the sample, and the level of confidence we wish to attain. Acceptance sampling can be used for both attribute and variable measures, though it is most commonly used for attributes. In this section we will look at the different types of sampling plans and at ways to evaluate how well sampling plans discriminate between good and bad lots.

Sampling Plans

A **sampling plan** is a plan for acceptance sampling that precisely specifies the parameters of the sampling process and the acceptance/rejection criteria. The variables to be specified include the size of the lot (N), the size of the sample inspected from the lot (n), the number of defects above which a lot is rejected (c), and the number of samples that will be taken.

► **Sampling plan**
A plan for acceptance sampling that precisely specifies the parameters of the sampling process and the acceptance/rejection criteria.

There are different types of sampling plans. Some call for *single sampling,* in which a random sample is drawn from every lot. Each item in the sample is examined and is labeled as either "good" or "bad." Depending on the number of defects or "bad" items found, the entire lot is either accepted or rejected. For example, a lot size of 50 cookies is evaluated for acceptance by randomly inspecting 10 cookies from the lot. The cookies may be inspected to make sure they are not broken or burned. If 4 or more of the 10 cookies inspected are bad, the entire lot is rejected. In this example, the lot size $N = 50$, the sample size $n = 10$, and the maximum number of defects at which a lot is accepted is $c = 4$. These parameters define the acceptance sampling plan.

Another type of acceptance sampling is called *double sampling.* This provides an opportunity to sample the lot a second time if the results of the first sample are inconclusive. In double sampling we first sample a lot of goods according to preset criteria for definite acceptance or rejection. However, if the results fall in the middle range, they are considered inconclusive and a second sample is taken. For example, a water treatment plant may sample the quality of the water ten times in random intervals throughout the day. Criteria may be set for acceptable or unacceptable water quality, such as .05 percent chlorine and .1 percent chlorine. However, a sample of water containing between .05 percent and .1 percent chlorine is inconclusive and calls for a second sample of water.

In addition to single and double-sampling plans, there are *multiple sampling plans.* Multiple sampling plans are similar to double sampling plans except that criteria are set for more than two samples. The decision as to which sampling plan to select has a great deal to do with the cost involved in sampling, the time consumed by sampling, and the cost of passing on a defective item. In general, if the cost of collecting a sample is relatively high, single sampling is preferred. An extreme example is collecting a biopsy from a hospital patient. Because the actual cost of getting the sample is high, we want to get a large sample and sample only once. The opposite is true when the cost of collecting the sample is low but the actual cost of testing is high. This may be the case with a water treatment plant, where collecting the water is inexpensive but the chemical analysis is costly. In this section we focus primarily on single sampling plans.

Operating Characteristic (OC) Curves

As we have seen, different sampling plans have different capabilities for discriminating between good and bad lots. At one extreme is 100 percent inspection, which has perfect discriminating power. However, as the size of the sample inspected decreases, so does the chance of accepting a defective lot. We can show the discriminating power of a sampling plan on a graph by means of an **operating characteristic (OC) curve.** This curve shows the probability or chance of accepting a lot given various proportions of defects in the lot.

▶ **Operating characteristic (OC) curve**
A graph that shows the probability or chance of accepting a lot given various proportions of defects in the lot.

Figure 6-11 shows a typical OC curve. The *x* axis shows the percentage of items that are defective in a lot. This is called "lot quality." The *y* axis shows the probability or chance of accepting a lot. You can see that if we use 100 percent inspection we are certain of accepting only lots with zero defects. However, as the proportion of defects in the lot increases, our chance of accepting the lot decreases. For example, we have a 90 percent probability of accepting a lot with 5 percent defects and an 80 percent probability of accepting a lot with 8 percent defects.

Regardless of which sampling plan we have selected, the plan is not perfect. That is, there is still a chance of accepting lots that are "bad" and rejecting "good" lots. The steeper the OC curve, the better our sampling plan is for discriminating between "good" and "bad." Figure 6-12 shows three different OC curves, A, B, and C. Curve A is the most discriminating and curve C the least. You can see that the steeper the slope of the curve, the more discriminating is the sampling plan. When 100 percent inspection is not possible, there is a certain amount of risk for consumers in accepting defective lots and a certain amount of risk for producers in rejecting good lots.

▶ **Acceptable quality level (AQL)**
The small percentage of defects that consumers are willing to accept.

▶ **Lot tolerance percent defective (LTPD)**
The upper limit of the percentage of defective items consumers are willing to tolerate.

There is a small percentage of defects that consumers are willing to accept. This is called the **acceptable quality level (AQL)** and is generally in the order of 1–2 percent. However, sometimes the percentage of defects that passes through is higher than the AQL. Consumers will usually tolerate a few more defects, but at some point the number of defects reaches a threshold level beyond which consumers will not tolerate them. This threshold level is called the **lot tolerance percent defective (LTPD).** The

FIGURE 6-11

Example of an operating characteristic (OC) curve

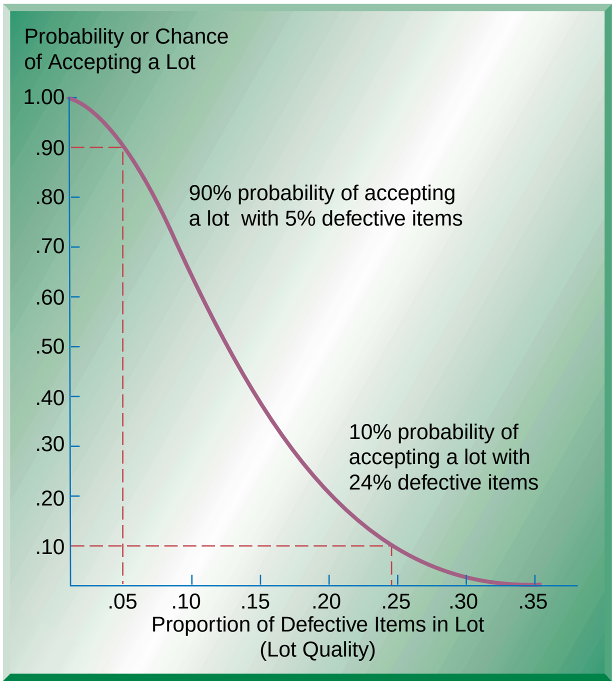
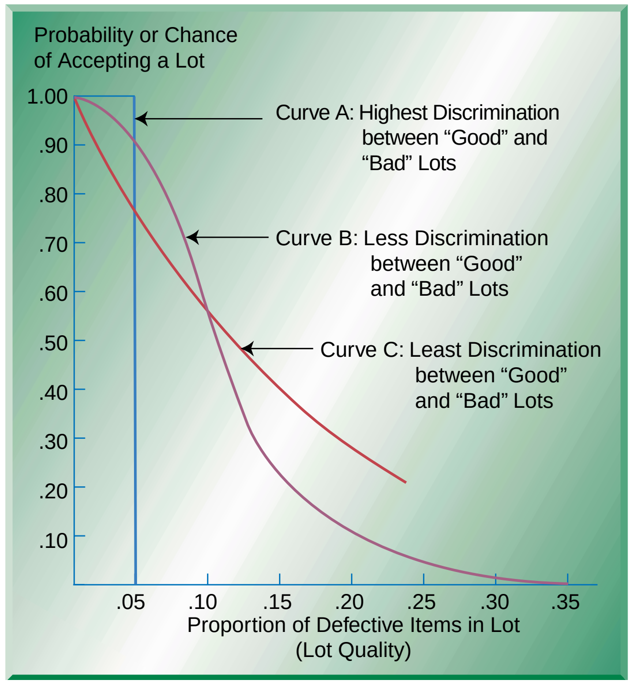


FIGURE 6-12

OC curves with different steepness levels and different levels of discrimination



LTPD is the upper limit of the percentage of defective items consumers are willing to tolerate.

Consumer's risk is the chance or probability that a lot will be accepted that contains a greater number of defects than the LTPD limit. This is the probability of making a Type II error—that is, accepting a lot that is truly “bad.” Consumer's risk or Type II error is generally denoted by beta (β). The relationships among AQL, LTPD, and β are shown in Figure 6-13. **Producer's risk** is the chance or probability that a lot containing an acceptable quality level will be rejected. This is the probability of making a Type I error—that is, rejecting a lot that is “good.” It is generally denoted by alpha (α). Producer's risk is also shown in Figure 6-13.

▶ **Consumer's risk** The chance of accepting a lot that contains a greater number of defects than the LTPD limit.

▶ **Producer's risk** The chance that a lot containing an acceptable quality level will be rejected.

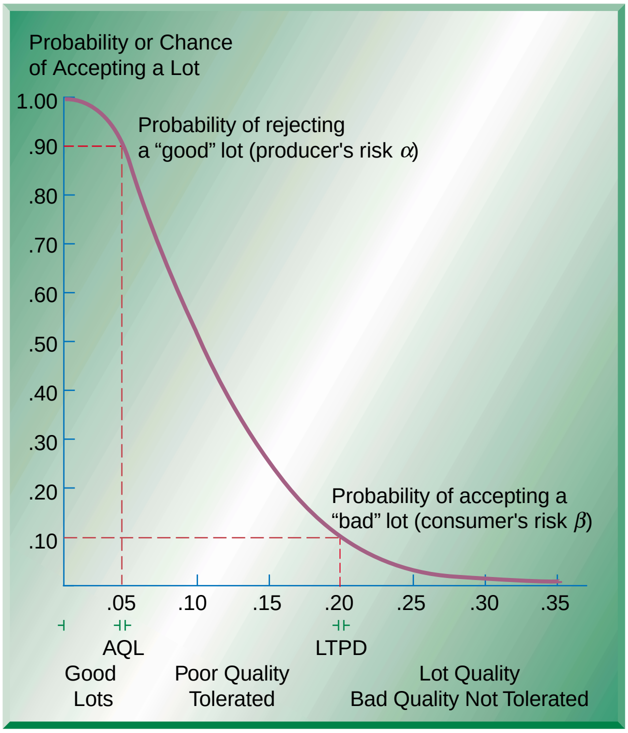
We can determine from an OC curve what the consumer's and producer's risks are. However, these values should not be left to chance. Rather, sampling plans are usually designed to meet specific levels of consumer's and producer's risk. For example, one common combination is to have a consumer's risk (β) of 10 percent and a producer's risk (α) of 5 percent, though many other combinations are possible.

Developing OC Curves

An OC curve graphically depicts the discriminating power of a sampling plan. To draw an OC curve, we typically use a cumulative binomial distribution to obtain

FIGURE 6-13

An OC curve showing producer's risk (α) and consumer's risk (β)



probabilities of accepting a lot given varying levels of lot defects.[1] The cumulative binomial table is found in Appendix C. A small part of this table is reproduced in Table 6-2. The top of the table shows values of *p*, which represents the proportion of defective items in a lot (5 percent, 10 percent, 20 percent, etc.). The left-hand column shows values of *n*, which represent the sample size being considered, and *x* represents the cumulative number of defects found. Let's use an example to illustrate how to develop an OC curve for a specific sampling plan using the information from Table 6-2.

TABLE 6-2

Partial Cumulative Binomial Probability Table

		Proportion of Items Defective (*p*)									
		.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
n	*x*										
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.9974	.9185	.8352	.7373	.6328	.5282	.4284	.3370	.2562	.1875
	2	.9988	.9914	.9734	.9421	.8965	.8369	.7648	.6826	.5931	.5000

[1]For $n \geq 20$ and $p \leq .05$ a Poisson distribution is generally used.

EXAMPLE 6.8

Constructing an OC Curve

Let's say that we want to develop an OC curve for a sampling plan in which a sample of $n = 5$ items is drawn from lots of $N = 1000$ items. The accept/reject criteria are set up in such a way that we accept a lot if *no more than* one defect ($c = 1$) is found.

• Solution

Let's look at the partial binomial distribution in Table 6-2. Since our criteria require us to sample $n = 5$, we will go to the row where n equals 5 in the left-hand column. The "x" column tells us the cumulative number of defects found at which we reject the lot. Since we are not allowing more than one defect, we look for an x value that corresponds to 1. The row corresponding to $n = 5$ and $x = 1$ tells us our chance or probability of accepting lots with various proportions of defects using this sampling plan. For example, with this sampling plan we have a 99.74% chance of accepting a lot with 5% defects. If we move down the row, we can see that we have a 91.85% chance of accepting a lot with 10% defects, a 83.52% chance of accepting a lot with 15% defects, and a 73.73% chance of accepting a lot with 20% defects. Using these values and those remaining in the row, we can construct an OC chart for $n = 5$ and $c = 1$. This is shown in Figure 6-14.

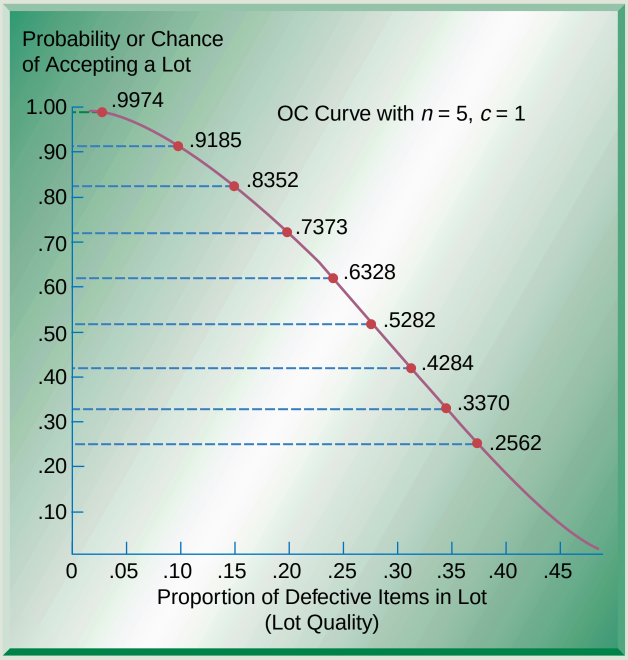


FIGURE 6-14

OC curve with $n = 5$ and $c = 1$

Average Outgoing Quality

As we observed with the OC curves, the higher the quality of the lot, the higher is the chance that it will be accepted. Conversely, the lower the quality of the lot, the greater is the chance that it will be rejected. Given that some lots are accepted and some rejected, it is useful to compute the **average outgoing quality (AOQ)** of lots to get a sense of the overall outgoing quality of the product. Assuming that all lots have the

► **Average outgoing quality (AOQ)**
The expected proportion of defective items that will be passed to the customer under the sampling plan.

same proportion of defective items, the average outgoing quality can be computed as follows:

$$\text{AOQ} = (P_{ac})p\left(\frac{N - n}{N}\right)$$

where P_{ac} = probability of accepting a given lot
p = proportion of defective items in a lot
N = the size of the lot
n = the sample size chosen for inspection

Usually we assume the fraction in the previous equation to equal 1 and simplify the equation to the following form:

$$\text{AOQ} = (P_{ac})p$$

We can then use the information from Figure 6-14 to construct an AOQ curve for different levels of probabilities of acceptance and different proportions of defects in a lot. As we will see, an AOQ curve is similar to an OC curve.

EXAMPLE 6.9

Constructing an AOQ Curve

Let's go back to our initial example, in which we sampled 5 items ($n = 5$) from a lot of 1000 ($N = 1000$) with an acceptance range of no more than 1($c = 1$) defect. Here we will construct an AOQ curve for this sampling plan and interpret its meaning.

• Solution

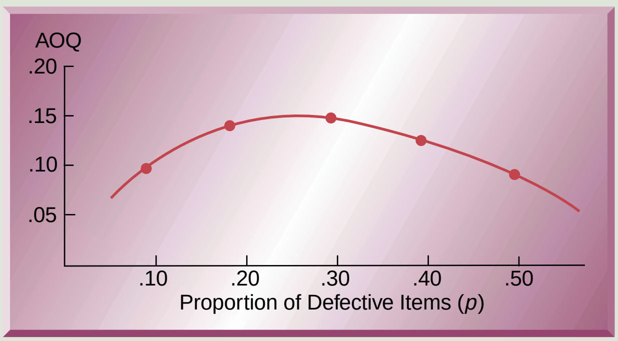
For the parameters $N = 1000$, $n = 5$, and $c = 1$, we can read the probabilities of P_{ac} from Figure 6-14. Then we can compute the value of AOQ as AOQ = $(P_{ac})\ p$.

p	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
P_{ac}	.9974	.9185	.8352	.7373	.6328	.5282	.4284	.3370	.2562	.1875
AOQ	.0499	.0919	.1253	.1475	.1582	.1585	.1499	.1348	.1153	.0938

Figure 6–15 shows a graphical representation of the AOQ values. The AOQ varies, depending on the proportion of defective items in the lot. The largest value of AOQ, called the average outgoing quality limit (AOQL), is around 15.85%. You can see from Figure 6-15 that the average outgoing quality

FIGURE 6-15

The AOQ for $n = 5$ and $c = 1$



will be high for lots that are either very good or very bad. For lots that have close to 30% of defective items, the AOQ is the highest. Managers can use this information to compute the worst possible value of their average outgoing quality given the proportion of defective items (p). Then this information can be used to develop a sampling plan with appropriate levels of discrimination.

IMPLICATIONS FOR MANAGERS

In this chapter we have learned about a variety of different statistical quality control (SQC) tools that help managers make decisions about product and process quality. However, to use these tools properly managers must make a number of decisions. In this section we discuss some of the most important decisions that must be made when implementing SPC.

How Much and How Often to Inspect

Consider Product Cost and Product Volume As you know, 100 percent inspection is rarely possible. The question then becomes one of how often to inspect in order to minimize the chances of passing on defects and still keep inspection costs manageable. This decision should be related to the *product cost* and *product volume* of what is being produced. At one extreme are high-volume, low-cost items, such as paper, pencils, nuts and bolts, for which 100 percent inspection would not be cost justified. Also, with such a large volume 100 percent inspection would not be possible because worker fatigue sets in and defects are often passed on. At the other extreme are low-volume, high-cost items, such as parts that will go into a space shuttle or be used in a medical procedure, that require 100 percent inspection.

Most items fall somewhere between the two extremes just described. For these items, frequency of inspection should be designed to consider the trade-off between the cost of inspection and the cost of passing on a defective item. Historically, inspections were set up to minimize these two costs. Today, it is believed that defects of any type should not be tolerated and that eliminating them helps reduce organizational costs. Still, the inspection process should be set up to consider issues of product cost and volume. For example, one company will probably have different frequencies of inspection for different products.

Consider Process Stability Another issue to consider when deciding how much to inspect is the stability of the process. Stable processes that do not change frequently do not need to be inspected often. On the other hand, processes that are unstable and change often should be inspected frequently. For example, if it has been observed that a particular type of drilling machine in a machine shop often goes out of tolerance, that machine should be inspected frequently. Obviously, such decisions cannot be made without historical data on process stability.

Consider Lot Size The size of the lot or batch being produced is another factor to consider in determining the amount of inspection. A company that produces a small number of large lots will have a smaller number of inspections than a company that produces a large number of small lots. The reason is that every lot should have some inspection, and when lots are large, there are fewer lots to inspect.

Where to Inspect

Since we cannot inspect every aspect of a process all the time, another important decision is to decide where to inspect. Some areas are less critical than others. Following are some points that are typically considered most important for inspection.

Inbound Materials Materials that are coming into a facility from a supplier or distribution center should be inspected before they enter the production process. It is important to check the quality of materials before labor is added to it. For example, it would be wasteful for a seafood restaurant not to inspect the quality of incoming lobsters only to later discover that its lobster bisque is bad. Another reason for checking inbound materials is to check the quality of sources of supply. Consistently poor quality in materials from a particular supplier indicates a problem that needs to be addressed.

Finished Products Products that have been completed and are ready for shipment to customers should also be inspected. This is the last point at which the product is in the production facility. The quality of the product represents the company's overall quality. The final quality level is what will be experienced by the customer, and an inspection at this point is necessary to ensure high quality in such aspects as fitness for use, packaging, and presentation.

Prior to Costly Processing During the production process it makes sense to check quality before performing a costly process on the product. If quality is poor at that point and the product will ultimately be discarded, adding a costly process will simply lead to waste. For example, in the production of leather armchairs in a furniture factory, chair frames should be inspected for cracks before the leather covering is added. Otherwise, if the frame is defective the cost of the leather upholstery and workmanship may be wasted.

Which Tools to Use

In addition to where and how much to inspect, managers must decide which tools to use in the process of inspection. As we have seen, tools such as control charts are best used at various points in the production process. Acceptance sampling is best used for inbound and outbound materials. It is also the easiest method to use for attribute measures, whereas control charts are easier to use for variable measures. Surveys of industry practices show that most companies use control charts, especially x-bar and R-charts, because they require less data collection than p-charts.

STATISTICAL QUALITY CONTROL IN SERVICES

Statistical quality control (SQC) tools have been widely used in manufacturing organizations for quite some time. Manufacturers such as Motorola, General Electric, Toyota, and others have shown leadership in SQC for many years. Unfortunately, service organizations have lagged behind manufacturing firms in their use of SQC. The primary reason is that statistical quality control requires measurement, and it is difficult to measure the quality of a service. Remember that services often provide an intangible product and that perceptions of quality are often highly subjective. For example, the quality of a service is often judged by such factors as friendliness and courtesy of the staff and promptness in resolving complaints.

A way to measure the quality of services is to devise quantifiable measurements of the important dimensions of a particular service. For example, the number of complaints received per month, the number of telephone rings after which a response is received, or customer waiting time can be quantified. These types of measurements are not subjective or subject to interpretation. Rather, they can be measured and recorded. As in manufacturing, acceptable control limits should be developed and the variable in question should be measured periodically.

Another issue that complicates quality control in service organizations is that the service is often consumed during the production process. The customer is often present during service delivery, and there is little time to improve quality. The workforce that interfaces with customers is part of the service delivery. The way to manage this issue is to provide a high level of workforce training and to empower workers to make decisions that will satisfy customers.

LINKS TO PRACTICE

The Ritz-Carlton Hotel Company, L.L.C.
www.ritzcarlton.com

Nordstrom, Inc.
www.nordstrom.com

One service organization that has demonstrated quality leadership is The Ritz-Carlton Hotel Company. This luxury hotel chain caters to travelers who seek high levels of customer service. The goal of the chain is to be recognized for outstanding service quality. To this end, computer records are kept of regular clients' preferences. To keep customers happy, employees are empowered to spend up to $2,000 on the spot to correct any customer complaint. Consequently, The Ritz-Carlton has received a number of quality awards including winning the Malcolm Baldrige National Quality Award twice. It is the only company in the service category to do so.

Another leader in service quality that uses the strategy of high levels of employee training and empowerment is Nordstrom Department Stores. Outstanding customer service is the goal of this department store chain. Its organizational chart places the customer at the head of the organization. Records are kept of regular clients' preferences, and employees are empowered to make decisions on the spot to satisfy customer wants. The customer is considered to always be right.

LINKS TO PRACTICE

Marriott International, Inc.
www.marriott.com

Service organizations, must also use statistical tools to measure their processes and monitor performance. For example, the Marriott is known for regularly collecting data in the form of guest surveys. The company randomly surveys as many as a million guests each year. The collected data is stored in a large database and continually examined for patterns, such as trends and changes in customer preferences. Statistical techniques are used to analyze the data and provide important information, such as identifying areas that have the highest impact on performance, and those areas that need improvement. This information allows Marriott to provide a superior level of customer service, anticipate customer demands, and put resources in service features most important to customers.

OM ACROSS THE ORGANIZATION

It is easy to see how operations managers can use the tools of SQC to monitor product and process quality. However, you may not readily see how these statistical techniques affect other functions of the organization. In fact, SQC tools require input from other functions, influence their success, and are actually used by other organizational functions in designing and evaluating their tasks.

Marketing plays a critical role in setting up product and service quality standards. It is up to marketing to provide information on current and future quality standards required by customers and those being offered by competitors. Operations managers can incorporate this information into product and process design. Consultation with marketing managers is essential to ensure that quality standards are being met. At the same time, meeting quality standards is essential to the marketing department, since sales of products are dependent on the standards being met.

Finance is an integral part of the statistical quality control process, because it is responsible for placing financial values on SQC efforts. For example, the finance department evaluates the dollar costs of defects, measures financial improvements that result from tightening of quality standards, and is actively involved in approving investments in quality improvement efforts.

Human resources becomes even more important with the implementation of TQM and SQC methods, as the role of workers changes. To understand and utilize SQC tools, workers need ongoing training and the ability to work in teams, take pride in their work, and assume higher levels of responsibility. The human resources department is responsible for hiring workers with the right skills and setting proper compensation levels.

Information systems is a function that makes much of the information needed for SQC accessible to all who need it. Information systems managers need to work closely with other functions during the implementation of SQC so that they understand exactly what types of information are needed and in what form. As we have seen, SQC tools are dependent on information, and it is up to information systems managers to make that information available. As a company develops ways of using TQM and SQC tools, information systems managers must be part of this ongoing evolution to ensure that the company's information needs are being met.

All functions need to work closely together in the implementation of statistical process control. Everyone benefits from this collaborative relationship: operations is able to produce the right product efficiently; marketing has the exact product customers are looking for; and finance can boast of an improved financial picture for the organization.

SQC also affects various organizational functions through its direct application in evaluating quality performance in all areas of the organization. SQC tools are not used only to monitor the production process and ensure that the product being produced is within specifications. As we have seen in the Motorola Six Sigma example, these tools can be used to monitor both quality levels and defects in accounting procedures, financial record keeping, sales and marketing, office administration, and other functions. Having high quality standards in operations does not guarantee high quality in the organization as a whole. The same stringent standards and quality evaluation procedures should be used in setting standards and evaluating the performance of all organizational functions.

INSIDE OM

The decision to increase the level of quality standard and reduce the number of product defects requires support from every function within operations management. Two areas of operations management that are particularly affected are product and process design. Process design needs to be modified to incorporate customer-defined quality and simplification of design. Processes need to be continuously monitored and changed to build quality into the process and reduce variation. Other areas that are affected are job design, as we expand the role of employees to become responsible for monitoring quality levels and to use statistical quality control tools. Supply chain management and inventory control are also affected as we increase quality standard requirements from our suppliers and change the materials we use. All areas of operations management are involved when increasing the quality standard of a firm.

Chapter Highlights

1. Statistical quality control (SQC) refers to statistical tools that can be used by quality professionals. Statistical quality control can be divided into three broad categories: descriptive statistics, acceptance sampling, and statistical process control (SPC).
2. Descriptive statistics are used to describe quality characteristics, such as the mean, range, and variance. Acceptance sampling is the process of randomly inspecting a sample of goods and deciding whether to accept or reject the entire lot. Statistical process control (SPC) involves inspecting a random sample of output from a process and deciding whether the process is producing products with characteristics that fall within preset specifications.
3. There are two causes of variation in the quality of a product or process: common causes and assignable causes. Common causes of variation are random causes that we cannot identify. Assignable causes of variation are those that can be identified and eliminated.
4. A control chart is a graph used in statistical process control that shows whether a sample of data falls within the normal range of variation. A control chart has upper and lower control limits that separate common from assignable causes of variation. Control charts for variables monitor characteristics that can be measured and have a continuum of values, such as height, weight, or volume. Control charts for attributes are used to monitor characteristics that have discrete values and can be counted.
5. Control charts for variables include x-bar charts and R-charts. X-bar charts monitor the mean or average value of a product characteristic. R-charts monitor the range or dispersion of the values of a product characteristic. Control charts for attributes include p-charts and c-charts. P-charts are used to monitor the proportion of defects in a sample. C-charts are used to monitor the actual number of defects in a sample.
6. Process capability is the ability of the production process to meet or exceed preset specifications. It is measured by the process capability index, C_p, which is computed as the ratio of the specification width to the width of the process variability.
7. The term *Six Sigma* indicates a level of quality in which the number of defects is no more than 3.4 parts per million.
8. The goal of acceptance sampling is to determine criteria for acceptance or rejection based on lot size, sample size, and the desired level of confidence. Operating characteristic (OC) curves are graphs that show the discriminating power of a sampling plan.
9. It is more difficult to measure quality in services than in manufacturing. The key is to devise quantifiable measurements for important service dimensions.

Key Terms

statistical quality control (SQC) 172
descriptive statistics 172
statistical process control (SPC) 173
acceptance sampling 173
common causes of variation 174
assignable causes of variation 174
mean (average) 174
range 175
standard deviation 175
control chart 176
out of control 176
variable 177
attribute 177
x-bar chart 178
R-chart 182
p-chart 185
c-chart 188
process capability 190
product specifications 190
process capability index 191
Six Sigma quality 195
sampling plan 197
operating characteristic (OC) curve 198
acceptable quality level (AQL) 198
lot tolerance percent defective (LTPD) 198
consumer's risk 199
producer's risk 199
average outgoing quality (AOQ) 201

Formula Review

1. Mean $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

2. Standard Deviation $\sigma = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

3. Control Limits for x-Bar Charts

Upper control limit (UCL) $= \bar{\bar{x}} + z\sigma_{\bar{x}}$

Lower control limit (LCL) $= \bar{\bar{x}} - z\sigma_{\bar{x}}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

4. Control Limits for x-Bar Charts Using Sample Range as an Estimate of Variability

Upper control limit (UCL) $= \bar{\bar{x}} + A_2\bar{R}$

Lower control limit (LCL) $= \bar{\bar{x}} - A_2\bar{R}$

5. Control Limits for R-Charts $\text{UCL} = D_4\bar{R}$

$\text{LCL} = D_3\bar{R}$

6. Control Limits for p-Charts $\text{UCL} = \bar{p} + z(\sigma_p)$

$\text{LCL} = \bar{p} - z(\sigma_p)$

7. Control Limits for c-Charts $\text{UCL} = \bar{c} + z\sqrt{\bar{c}}$

$\text{LCL} = \bar{c} + z\sqrt{\bar{c}}$

8. Measures for Process Capability

$$C_p = \frac{\text{specification width}}{\text{process width}} = \frac{\text{USL} - \text{LSL}}{6\sigma}$$

$$C_{pk} = \min\left(\frac{\text{USL} - \mu}{3\sigma}, \frac{\mu - \text{LSL}}{3\sigma}\right)$$

9. Average Outgoing Quality $\text{AOQ} = (P_{ac})p$

Solved Problems

• Problem 1

A quality control inspector at the Crunchy Potato Chip Company has taken 3 samples with 4 observations each of the volume of bags filled. The data and the computed means are shown in the following table:

Sample of Potato Chip Bag Volume in Ounces

Sample Number	Observations 1	2	3	4
1	12.5	12.3	12.6	12.7
2	12.8	12.4	12.4	12.8
3	12.1	12.6	12.5	12.4
4	12.2	12.6	12.5	12.3
5	12.4	12.5	12.5	12.5
6	12.3	12.4	12.6	12.6
7	12.6	12.7	12.5	12.8
8	12.4	12.3	l2.6	12.5
9	12.6	12.5	l2.3	12.6
10	12.1	12.7	12.5	12.8
Mean $\bar{x}$	12.4	12.5	12.5	12.6

If the standard deviation of the bagging operation is 0.2 ounces, use the information in the table to develop control limits of 3 standard deviations for the bottling operation.

• Solution

The center line of the control data is the average of the samples:

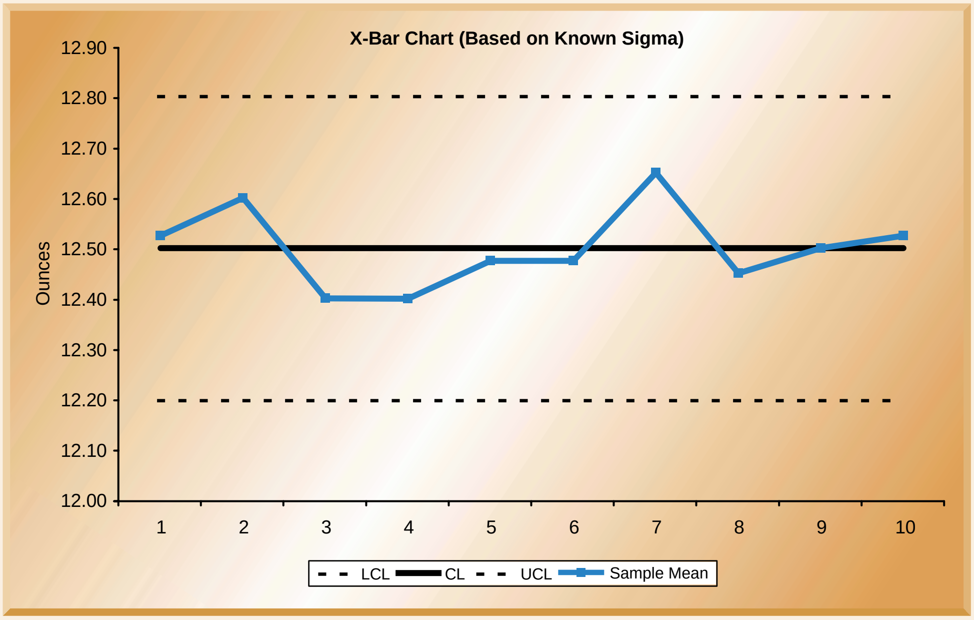
$$\bar{\bar{x}} = \frac{12.4 + 12.5 + 12.5 + 12.6}{4} = 12.5 \text{ ounces}$$

The control limits are:

$$\text{UCL} = \bar{\bar{x}} + z\sigma_{\bar{x}} = 12.5 + 3\left(\frac{.2}{\sqrt{4}}\right) = 12.80$$

$$\text{LCL} = \bar{\bar{x}} - z\sigma_{\bar{x}} = 12.5 - 3\left(\frac{.2}{\sqrt{4}}\right) = 12.20$$

Following is the associated control chart:



The problem can also be solved using a spreadsheet.

	A	B	C	D	E	F	G
1							
2	**Crunchy Potato Chips Company**						
3							
4							
5		**Bottle Volume in Ounces**					
6	Sample Num	Obs 1	Obs 2	Obs 3	Obs 4	Average	
7	1	12.50	12.30	12.60	12.70	12.53	
8	2	12.80	12.40	12.40	12.80	12.60	
9	3	12.10	12.60	12.50	12.40	12.40	
10	4	12.20	12.60	12.50	12.30	12.40	
11	5	12.40	12.50	12.50	12.50	12.48	
12	6	12.30	12.40	12.60	12.60	12.48	
13	7	12.60	12.70	12.50	12.80	12.65	
14	8	12.40	12.30	12.60	12.50	12.45	
15	9	12.60	12.50	12.30	12.60	12.50	
16	10	12.10	12.70	12.50	12.80	12.53	
17						12.50	
18		Number of Samples		10		Xbar-bar	
19	Number of Observations per Sample			4			
20							
21							
22	**Computations for X-Bar Chart**						
23			Overall Mean (Xbar-bar) =	12.50			
24			Sigma for Process =	0.2	ounces		
25			Standard Error of the Mean =	0.1			
26			Z-value for control charts =	3			
27							
28			CL: Center Line =	**12.50**			
29			LCL: Lower Control Limit =	**12.20**			
30			UCL: Upper Control Limit =	**12.80**			

F7: =AVERAGE(B7:E7)

F17: =AVERAGE(F7:F16)

D23: =F17

D25: =D24/SQRT(D19)

D28: =D23

D29: =D23-D26*D25

D30: =D23+D26*D25

• Problem 2

Use of the sample range to estimate variability can also be applied to the Crunchy Potato Chip operation. A quality control inspector has taken 4 samples with 5 observations each, measuring the volume of chips per bag. If the average range for the 4 samples is .2 ounces and the average mean of the observations is 12.5 ounces, develop three-sigma control limits for the bottling operation.

• *Solution*

$$\bar{\bar{x}} = 12.5 \text{ ounces}$$

$$\bar{R} = .2$$

The value of A_2 is obtained from Table 6-1. For $n = 5$, $A_2 = .58$. This leads to the following limits:

The center of the control chart is CL = 12.5 ounces

$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 12.5 + (.58)(.2) = 12.62$$

$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 12.5 - (.58)(.2) = 12.38$$

• Problem 3

Ten samples with 5 observations each have been taken from the Crunchy Potato Chip Company plant in order to test for volume dispersion in the bagging process. The average sample range was found to be .3 ounces. Develop control limits for the sample range.

• *Solution*

$$\bar{R} = .3 \text{ ounces}$$

$$n = 5$$

From Table 6-1 for $n = 5$:

$$D_4 = 2.11$$

$$D_3 = 0$$

Therefore,

$$\text{UCL} = D_4\bar{R} = 2.11(.3) = .633$$

$$\text{LCL} = D_3\bar{R} = 0(.3) = 0$$

• Problem 4

A production manager at a light bulb plant has inspected the number of defective light bulbs in 10 random samples with 30 observations each. Following are the numbers of defective light bulbs found:

Sample	Number Defective	Number of Observations in Sample
1	1	30
2	3	30
3	3	30
4	1	30
5	0	30
6	5	30
7	1	30
8	1	30
9	1	30
10	1	30
Total	17	300

Construct a three-sigma control chart ($z = 3$) with this information.

• *Solution*

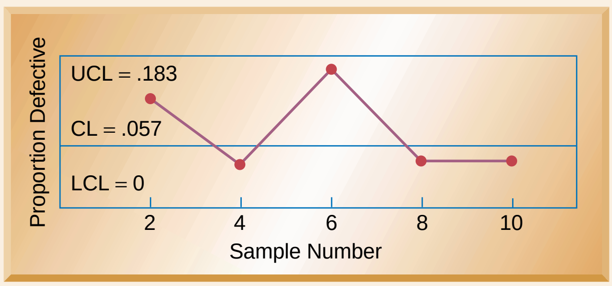
The center line of the chart is:

$$\text{CL} = \bar{p} = \frac{\text{number defective}}{\text{number of observations}} = \frac{17}{300} = .057$$

$$\sigma_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{(.057)(.943)}{30}} = .042$$

$$\text{UCL} = \bar{p} + z(\sigma_p) = .057 + 3(.042) = .183$$

$$\text{LCL} = \bar{p} - z(\sigma_p) = .057 - 3(.042) = -.069 \longrightarrow 0$$



This is also solved using a spreadsheet.

	A	B	C	D	E	F	G
1							
2	**p-Chart for Light Bulb Quality**						
3							
4		Sample Size	30				
5		Number Samples	10				
6							
7	**Sample #**	**# Defectives**	**p**	C8: =B8/C$4			
8	1	1	0.03333333				
9	2	3	0.1				
10	3	3	0.1				
11	4	1	0.03333333				
12	5	0	0				
13	6	5	0.16666667				
14	7	1	0.03333333				
15	8	1	0.03333333				
16	9	1	0.03333333				
17	10	1	0.03333333				
18				C19: =SUM(B8:B17)/(C4*C5)			
19		p bar =	0.05666667	C20: =SQRT((C19*(1-C19))/C4)			
20		Sigma_p =	0.04221199				
21	Z-value for control charts =		3				
22				C23: =C19			
23	CL: Center Line =		**0.05666667**	C24: =MAX(C$19-C$21*C$20,0)			
24	LCL: Lower Control Limit =		**0**	C25: =C$19+C$21*C$20			
25	UCL: Upper Control Limit =		**0.18330263**				

• Problem 5

Kinder Land Child Care uses a c-chart to monitor the number of customer complaints per week. Complaints have been recorded over the past 20 weeks. Develop a control chart with three-sigma control limits using the following data:

Week	Number of Complaints	Week	Number of Complaints
1	0	11	4
2	3	12	3
3	4	13	1
4	1	14	1
5	0	15	1
6	0	16	0
7	3	17	2
8	1	18	1
9	1	19	2
10	0	20	2
		Total	30

• *Solution*

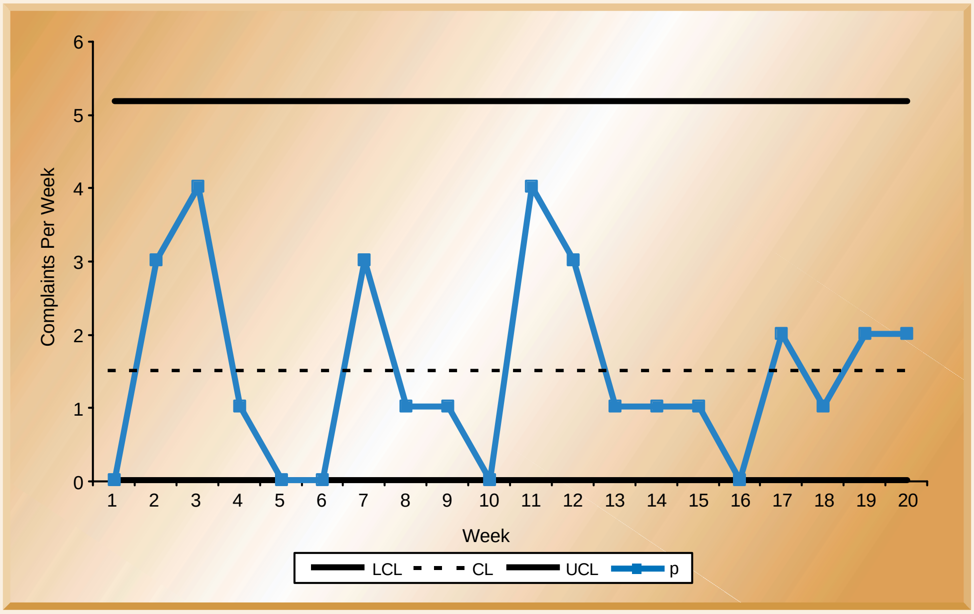
The average weekly number of complaints is $\frac{30}{20} = 1.5$

Therefore,

$$\text{UCL} = \bar{c} + z\sqrt{\bar{c}} = 1.5 + 3\sqrt{1.5} = 5.17$$

$$\text{LCL} = \bar{c} - z\sqrt{\bar{c}} = 1.5 - 3\sqrt{1.5} = -2.17 \longrightarrow 0$$

The resulting control chart is:



• Problem 6

Three bagging machines at the Crunchy Potato Chip Company are being evaluated for their capability. The following data are recorded:

Bagging Machine	Standard Deviation
A	.2
B	.3
C	.05

If specifications are set between 12.35 and 12.65 ounces, determine which of the machines are capable of producing within specification.

• Solution

To determine the capability of each machine we need to divide the specification width (USL − LSL = 12.65 − 12.35 = .3) by 6σ for each machine:

Bagging Machine	σ	USL − LSL	6σ	$C_p = \frac{\text{USL} - \text{LSL}}{6\sigma}$
A	.2	.3	1.2	0.25
B	.3	.3	1.8	0.17
C	.05	.3	.3	1.00

Looking at the C_p values, only machine C is capable of bagging the potato chips within specifications, because it is the only machine that has a C_p value at or above 1.

• Problem 7

Compute the C_{pk} measure of process capability for the following machine and interpret the findings. What value would you have obtained with the C_p measure?

Machine Data: USL = 80
LSL = 50
Process σ = 5
Process μ = 60

• Solution

To compute the C_{pk} measure of process capability:

$$C_{pk} = \min\left(\frac{\text{USL} - \mu}{3\sigma}, \frac{\mu - \text{LSL}}{3\sigma}\right)$$

$$= \min\left(\frac{80 - 60}{3(5)}, \frac{60 - 50}{3(5)}\right)$$

$$= \min(1.33, 0.67)$$

$$= 0.67$$

This means that the process is not capable. The C_p measure of process capability gives us the following measure:

$$C_p = \frac{30}{6(5)} = 1.0$$

which leads us to believe that the process is capable.

Discussion Questions

1. Explain the three categories of statistical quality control (SQC). How are they different, what different information do they provide, and how can they be used together?

2. Describe three recent situations in which you were directly affected by poor product or service quality.

3. Discuss the key differences between common and assignable causes of variation. Give examples.

4. Describe a quality control chart and how it can be used. What are upper and lower control limits? What does it mean if an observation falls outside the control limits?

5. Explain the differences between x-bar and R-charts. How can they be used together and why would it be important to use them together?

6. Explain the use of p-charts and c-charts. When would you use one rather than the other? Give examples of measurements for both p-charts and c-charts.

7. Explain what is meant by process capability. Why is it important? What does it tell us? How can it be measured?

8. Describe the process of acceptance sampling. What types of sampling plans are there? What is acceptance sampling used for?

9. Describe the concept of Six Sigma quality. Why is such a high quality level important?

Problems

1. A quality control manager at a manufacturing facility has taken 4 samples with 4 observations each of the diameter of a part.
   (a) Compute the mean of each sample.
   (b) Compute an estimate of the mean and standard deviation of the sampling distribution.
   (c) Develop control limits for 3 standard deviations of the product diameter.

Samples of Part Diameter in Inches

1	2	3	4
5.8	6.2	6.1	6.0
5.9	6.0	5.9	5.9
6.0	5.9	6.0	5.9
6.1	5.9	5.8	6.1

2. A quality control inspector at the Beautiful Shampoo Company has taken 3 samples with 4 observations each of the volume of shampoo bottles filled. The data collected by the inspector and the computed means are shown here:

Samples of Shampoo Bottle Volume in Ounces

Observation	1	2	3
1	19.7	19.7	19.7
2	20.6	20.2	18.7
3	18.9	18.9	21.6
4	20.8	20.7	20.0
Mean	20.0	19.875	20.0

If the standard deviation of the shampoo bottle filling operation is .2 ounces, use the information in the table to develop control limits of 3 standard deviations for the operation.

3. A quality control inspector has taken 4 samples with 5 observations each at the Beautiful Shampoo Company, measuring the volume of shampoo per bottle. If the average range for the 4 samples is .4 ounces and the average mean of the observations is 19.8 ounces, develop three sigma control limits for the bottling operation.

4. A production manager at Ultra Clean Dishwashing company is monitoring the quality of the company's production process. There has been concern relative to the quality of the operation to accurately fill the 16 ounces of dishwashing liquid. The product is designed for a fill level of 16.00 ± 0.30. The company collected the following sample data on the production process:

	Observations			
Sample	1	2	3	4
1	16.40	16.11	15.90	15.78
2	15.97	16.10	16.20	15.81
3	15.91	16.00	16.04	15.92
4	16.20	16.21	15.93	15.95
5	15.87	16.21	16.34	16.43
6	15.43	15.49	15.55	15.92
7	16.43	16.21	15.99	16.00
8	15.50	15.92	l6.12	16.02
9	16.13	16.21	16.05	16.01
10	15.68	16.43	16.20	15.97

   (a) Are the process mean and range in statistical control?
   (b) Do you think this process is capable of meeting the design standard?

5. Ten samples with 5 observations each have been taken from the Beautiful Shampoo Company plant in order to test for volume dispersion in the shampoo bottle filling process. The average sample range was found to be .3 ounces. Develop control limits for the sample range.

6. The Awake Coffee Company produces gourmet instant coffee. The company wants to be sure that the average fill of coffee containers is 12.0 ounces. To make sure the process is in control, a worker periodically selects at random a box containing 6 containers of coffee and measures their weight. When the process is in control, the range of the weight of coffee samples averages .6 ounces.
   (a) Develop an R-chart and an $\bar{x}$-chart for this process.
   (b) The measurements of weight from the last five samples taken of the 6 containers are shown below:

Is the process in control? Explain your answer.

Sample	$\bar{x}$	R
1	12.1	.7
2	11.8	.4
3	12.3	.6
4	11.5	.4
5	11.6	.9

7. A production manager at a Contour Manufacturing plant has inspected the number of defective plastic molds in 5 random samples of 20 observations each. Following are the number of defective molds found in each sample:

Sample	Number of Defects	Number of Observations in Sample
1	1	20
2	2	20
3	2	20
4	1	20
5	0	20
Total	6	100

Construct a three-sigma control chart ($z = 3$) with this information.

8. A tire manufacturer has been concerned about the number of defective tires found recently. In order to evaluate the true magnitude of the problem, a production manager selected ten random samples of 20 units each for inspection. The number of defective tires found in each sample are as follows:

(a) Develop a p-chart with a $z = 3$.

(b) Suppose that the next 4 samples selected had 6, 3, 3, and 4 defects. What conclusion can you make?

Sample	Number Defective
1	1
2	3
3	2
4	1
5	4
6	1
7	2
8	0
9	3
10	1

9. U-learn University uses a c-chart to monitor student complaints per week. Complaints have been recorded over the past 10 weeks. Develop three-sigma control limits using the following data:

Week	Number of Complaints
1	0
2	3
3	1
4	1
5	0
6	0
7	3
8	1
9	1
10	2

10. University Hospital has been concerned with the number of errors found in its billing statements to patients. An audit of 100 bills per week over the past 12 weeks revealed the following number of errors:

Week	Number of Errors
1	4
2	5
3	6
4	6
5	3
6	2
7	6
8	7
9	3
10	4
11	4
12	4

(a) Develop control charts with $z = 3$.

(b) Is the process in control?

11. Three ice cream packing machines at the Creamy Treat Company are being evaluated for their capability. The following data are recorded:

Packing Machine	Standard Deviation
A	.2
B	.3
C	.05

If specifications are set between 15.8 and 16.2 ounces, determine which of the machines are capable of producing within specifications.

12. Compute the C_{pk} measure of process capability for the following machine and interpret the findings. What value would you have obtained with the C_p measure?

Machine Data: USL $= 100$
LSL $= 70$
Process $\sigma = 5$
Process $\mu = 80$

13. Develop an OC curve for a sampling plan in which a sample of $n = 5$ items is drawn from lots of $N = 1000$ items. The accept/reject criteria are set up in such a way that we accept a lot if no more than one defect ($c = 1$) is found.

14. Quality Style manufactures self-assembling furniture. To reduce the cost of returned orders, the manager of its quality control department inspects the final packages each day using randomly selected samples. The defects include wrong parts, missing connection parts, parts with apparent painting problems, and parts with rough surfaces. The average defect rate is three per day.

(a) Which type of control chart should be used? Construct a control chart with three-sigma control limits.
(b) Today the manager discovered nine defects. What does this mean?

15. Develop an OC curve for a sampling plan in which a sample of $n = 10$ items is drawn from lots of $N = 1000$. The accept/reject criteria is set up in such a way that we accept a lot if no more than one defect ($c = 1$) is found.

16. The Fresh Pie Company purchases apples from a local farm to be used in preparing the filling for their apple pies. Sometimes the apples are fresh and ripe. Other times they can be spoiled or not ripe enough. The company has decided that they need an acceptance sampling plan for the purchased apples. Fresh Pie has decided that the acceptable quality level is 5 defective apples per 100, and the lot tolerance proportion defective is 5%. Producer's risk should be no more than 5% and consumer's risk 10% or less.

(a) Develop a plan that satisfies the above requirements.
(b) Determine the AOQL for your plan, assuming that the lot size is 1000 apples.

17. A computer manufacturer purchases microchips from a world-class supplier. The buyer has a lot tolerance proportion defective of 10 parts in 5000, with a consumer's risk of 15%. If the computer manufacturer decides to sample 2000 of the microchips received in each shipment, what acceptance number, c, would they want?

18. Joshua Simms has recently been placed in charge of purchasing at the Med-Tech Labs, a medical testing laboratory. His job is to purchase testing equipment and supplies. Med-Tech currently has a contract with a reputable supplier in the industry. Joshua's job is to design an appropriate acceptance sampling plan for Med-Tech. The contract with the supplier states that the acceptable quality level is 1% defective. Also, the lot tolerance proportion defective is 4%, the producer's risk is 5%, and the consumer's risk is 10%.

(a) Develop an acceptance sampling plan for Joshua that meets the stated criteria.
(b) Draw the OC curve for the plan you developed.
(c) What is the AOQL of your plan, assuming a lot size of 1000?

19. Breeze Toothpaste Company makes tubes of toothpaste. The product is produced and then pumped into tubes and capped. The production manager is concerned whether the filling process for the tubes of toothpaste is in statistical control. The process should be centered on 6 ounces per tube. Six samples of 5 tubes were taken and each tube was weighed. The weights are:

	Ounces of Toothpaste per Tube				
Sample	**1**	**2**	**3**	**4**	**5**
1	5.78	6.34	6.24	5.23	6.12
2	5.89	5.87	6.12	6.21	5.99
3	6.22	5.78	5.76	6.02	6.10
4	6.02	5.56	6.21	6.23	6.00
5	5.77	5.76	5.87	5.78	6.03
6	6.00	5.89	6.02	5.98	5.78

(a) Develop a control chart for the mean and range for the available toothpaste data.
(b) Plot the observations on the control chart and comment on your findings.

20. Breeze Toothpaste Company has been having a problem with some of the tubes of toothpaste leaking. The tubes are packed in containers with 100 tubes each. Ten containers of toothpaste have been sampled. The following number of toothpaste tubes were found to have leaks:

Sample	Number of Leaky Tubes	Sample	Number of Leaky Tubes
1	4	6	6
2	8	7	10
3	12	8	9
4	11	9	5
5	12	10	8
		Total	85

Develop a p-chart with three-sigma control limits and evaluate whether the process is in statistical control.

21. The Crunchy Potato Chip Company packages potato chips in a process designed for 10.0 ounces of chips with an upper specification limit of 10.5 ounces and a lower specification limit of 9.5 ounces. The packaging process results in bags with an average net weight of 9.8 ounces and a standard deviation of 0.12 ounces. The company wants to determine if the process is capable of meeting design specifications.

22. The Crunchy Potato Chip Company sells chips in boxes with a net weight of 30 ounces per box (850 grams). Each box contains 10 individual 3-ounce packets of chips. Product design specifications call for the packet-filling process average to be set at 86.0 grams so that the average net weight per box will be 860 grams. Specification width is set for the box to weigh 850 ± 12 grams. The standard deviation of the packet-filling process is 8.0 grams. The target process capability ratio is 1.33. The production manager has just learned that the packet-filling process average weight has dropped down to 85.0 grams. Is the packaging process capable? Is an adjustment needed?

CASE: Scharadin Hotels

Scharadin Hotels are a national hotel chain started in 1957 by Milo Scharadin. What started as one upscale hotel in New York City turned into a highly reputable national hotel chain. Today Scharadin Hotels serve over 100 1ocations and are recognized for their customer service and quality. Scharadin Hotels are typically located in large metropolitan areas close to convention centers and centers of commerce. They cater to both business and nonbusiness customers and offer a wide array of services. Maintaining high customer service has been considered a priority for the hotel chain.

A Problem with Quality

The Scharadin Hotel in San Antonio, Texas, had recently been experiencing a large number of guest complaints due to billing errors. The complaints seem to center around guests disputing charges on their final hotel bill. Guest complaints ranged from extra charges, such as meals or services that were not purchased, to confusion for not being charged at all. Most hotel guests use express checkout on their day of departure. With express checkout the hotel bill is left under the guest's door in the early morning hours and, if all is in order, does not require any additional action on the guest's part. Express checkout is a welcome service by busy travelers who are free to depart the hotel at their convenience. However, the increased number of billing errors began creating unnecessary delays and frustration for the guests who unexpectedly needed to settle their bill with the front desk. The hotel staff often had to calm frustrated guests who were rushing to the airport and were aggravated that they were getting charged for items they had not purchased.

Identifying the Source of the Problem

Larraine Scharadin, Milo Scharadin's niece, had recently been appointed to run the San Antonio hotel. A recent business school graduate, Larraine had grown up in the hotel business. She was poised and confident, and understood the importance of high quality for the hotel. When she became aware of the billing problem, she immediately called a staff meeting to uncover the source of the problem.

During the staff meeting discussion quickly turned to problems with the new computer system and software that had been put in place. Tim Coleman, head of MIS, defended the system, stating that the system was sound and the problems were exaggerated. Tim claimed that a few hotel guests made an issue of a few random problems. Scott Schultz, head of operations, was not so sure. Scott said that he noticed that the number of complaints seem to have significantly increased since the new system was installed. He said that he had asked his team to perform an audit of 50 random bills per day over the past 30 days. Scott showed the following numbers to Larraine, Tim, and the other staff members.

Day	Number of Incorrect Bills	Day	Number of Incorrect Bills	Day	Number of Incorrect Bills
1	2	11	1	21	3
2	2	12	2	22	3
3	1	13	3	23	3
4	2	14	3	24	4
5	2	15	2	25	5
6	3	16	3	26	5
7	2	17	2	27	6
8	2	18	2	28	5
9	1	19	1	29	5
10	2	20	3	30	5

Everyone looked at the data that had been presented. Then Tim exclaimed: "Notice that the number of errors increases in the last third of the month. The computer system had been in place for the entire month so that can't be the problem. Scott, it is probably the new employees you have on staff that are not entering the data properly." Scott quickly retaliated: "The employees are trained properly! Everyone knows the problem is the computer system!"

The argument between Tim and Scott become heated, and Larraine decided to step in. She said, "Scott, I think it is best if you perform some statistical analysis of that data and send us your findings. You know that we want a high-quality standard. We can't be Motorola with six-sigma quantity, but let's try for three-sigma. Would you develop some control charts with the data and let us know if you think the process is in control?"

Case Questions

1. Set up three-sigma control limits with the given data.
2. Is the process in control? Why?
3. Based on your analysis do you think the problem is the new computer system or something else?
4. What advice would you give to Larraine based on the information that you have?

CASE: Delta Plastics, Inc. (B)

Jose De Costa, Director of Manufacturing at Delta Plastics, sat at his desk looking at the latest production quality report, showing the number and type of product defects per week (see the quality report in Delta Plastics, Inc. Case A, Chapter 5). He was faced with the task of evaluating production quality for products made with two different materials. One of the materials was new and called "super plastic" due to its ability to sustain large temperature changes. The other material was the standard plastic that had been successfully used by Delta for many years.

The company had started producing products with the new "super plastic" material only a month earlier. Jose suspected that the new material could result in more defects during the production process than the standard material they had been using. Jose was opposed to starting production until R&D had fully completed testing and refining the new material. However, the CEO of Delta ordered production despite objections from manufacturing and R&D. Jose carefully looked at the report in front of him and prepared to analyze the results.

Case Questions

1. Prepare a three-sigma control chart for both production processes, using the new and standard material (use the quality report in Delta Plastics, Inc. Case A, Chapter 5). Are both processes in control? What can you conclude?
2. Are both materials equally subject to the defects?
3. Given your findings, what advice would you give Jose?

Interactive Learning

Enhance and test your knowledge of Chapter 6. Use the CD and visit our Web site, www.wiley.com/college/reid, for additional resources and information.

1. **Spreadsheets** *Solved Problems 1 and 4*
2. **Company Tours**
   Rickenbacker International Corporation
   Genesis Technologies, Inc.
   Canadian Springs Water Company
3. **Additional Web Resources**
   American Society for Quality Control, www.asqc.org
   Australian Quality Council, www.aqc.org.au
4. **Internet Challenge** *Safe-Air*

To gain business experience, you have volunteered to work at Safe-Air, a nonprofit agency that monitors airline safety records and customer service. Your first assignment is to compare three airlines based on their on-time arrivals and departures. Your manager has asked you to get your information from the Internet. Select any three airlines. For an entire week check the daily arrival and departure schedules of the three airlines from your city or closest airport. Remember that it is important to compare the arrivals and departures from the same location and during the same time period to account for factors such as the weather. Record the data that you collect for each airline. Then decide which types of statistical quality control tools you are going to use to evaluate the airlines' performances. Based on your findings, draw a conclusion regarding the on-time arrivals and departures of each of the airlines. Which is best and which is worst? Are there large differences in performance among the airlines? Also describe the statistical quality control tools you have decided to use to monitor performance. If you have chosen to use more than one tool, are you finding the tools equally useful or is one better at capturing differences in performance? Finally, based on what you have learned so far, how would you perform this analysis differently in the future?

Virtual Company: Valley Memorial Hospital

Assignment: *Statistical Quality Control* This assignment involves controlling nursing hours at Valley Memorial Hospital. Lee Jordan, director of the hospital's Medical/Surgical Nursing Unit, has already told you that VMH employs more than 500 nurses, with an annual nursing budget of $5,000,000. "We're trying for a five

www.wiley.com/college/reid

percent reduction in nursing FTEs—full-time equivalents," he says. "I've been personally recording the nursing hours per patient per day for over three months in Med/Surg. I would like you to look at the numbers and see if you can tell me how to meet our goals.

To complete this assignment, go to **www.wiley.com/college/reid** to get more details on the following projects:

1. Develop upper and lower limits for FTEs within which the Medical/Surgical Nursing Unit will be efficient and will maintain quality at least 95 percent of the time.
2. Look at the data and determine whether Jordan is really in control of nursing hours. If he isn't, tell him why.
3. Determine how the Medical/Surgical Nursing Unit can bring nursing hours per patient day (NHPPD) down to 8.00. Also, provide some advice on how Jordan can get his staff to buy into the concept of an NHPPD target of 8.00.
4. Jot down your thoughts on the three statistical problems, which are contained in memos Jordan received from other VMH staff:
   - Will Hartmann, in the Business Office has kept track of billing errors for the past 21 weeks. Based on this data, determine control limits for billing errors. Also, is the percentage of defective bills a valid measure for this analysis?
   - Analyze trends in patient surveys about the meals served at VMH. Doug Jennings, in Dietary, thinks the number of OUTSTANDING responses has been declining, but he's not sure if that decline is statistically significant.
   - Margot Hamilton, in Housekeeping, has been keeping track of defects in room cleaning. Based on her data, develop some recommendations on how she can get better results.

To access the Web site:

- Go to **www.wiley.com/college/reid**
- Click **Student Companion Site**
- Click **Virtual Company**
- Click **Kaizen Consulting, Inc.**
- Click **Consulting Assignments**
- Click **Statistical Quality Control**

Bibliography

Brue, G. *Six Sigma for Managers.* New York: McGraw-Hill, 2002.

Duncan, A. J. *Quality Control and Industrial Statistics.* 5th ed. Homewood, Ill.: Irwin, 1986.

Evans, James R., and William M. Lindsay. *The Management and Control of Quality.* 4th ed. Cincinnati: South-Western, 1999.

Feigenbaum, A. V. *Total Quality Control.* New York: McGraw-Hill, 1991.

Grant, E. L., and R. S. Leavenworth. *Statistical Quality Control.* 6th ed. New York: McGraw-Hill, 1998.

Hoyer, R. W., and C. E. Wayne. "A Graphical Exploration of SPC, Part 1." *Quality Progress,* 29, no. 5 (May 1996), 65-73.

Juran, J. M., and F. M. Gryna. *Quality Planning and Analysis.* 2nd ed. New York: McGraw-Hill, 1980.

Wadsworth, H. M., K. S. Stephens, and A. B. Godfrey. *Modern Methods for Quality Control and Improvement.* New York: Wiley, 1986.