almost entirely to the study of games of chance and to combinatorial analysis.

8. PROBLEMS FOR SOLUTION

1. Among the digits 1, 2, 3, 4, 5 first one is chosen, and then a second selection is made among the remaining four digits. Assume that all twenty possible results have the same probability. Find the probability that an odd digit will be selected (*a*) the first time, (*b*) the second time, (*c*) both times.

2. In the sample space of example (2.*a*) attach equal probabilities to all 27 points. Using the notation of example (4.*d*), verify formula (7.4) for the two events $A_1 = S_1$ and $A_2 = S_2$. How many points does S_1S_2 contain?

3. Consider the 24 possible arrangements (permutations) of the symbols 1234 and attach to each probability $\frac{1}{24}$. Let A_i be the event that the digit i appears at its natural place (where $i = 1, 2, 3, 4$). Verify formula (7.4).

4. A coin is tossed until for the first time the same result appears twice in succession. To every possible outcome requiring n tosses attribute probability $1/2^{n-1}$. Describe the sample space. Find the probability of the following events: (*a*) the experiment ends before the sixth toss, (*b*) an *even* number of tosses is required.

5. In the sample space of example (5.*b*) let us attribute to each point of (*) containing exactly k letters probability $1/2^k$. (In other words, *aa* and *bb* carry probability $\frac{1}{4}$, *acb* has probability $\frac{1}{8}$, etc.) (*a*) Show that the probabilities of the points of (*) add up to unity, whence the two points (**) receive probability zero. (*b*) Show that the probability that a wins is $\frac{5}{14}$. The probability of b winning is the same, and c has probability $\frac{2}{7}$ of winning. (*c*) The probability that no decision is reached at or before the kth turn (game) is $1/2^{k-1}$.

6. Modify example (5.*b*) to take account of the possibility of ties at the individual games. Describe the appropriate sample space. How would you define probabilities?

7. In problem 3 show that $A_1A_2A_3 \subset A_4$ and $A_1A_2A_3' \subset A_4'$.

8. Using the notations of example (4.*d*) show that (*a*) $S_1S_2D_3 = 0$; (*b*) $S_1D_2 \subset E_3$; (*c*) $E_3 - D_2S_1 \supset S_2D_1$.

9. Two dice are thrown. Let A be the event that the sum of the faces is odd, B the event of at least one ace. Describe the events AB, $A \cup B$, AB'. Find their probabilities assuming that all 36 sample points have equal probabilities.

10. In example (2.*g*), discuss the meaning of the following events:
(*a*) ABC, (*b*) $A - AB$, (*c*) $AB'C$.

11. In example (2.*g*), verify that $AC' \subset B$.

12. *Bridge* (cf. footnote 1). For $k = 1, 2, 3, 4$ let N_k be the event that North has at least k aces. Let S_k, E_k, W_k be the analogous events for South, East, West. Discuss the number x of aces in West's possession in the events
(*a*) W_1', (*b*) N_2S_2, (*c*) $N_1'S_1'E_1'$, (*d*) $W_2 - W_3$,
(*e*) $N_1S_1E_1W_1$, (*f*) N_3W_1, (*g*) $(N_2 \cup S_2)E_2$.

13. In the preceding problem verify that
(*a*) $S_3 \subset S_2$, (*b*) $S_3W_2 = 0$, (*c*) $N_2S_1E_1W_1 = 0$,

14. Verify the following relations.[4]
(*a*) $(A \cup B)' = A'B'$. (*b*) $(A \cup B) - B = A - AB = AB'$.
(*c*) $AA = A \cup A = A$. (*d*) $(A - AB) \cup B = A \cup B$.
(*e*) $(A \cup B) - AB = AB' \cup A'B$. (*f*) $A' \cup B' = (AB)'$.
(*g*) $(A \cup B)C = AC \cup BC$.

15. Find simple expressions for
(*a*) $(A \cup B)(A \cup B')$, (*b*) $(A \cup B)(A' \cup B)(A \cup B')$, (*c*) $(A \cup B)(B \cup C)$.

16. State which of the following relations are correct and which incorrect:
(*a*) $(A \cup B) - C = A \cup (B - C)$.
(*b*) $ABC = AB(C \cup B)$.
(*c*) $A \cup B \cup C = A \cup (B - AB) \cup (C - AC)$.
(*d*) $A \cup B = (A - AB) \cup B$.
(*e*) $AB \cup BC \cup CA \supset ABC$.
(*f*) $(AB \cup BC \cup CA) \subset (A \cup B \cup C)$.
(*g*) $(A \cup B) - A = B$.
(*h*) $AB'C \subset A \cup B$.
(*i*) $(A \cup B \cup C)' = A'B'C'$.
(*j*) $(A \cup B)'C = A'C \cup B'C$.
(*k*) $(A \cup B)'C = A'B'C$.
(*l*) $(A \cup B)'C = C - C(A \cup B)$.

17. Let A, B, C be three arbitrary events. Find expressions for the events that of A, B, C:
(*a*) Only A occurs. (*b*) Both A and B, but not C, occur.
(*c*) All three events occur. (*d*) At least one occurs.
(*e*) At least two occur. (*f*) One and no more occurs.
(*g*) Two and no more occur. (*h*) None occurs.
(*i*) Not more than two occur.

18. The union $A \cup B$ of two events can be expressed as the union of two mutually exclusive events, thus: $A \cup B = A \cup (B - AB)$. Express in a similar way the union of three events A, B, C.

19. Using the result of problem 18 prove that

$$\mathbf{P}\{A \cup B \cup C\} = \mathbf{P}\{A\} + \mathbf{P}\{B\} + \mathbf{P}\{C\} - \mathbf{P}\{AB\} - \mathbf{P}\{AC\} - \mathbf{P}\{BC\} + \mathbf{P}\{ABC\}$$

[This is a special case of IV, (1.5).]

[4] Notice that $(A \cup B)'$ denotes the complement of $A \cup B$, which is not the same as $A' \cup B'$. Similarly, $(AB)'$ is not the same as $A'B'$.

In the above problems AB stands for $A \cap B$.

Refinements. The inequality (9.10) has a companion inequality in the reverse direction. Indeed, from (9.9) it is obvious that

$$d_n - d_{n+1} > \frac{1}{3(2n+1)^2} > \frac{1}{12n+1} - \frac{1}{12(n+1)+1}. \tag{9.13}$$

It follows that the sequence $\{d_n - (12n+1)^{-1}\}$ decreases. Since $\{d_n - (12n)^{-1}\}$ increases this implies the double inequality

$$C + \frac{1}{12n+1} < d_n < C + \frac{1}{12n}. \tag{9.14}$$

Substituting into (9.6), and anticipating that $e^C = \sqrt{2\pi}$, we get

$$\sqrt{2\pi}\, n^{n+\frac{1}{2}}e^{-n}\cdot e^{(12n+1)^{-1}} < n! < \sqrt{2\pi}\, n^{n+\frac{1}{2}}e^{-n}\cdot e^{(12n)^{-1}} \tag{9.15}$$

This double inequality supplements Stirling's formula in a remarkable manner. The ratio of the extreme members is close to $1-(12n^2)^{-1}$, and hence *the right-hand member in (9.15) overestimates $n!$, but with an error of less than $9n^{-2}$ per cent.* In reality the error is much smaller;[17] for $n = 2$ the right side in (9.15) yields 2.0007, for $n = 5$ we get 120.01.

PROBLEMS FOR SOLUTION

Note: Sections 11 and 12 contain problems of a different character and diverse complements to the text.

10. EXERCISES AND EXAMPLES

Note: *Assume in each case that all arrangements have the same probability.*

1. How many different sets of initials can be formed if every person has one surname and (*a*) exactly two given names, (*b*) at most two given names, (*c*) at most three given names?

2. Letters in the Morse code are formed by a succession of dashes and dots with repetitions permitted. How many letters is it possible to form with ten symbols or less?

3. Each domino piece is marked by two numbers. The pieces are symmetrical so that the number-pair is not ordered. How many different pieces can be made using the numbers $1, 2, \ldots, n$?

4. The numbers $1, 2, \ldots, n$ are arranged in random order. Find the probability that the digits (*a*) 1 and 2, (*b*) 1, 2, and 3, appear as neighbors in the order named.

[17] Starting from (9.9) it is possible to show that $d_n = C + (12n)^{-1} - (360n^3)^{-1} + \cdots$ where the dots indicate terms dominated by a multiple of n^{-4}.

5. A throws six dice and wins if he scores at least one ace. B throws twelve dice and wins if he scores at least two aces. Who has the greater probability to win?[18]

Hint: Calculate the probabilities to lose.

6. (*a*) Find the probability that among three random digits there appear exactly 1, 2, or 3 different ones. (*b*) Do the same for four random digits.

7. Find the probabilities p_r that in a sample of r random digits no two are equal. Estimate the numerical value of p_{10}, using Stirling's formula.

8. What is the probability that among k random digits (*a*) 0 does not appear; (*b*) 1 does not appear; (*c*) neither 0 nor 1 appears; (*d*) at least one of the two digits 0 and 1 does not appear? Let A and B represent the events in (*a*) and (*b*). Express the other events in terms of A and B.

9. If n balls are placed at random into n cells, find the probability that exactly one cell remains empty.

10. At a parking lot there are twelve places arranged in a row. A man observed that there were eight cars parked, and that the four empty places were adjacent to each other (formed *one* run). Given that there are four empty places, is this arrangement surprising (indicative of non-randomness)?

11. A man is given n keys of which only one fits his door. He tries them successively (sampling without replacement). This procedure may require $1, 2, \ldots, n$ trials. Show that each of these n outcomes has probability n^{-1}.

12. Suppose that each of n sticks is broken into one long and one short part. The $2n$ parts are arranged into n pairs from which new sticks are formed. Find the probability (*a*) that the parts will be joined in the original order, (*b*) that all long parts are paired with short parts.[19]

13. *Testing a statistical hypothesis.* A Cornell professor got a ticket twelve times for illegal overnight parking. All twelve tickets were given either Tuesdays or Thursdays. Find the probability of this event. (Was his renting a garage only for Tuesdays and Thursdays justified?)

14. *Continuation.* Of twelve police tickets none was given on Sunday. Is this evidence that no tickets are given on Sundays?

15. A box contains ninety good and ten defective screws. If ten screws are used, what is the probability that none is defective?

16. From the population of five symbols a, b, c, d, e, a sample of size 25 is taken. Find the probability that the sample will contain five symbols of each

[18] This paraphrases a question addressed in 1693 to I. Newton by the famous Samuel Pepys. Newton answered that "an easy computation" shows A to be at an advantage. On prodding he later submitted the calculations, but he was unable to convince Pepys. For a short documented account see E. D. Schell, *Samuel Pepys, Isaac Newton, and probability*, The Amer. Statistician, vol. 14 (1960), pp. 27–30. There reference is made to *Private correspondence and miscellaneous papers of Samuel Pepys*, London (G. Bell and Sons), 1926.

[19] When cells are exposed to harmful radiation, some chromosomes break and play the role of our "sticks." The "long" side is the one containing the so-called centromere. If two "long" or two "short" parts unite, the cell dies. See D. G. Catcheside, *The effect of X-ray dosage upon the frequency of induced structural changes in the chromosomes of* Drosophila Melanogaster, Journal of Genetics, vol. 36 (1938), pp. 307–320.

kind. Check the result in tables of random numbers,[20] identifying the digits 0 and 1 with a, the digits 2 and 3 with b, etc.

17. If n men, among whom are A and B, stand in a row, what is the probability that there will be exactly r men between A and B? If they stand in a ring instead of in a row, show that the probability is independent of r and hence $1/(n-1)$. (In the circular arrangement consider only the arc leading from A to B in the positive direction.)

18. What is the probability that two throws with three dice each will show the same configuration if (*a*) the dice are distinguishable, (*b*) they are not?

19. Show that it is more probable to get at least one ace with four dice than at least one double ace in 24 throws of two dice. The answer is known as de Méré's paradox.[21]

20. From a population of n elements a sample of size r is taken. Find the probability that none of N prescribed elements will be included in the sample, assuming the sampling to be (*a*) without, (*b*) with replacement. Compare the numerical values for the two methods when (i) $n = 100$, $r = N = 3$, and (ii) $n = 100$, $r = N = 10$.

21. *Spread of rumors.* In a town of $n + 1$ inhabitants, a person tells a rumor to a second person, who in turn repeats it to a third person, etc. At each step the recipient of the rumor is chosen at random from the n people available. Find the probability that the rumor will be told r times without: (*a*) returning to the originator, (*b*) being repeated to any person. Do the same problem when at each step the rumor is told by one person to a gathering of N randomly chosen people. (The first question is the special case $N = 1$.)

22. *Chain letters.* In a population of $n + 1$ people a man, the "progenitor," sends out letters to two distinct persons, the "first generation." These repeat the performance and, generally, for each letter received the recipient sends out two letters to two persons chosen at random without regard to the past development. Find the probability that the generations number $1, 2, \ldots, r$ will not include the progenitor. Find the median of the distribution, supposing n to be large.

23. *A family problem.* In a certain family four girls take turns at washing dishes. Out of a total of four breakages, three were caused by the youngest girl, and she was thereafter called clumsy. Was she justified in attributing the frequency of her breakages to chance? Discuss the connection with random placements of balls.

24. What is the probability that (*a*) the birthdays of twelve people will fall in twelve different calendar months (assume equal probabilities for the twelve months), (*b*) the birthdays of six people will fall in exactly two calendar months?

[20] They are occasionally miraculously obliging: see J. A. Greenwood and E. E. Stuart, *Review of Dr. Feller's critique*, Journal for Parapsychology, vol. 4 (1940), pp. 298–319, in particular p. 306.

[21] An often repeated story asserts that the problem arose at the gambling table and that in 1654 de Méré proposed it to Pascal. This incident is supposed to have greatly stimulated the development of probability theory. The problem was in fact treated by Cardano (1501–1576). See O. Ore, *Pascal and the invention of probability theory*, Amer. Math. Monthly, vol. 67 (1960), pp. 409–419, and *Cardano, the gambling scholar*, Princeton (Princeton Univ. Press), 1953.

25. Given thirty people, find the probability that among the twelve months there are six containing two birthdays and six containing three.

26. A closet contains n pairs of shoes. If $2r$ shoes are chosen at random (with $2r < n$), what is the probability that there will be (*a*) no complete pair, (*b*) exactly one complete pair, (*c*) exactly two complete pairs among them?

27. A car is parked among N cars in a row, not at either end. On his return the owner finds that exactly r of the N places are still occupied. What is the probability that both neighboring places are empty?

28. A group of $2N$ boys and $2N$ girls is divided into two equal groups. Find the probability p that each group will be equally divided into boys and girls. Estimate p, using Stirling's formula.

29. In bridge, prove that the probability p of West's receiving exactly k aces is the same as the probability that an arbitrary hand of thirteen cards contains exactly k aces. (This is intuitively clear. Note, however, that the two probabilities refer to two different experiments, since in the second case thirteen cards are chosen at random and in the first case all 52 are distributed.)

30. The probability that in a bridge game East receives m and South n spades is the same as the probability that of two hands of thirteen cards each, drawn at random from a deck of bridge cards, the first contains m and the second n spades.

31. What is the probability that the bridge hands of North and South together contain exactly k aces, where $k = 0, 1, 2, 3, 4$?

32. Let a, b, c, d be four non-negative integers such that $a + b + c + d = 13$. Find the probability $p(a, b, c, d)$ that in a bridge game the players North, East, South, West have a, b, c, d spades, respectively. Formulate a scheme of placing red and black balls into cells that contains the problem as a special case.

33. Using the result of problem 32, find the probability that some player receives a, another b, a third c, and the last d spades if (*a*) $a = 5$, $b = 4$, $c = 3$, $d = 1$; (*b*) $a = b = c = 4$, $d = 1$; (*c*) $a = b = 4$, $c = 3$, $d = 2$. Note that the three cases are essentially different.

34. Let a, b, c, d be integers with $a + b + c + d = 13$. Find the probability $q(a, b, c, d)$ that a hand at bridge will consist of a spades, b hearts, c diamonds, and d clubs and show that the problem does *not* reduce to one of placing, at random, thirteen balls into four cells. Why?

35. *Distribution of aces among r bridge cards.* Calculate the probabilities $p_0(r), p_1(r), \ldots, p_4(r)$ that among r bridge cards drawn at random there are $0, 1, \ldots, 4$ aces, respectively. Verify that $p_0(r) = p_4(52-r)$.

36. *Continuation: waiting times.* If the cards are drawn one by one, find the probabilities $f_1(r), \ldots, f_4(r)$ that the first, ..., fourth ace turns up at the rth trial. *Guess* at the *medians* of the waiting times for the first, ..., fourth ace and then calculate them.

37. Find the probability that each of two hands contains exactly k aces if the two hands are composed of r bridge cards each, and are drawn (*a*) from the same deck, (*b*) from two decks. Show that when $r = 13$ the probability in part (*a*) is the probability that two preassigned bridge players receive exactly k aces each.

38. *Misprints.* Each page of a book contains N symbols, possibly misprints. The book contains $n = 500$ pages and $r = 50$ misprints. Show that

A comparison of (7.3) and (7.4) shows that

$$p_{n+1} = \frac{u_{n+1} + v_{n+1}}{1 - w_{n+1}} = \frac{p_n}{1 - q_n^2} = \frac{1}{1 + q_n} \tag{7.5}$$

and similarly

$$q_{n+1} = \frac{v_{n+1}}{1 - w_{n+1}} = \frac{q_n}{1 + q_n}. \tag{7.6}$$

From (7.6) we can calculate q_n explicitly. In fact, taking reciprocals we get

$$q_{n+1}^{-1} = 1 + q_n^{-1} \tag{7.7}$$

whence successively

$$q_1^{-1} = 1 + q^{-1}, \qquad q_2^{-1} = 2 + q^{-1}, \qquad q_3^{-1} = 3 + q^{-1}, \quad \ldots, \quad q_n^{-1} = n + q^{-1} \tag{7.8}$$

or

$$q_n = \frac{q}{1 + nq}, \qquad w_{n+1} = \left(\frac{q}{1 + qn}\right)^2. \tag{7.9}$$

We see that the unproductive (or undesirable) genotype gradually drops out, but the process is extremely slow. For $q = 0.1$ it takes ten generations to reduce the frequency of a-genes by one-half; this reduces the frequency of the aa-type approximately from 1 to $\frac{1}{4}$ per cent. (If a is sex-linked, the elimination proceeds much faster; see problem 29. For a generalized selection scheme see problem 30.)[11]

8. PROBLEMS FOR SOLUTION

1. Three dice are rolled. If no two show the same face, what is the probability that one is an ace?

2. Given that a throw with ten dice produced at least one ace, what is the probability p of two or more aces?

3. *Bridge*. In a bridge party West has no ace. What probability should he attribute to the event of his partner having (*a*) no ace, (*b*) two or more aces? Verify the result by a direct argument.

4. *Bridge*. North and South have ten trumps between them (trumps being cards of a specified suit). (*a*) Find the probability that all three remaining trumps are in the same hand (that is, either East or West has no trumps). (*b*) If it is known that the king of trumps is included among the three, what is the probability that he is "unguarded" (that is, one player has the king, the other the remaining two trumps)?

5. Discuss the key problem in example II, (7.*b*) in terms of conditional probabilities following the pattern of example (2.*a*).

6. In a bolt factory machines A, B, C manufacture, respectively, 25, 35, and 40 per cent of the total. Of their output 5, 4, and 2 per cent are defective bolts. A bolt is drawn at random from the produce and is found defective. What are the probabilities that it was manufactured by machines A, B, C?

7. Suppose that 5 men out of 100 and 25 women out of 10,000 are colorblind. A colorblind person is chosen at random. What is the probability of his being male? (Assume males and females to be in equal numbers.)

8. Seven balls are distributed randomly in seven cells. Given that two cells are empty, show that the (conditional) probability of a triple occupancy of some cells equals $\frac{1}{4}$. Verify this numerically using table 1 of II, 5.

9. A die is thrown as long as necessary for an ace to turn up. Assuming that the ace does not turn up at the first throw, what is the probability that more than three throws will be necessary?

10. *Continuation*. Suppose that the number, n, of throws is even. What is the probability that $n = 2$?

11. Let[12] the probability p_n that a family has exactly n children be αp^n when $n \geq 1$, and $p_0 = 1 - \alpha p(1 + p + p^2 + \cdots)$. Suppose that all sex distributions of n children have the same probability. Show that for $k \geq 1$ the probability that a family has exactly k boys is $2\alpha p^k/(2-p)^{k+1}$.

12. *Continuation*. Given that a family includes at least one boy, what is the probability that there are two or more?

13. Die A has four red and two white faces, whereas die B has two red and four white faces. A coin is flipped *once*. If it falls heads, the game continues by throwing die A alone; if it falls tails, die B is to be used. (*a*) Show that the probability of red at any throw is $\frac{1}{2}$. (*b*) If the first two throws resulted in red, what is the probability of red at the third throw? (*c*) If red turns up at the first n throws, what is the probability that die A is being used? (*d*) To which urn model is this game equivalent?

14. In example (2.*a*) let x_n be the conditional probability that the winner of the nth trial wins the entire game given that the game does not terminate at the nth trial; let y_n and z_n be the corresponding probabilities of victory for the losing and the pausing player, respectively, of the nth trial. (*a*) Show that

$$x_n = \tfrac{1}{2} + \tfrac{1}{2}y_{n+1}, \qquad y_n = \tfrac{1}{2}z_{n+1}, \qquad z_n = \tfrac{1}{2}x_{n+1}. \tag{$*$}$$

(*b*) Show by a direct simple argument that in reality $x_n = x$, $y_n = y$, $z_n = z$ are independent of n. (*c*) Conclude that the probability that player a wins the game is $\frac{5}{14}$ (in agreement with problem 5 in I, 8). (*d*) Show that $x_n = \frac{4}{7}$, $y_n = \frac{1}{7}$, $z_n = \frac{2}{7}$ is the only bounded solution of ($*$).

[11] For a further analysis of various eugenic effects (which are frequently different from the ideas of enthusiastic proponents of sterilization laws) see G. Dahlberg, *Mathematical methods for population genetics*, New York and Basel, 1948.

[12] According to A. J. Lotka, American family statistics satisfies our hypothesis with $p = 0.7358$. See *Théorie analytique des associations biologiques II*, Actualités scientifiques et industrielles, no. 780, Paris, 1939.

*8. THE CORRELATION COEFFICIENT

Let X and Y be any two random variables with means μ_x and μ_y and positive variances σ_x^2 and σ_y^2. We introduce the corresponding normalized variables X* and Y* defined by (4.6). Their covariance is called *the correlation coefficient of* X, Y *and is denoted by* $\rho(X, Y)$. Thus, using (5.4),

$$\rho(X, Y) = \text{Cov}(X^*, Y^*) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} \tag{8.1}$$

Clearly this correlation coefficient is independent of the origins and units of measurements, that is, for any constants a_1, a_2, b_1, b_2, with $a_1 > 0$, $a_2 > 0$, we have $\rho(a_1X+b_1, a_2Y+b_2) = \rho(X, Y)$.

The use of the correlation coefficient amounts to a fancy way of writing the covariance.[8] Unfortunately, the term correlation is suggestive of implications which are not inherent in it. We know from section 5 that $\rho(X, Y) = 0$ whenever X and Y are independent. It is important to realize that the converse is not true. In fact, *the correlation coefficient* $\rho(X, Y)$ *can vanish even if* Y *is a function of* X.

Examples. (*a*) Let X assume the values ± 1, ± 2 each with probability $\frac{1}{4}$. Let $Y = X^2$. The joint distribution is given by $p(-1, 1) = p(1, 1) = p(2, 4) = p(-2, 4) = \frac{1}{4}$. For reasons of symmetry $\rho(X, Y) = 0$ even though we have a direct functional dependence of Y on X.

(*b*) Let U and V have the same distribution, and let $X = U + V$, $Y = U - V$. Then $E(XY) = E(U^2) - E(V^2) = 0$ and $E(Y) = 0$. Hence $\text{Cov}(X, Y) = 0$ and therefore also $\rho(X, Y) = 0$. For example, X and Y may be the sum and difference of points on two dice. Then X and Y are either both odd or both even and therefore dependent. ▶

It follows that the correlation coefficient is by no means a general measure of dependence between X and Y. However, $\rho(X, Y)$ is connected with the *linear* dependence of X and Y.

Theorem. *We have always* $|\rho(X, Y)| \leq 1$; *furthermore,* $\rho(X, Y) = \pm 1$ *only if there exist constants* a *and* b *such that* $Y = aX + b$, *except, perhaps, for values of* X *with zero probability.*

Proof. Let X* and Y* be the normalized variables. Then

$$\text{Var}(X^* \pm Y^*) = \text{Var}(X^*) \pm 2\,\text{Cov}(X^*, Y^*) + \text{Var}(Y^*) = 2(1 \pm \rho(X, Y)). \tag{8.2}$$

* This section treats a special topic and may be omitted at first reading.

[8] The physicist would define the correlation coefficient as "dimensionless covariance."

The left side cannot be negative; hence $|\rho(X, Y)| \leq 1$. For $\rho(X, Y) = 1$ it is necessary that $\text{Var}(X^* - Y^*) = 0$ which means that with unit probability the variable $X^* - Y^*$ assumes only one value. In this case $X^* - Y^* = \text{const.}$, and hence $Y = aX + \text{const.}$ with $a = \sigma_y/\sigma_x$. A similar argument applies to the case $\rho(X, Y) = -1$. ▶

9. PROBLEMS FOR SOLUTION

1. Seven balls are distributed randomly in seven cells. Let X_i be the number of cells containing exactly i balls. Using the probabilities tabulated in II, 5, write down the joint distribution of (X_2, X_3).

2. Two ideal dice are thrown. Let X be the score on the first die and Y be the larger of two scores. (*a*) Write down the joint distribution of X and Y. (*b*) Find the means, the variances, and the covariance.

3. In five tosses of a coin let X, Y, Z be, respectively, the number of heads, the number of head runs, the length of the largest head run. Tabulate the 32 sample points together with the corresponding values of X, Y, and Z. By simple counting derive the joint distributions of the pairs (X, Y), (X, Z), (Y, Z) and the distributions of X + Y and XY. Find the means, variances, covariances of the variables.

4. Let X, Y, and Z be independent random variables with the same geometric distribution $\{q^kp\}$. Find (*a*) $P\{X = Y\}$; (*b*) $P\{X \geq 2Y\}$; and (*c*) $P\{X + Y \leq Z\}$.

5. *Continuation.* Let U be the larger of X and Y, and put $V = X - Y$. Show that U and V are independent.[9]

6. Let X_1 and X_2 be independent random variables with Poisson distributions $\{p(k; \lambda_1)\}$ and $\{p(k; \lambda_2)\}$.
(*a*) Prove that $X_1 + X_2$ has the Poisson distribution $\{p(k; \lambda_1 + \lambda_2)\}$.
(*b*) Show that the *conditional distribution of* X_1 *given* $X_1 + X_2$ *is binomial*, namely

$$P\{X_1 = k \mid X_1 + X_2 = n\} = b\left(k; n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right). \tag{9.1}$$

7. Let X_1 and X_2 be independent and have the common geometric distribution $\{q^kp\}$ (as in problem 4). Show without calculations that the *conditional distribution of* X_1 *given* $X_1 + X_2$ *is uniform*, that is,

$$P\{X_1 = k \mid X_1 + X \ = n\} = \frac{1}{n+1}, \qquad k = 0, \ldots, n. \tag{9.2}$$

8. Let $X_1, \ldots, X_r$ be mutually independent random variables, each having the *uniform distribution* $P\{X_i = k\} = 1/N$ for $k = 1, 2, \ldots, N$. Let U_n be the smallest among the $X_1, \ldots, X_n$ and V_n the largest. Find the distributions of U_n and V_n. What is the connection with the *estimation problem* (3.e)?

[9] The geometric distribution is the only probability distribution on the integers for which this is true. See T. S. Ferguson, *A characterization of the geometric distribution*, Amer. Math. Monthly, vol. 72 (1965), pp. 256–260.

9. *Continuation to the estimation problem in example* (3.*e*). (*a*) Find the joint distribution of the largest and the smallest observation. Specialize to $n = 2$. (*Hint:* Calculate first $P\{X \leq r, Y \geq s\}$.)

(*b*) Find the conditional probability that the first two observations are j and k, given that $X = r$.

(*c*) Find $E(X^2)$ and hence an asymptotic expression for Var (X) as $N \to \infty$ (with n fixed).

10. *Simulating a perfect coin.* Given a biased coin such that the probability of heads is α, we simulate a perfect coin as follows. Throw the biased coin twice. Interpret HT as success and TH as failure; if neither event occurs repeat the throws until a decision is reached. (*a*) Show that this model leads to Bernoulli trials with $p = \frac{1}{2}$. (*b*) Find the distribution and the expectation of the number of throws required to reach a decision.

11. *The problem of Banach's match boxes, example* VI,(8.*a*). Show that the expectation of the distribution $\{u_r\}$ is given by $\mu = (2N+1)u_0 - 1$. Using Stirling's formula show that this is approximately $2\sqrt{N/\pi} - 1$. (For $N = 50$ the mean is about 7.04.)

Hint: Start from the relation

$$(N-r)u_r = \tfrac{1}{2}(2N+1)u_{r+1} - \tfrac{1}{2}(r+1)u_{r+1}.$$

Use the fact[10] that $\sum u_r = 1$.

12. *Sampling inspection.* Suppose that items with a probability p of being acceptable are subjected to inspection in such a way that the probability of an item being inspected is p'. We have four classes, namely, "acceptable and inspected," "acceptable but not inspected," etc. with corresponding probabilities pp', pq', $p'q$, qq' where $q = 1 - p$, $q' = 1 - p'$. We are concerned with double Bernoulli trials [see example VI,(9.*c*)]. Let N be the number of items passing the inspection desk (both inspected and uninspected) before the first defective is found, and let K be the (undiscovered) number of defectives among them. Find the joint distributions of N and K and the marginal distributions.

13. *Continuation.* Find $E\left(\frac{K}{N+1}\right)$ and Cov (K, N). [In industrial practice the discovered defective item is replaced by an acceptable one so that $K/(N+1)$ is the fraction of defectives and measures the quality of the lot. Note that $E\left(\frac{K}{N+1}\right)$ is not $E(K)/E(N+1)$.]

14. In a sequence of Bernoulli trials let X be the length of the run (of either successes or failures) started by the first trial. (*a*) Find the distribution of X, E(X), Var (X). (*b*) Let Y be the length of the *second* run. Find the distribution of Y, E(Y), Var (Y), and the joint distribution of X, Y.

15. Let X and Y have a common negative binomial distribution. Find the conditional probability $P\{X = j \mid X + Y = k\}$ and show that the identity II, (12.16) now becomes obvious without any calculations.[11]

[10] This fact is not obvious analytically; it may be verified by induction on N.

[11] This derivation permits generalizations to more than two factors. It is due to T. K. M. Wisniewski, Amer. Statistician, vol. 20 (1966), p. 25.

16. If two random variables X and Y assume only two values each, and if Cov (X, Y) = 0, then X and Y are independent.

17. *Birthdays.* For a group of n people find the expected number of days of the year which are birthdays of exactly k people. (Assume 365 days and that all arrangements are equally probable.)

18. *Continuation.* Find the expected number of multiple birthdays. How large should n be to make this expectation exceed 1?

19. A man with n keys wants to open his door and tries the keys independently and at random. Find the mean and variance of the number of trials (*a*) if unsuccessful keys are not eliminated from further selections; (*b*) if they are. (Assume that only one key fits the door. The exact distributions are given in II, 7, but are not required for the present problem.)

20. Let (X, Y) be random variables whose joint distribution is the trinomial defined by (1.8). Find E(X), Var (X), and Cov (X, Y) (*a*) by direct computation, (*b*) by representing X and Y as sums of n variables each and using the methods of section 5.

21. Find the covariance of the number of ones and sixes in n throws of a die.

22. In the animal trapping problem 24 of VI, 10, prove that the expected number of animals trapped at the νth trapping is $nqp^{\nu-1}$.

23. If X has the *geometric* distribution $P\{X = k\} = q^k p$ (where $k = 0, 1, \ldots$), show that $\text{Var}(X) = qp^{-2}$. Conclude that the *negative* binomial distribution $\{f(k; r, p)\}$ has variance rqp^{-2} provided r is a positive integer. Prove by direct calculation that the statement remains true for all $r > 0$.

24. In the *waiting time problem* (3.*d*) prove that

$$\text{Var}(S_r) = N\left\{\frac{1}{(N-1)^2} + \frac{2}{(N-2)^2} + \cdots + \frac{r-1}{(N-r+1)^2}\right\}.$$

Conclude that $N^{-2}E(S_N) \sim \sum k^{-2}$. (Incidentally, the value of this series is $\pi^2/6$.) *Hint:* Use the variance of the geometric distribution found in the preceding problem.

25. *Continuation.* Let Y_r be the number of drawings required to include r preassigned elements (instead of any r different elements as in the text). Find $E(Y_r)$ and $\text{Var}(Y_r)$. (*Note:* The exact distribution of Y_r was found in problem 12 of II, 11 but is not required for the present purpose.)

26. *The blood-testing problem.*[12] A large number, N, of people are subject to a blood test. This can be administered in two ways. (i) Each person can be

[12] This problem is based on a technique developed during World War II by R. Dorfman. In army practice Dorfman achieved savings up to 80 per cent. When the problem appeared in the first edition it caught widespread attention and led to various generalizations as well as to new industrial and biological applications. The main improvement consists in introducing more than two stages. See, for example, M. Sobel and P. A. Groll, *Group testing to eliminate efficiently all defectives in a binomial sample*, The Bell System Journal, vol. 38 (1959), pp. 1179–1252; G. S. Watson, *A study of the group screening method*, Technometrics, vol. 3 (1961), pp. 371–388; H. M. Finucan, *The blood-testing problem*, Applied Statistics, vol. 13 (1964), pp. 43–50.