Intro to Probability Theory

Yiqiao YIN
Statistics Department
Columbia University
Notes in LaTeX

July 4, 2017

Abstract

This is the lecture notes from Probability Theory class offered in Statistics Department at Columbia University. Topics include Combinatorial, Axioms of Probability, Conditional Probability and Independence, Random Variables, Continuous Random Variables, Jointly Distributed Random Variables, and Properties of Expectation. This document is a small note from conventional probability text book, *Probability and Statistics*, by Ross.

This note is dedicated to Professor Y. Kim.

Contents

1 COMBINATORIAL ANALYSIS

Go back to Table of Contents. Please click TOC

1.1 Principle of Counting

Principle of counting states that if one experiment can result in any of m possible outcomes and if another experiment can result in any of n possible outcomes, then there are mn possible outcomes of the two experiments.

The basic principle of counting states the following. Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of m possible outcomes and if, for each outcome of experiment 1, there are n possible outcomes of experiment 2, then together there are mn possible outcomes of the two experiments.

A small proof for basic counting principle: The basic principle may be proved by enumerating all the possible outcomes of the two experiments; that is,

$$\begin{array}{cccc} (1,1), & (1,2), & ..., & (1,n) \\ (2,1), & (2,2), & ..., & (2,n) \\ ..., & & & \\ (m,1), & (m,2), & ..., & (m,n) \end{array}$$

Example 1.1. A small coummunity consists of 10 women, each whom has 3 children. If one woman and one of her children are to be chosen as mother and child of the year, how many different choices are possible?

Solution By regarding the choice of the woman as the outcome of the first experiment and the subsequent choice of one of her children as the outcome of the second experiment, we see from basic principle that there are $10 \times 3 = 30$ possible choices.

□

The generalized basic principle of count states the following. If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes; and if, for each of these n_1 possible outcomes, there are n_2 possible outcomes of the second experiment; and if, for each of the possible outcomes of the first two experiments, htere are n_3 possible outcomes of the third experiment; and if ..., then there is a total of $n_1 \cdot n_2 \dots n_r$ possible outcomes of the r experiments.

Example 1.2. A college planning committee consists of 3 freshmen, 4 sophomores, 5 juniors, and 2 seniors. A subcommittee of 4, consisting of 1 person from each class, is to be chosen. How many different subcommittees are possible?

Solution We regard the choice of a subcomittee as the combined outcome of the four separate experiments of choosing a single representative from each of the classes. It then follows from the generalized version of the basic principle that there are $3 \times 4 \times 5 \times 2 = 120$ possible subcommittees.

□

1.2 Permutations

How many different ordered arrangements of the letters a, b, and c are possible? By direct enumeration we see that there are 6, namely, abc, acb, bac, bca, cab, and cba. Each arrangement is known as a permutation. Thus, there are 6 possible permutations of a set of 3 objects.

Suppose now that we have n objects. Reasoning similar to that we have just used for the 3 letters then shows that there are

$$n(n-1)(n-2)\dots 3 \cdot 2 \cdot 1 = n!$$

different permutations of the n objects.

Example 1.3. How many different batting orders are possible for a baseball team consisting of 9 players?

Solution There are $9! = 362,880$ possible batting orders.

□

Example 1.4. A class in probability theory consists of 6 men and 4 women. An examination is given, and the students are ranked according to their performance. Assume that no two students obtain the same score.

(a) How many different rankings are possible?

(b) If the men are ranked just among themselves and the women just among themselves, how many different rankings are possible?

Solution (a) Because each ranking corresponds to a particular ordered arrangement of the 10 people, the answer is $10! = 3,628,800$. (b) Since there are 6! possible rankins of the men among themselves and 4! possible rankings of the women among themselves, it follows from the basic principle that there are $(6!)(4!) = (720)(24) = 17,280$ possible rankings in the case.

□

In general, the same reasoning as that used in exmpales show that there are

$$\frac{n!}{n_1!n_2!\dots n_r!}$$

different permutations of n objects, of which n_1 are alike, n_2 are alike, ..., n_2 are alike.

1.3 Combinations

We are often interested in determining the number of different groups of r objects that could be formed from a total of n objects. For instance, how many different groups of 3 could be selected from the 5 items A, B, C, D, and E? To answer this question, reason as follows: since there are 5 ways to select the initial item, 4 ways to then select the next item, and 3 ways to select the final item, there are thus $5 \cdot 4 \cdot 3$ ways of selecting the group of 3 when the order in which the items are selected is relevant. However, since every group of 3 — say, the group consisting of items A, B, and C — will be counted 6 times (that is, all of the permutations ABC, ACB, BAC, BCA, CAB, and CBA will be counted when the order of selection is relevant), it follows that the total number of groups that can be formed is

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10$$

Definition 1.5. Notation and terminology for combinations. We define $\binom{n}{r}$, for $r \leq n$, by

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

and say that $\binom{n}{r}$ (read as "n choose r") represents the number of possible combinations of n objects taken r at a time.

Thus, $\binom{n}{r}$ represents the number of different groups of size r that could be selected from a set of n objects when the order of selection is not considered relevant. Equivalently, $\binom{n}{r}$ is the number of subsets of size r that can be chosen from a set of size n. Using that $0! = 1$, note that $\binom{n}{n} = \binom{n}{0} = \frac{n!}{0!n!} = 1$, which consistent with the preceding interpretation because in a set of size n there is exactly 1 subset of size n, and exactly one subset of size 0 (namely the empty set).

Example 1.6. A committee of 3 is to be formed from a group of 20 people. How many different committees are possible?

Solution There are $\binom{20}{3} = \frac{20 \cdot 19 \cdot 18}{3 \cdot 2 \cdot 1} = 1140$ possible committees.

□

Example 1.7. From a group of 5 women and 7 men, how many different committees consisting of 2 women and 3 men can be formed? What if 2 of the men are feuding and refuse to serve on the committee together?

Solution As there are $\binom{5}{2}$ possible groups of 2 women, and $\binom{7}{3}$ possible groups of 3 men, it follows from the basic principle that there are $\binom{5}{2}\binom{7}{3} = \left(\frac{5 \cdot 4}{2 \cdot 1}\right)\frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 350$ possible committees consisting of 2 women and 3 men.

Now suppose that 2 of the men refuse to serve together. Because a total of $\binom{2}{2}\binom{5}{1} = 5$ out of the $\binom{7}{3} = 35$ possible groups of 3 men contain both of the feuding men, it follows that there are $35 - 5 = 30$ groups that do not contain both of the feuding men. Because there are still $\binom{5}{2} = 10$ ways to choose 2 women, there are $30 \cdot 10 = 300$ possibloe committees in this case.

□

A useful combinatorial identity is

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r} \quad 1 \leq r \leq n$$

Theorem 1.8. *The Binomial Theorem.*

$$(x+y)^n = \sum_{k=1}^{n} \binom{n}{k} x^k y^{n-k} \tag{1}$$

Proof: When $n = 1$, equation 1 reduces to

$$x + y = \binom{1}{0} x^0 y^1 + \binom{1}{1} x^1 y^0 = y + x$$

Assume equation 1 for $n - 1$. Now,

$$\begin{aligned} (x+y)^n &= (x+y)(x+y)^{n-1} \\ &= (x+y)\sum_{k=0}^{n-1} \tbinom{n-1}{k} x^k y^{n-1-k} \\ &= \sum_{k=0}^{n-1} \tbinom{n-1}{k} x^{k+1} y^{n-1-k} + \sum_{k=0}^{n-1} \tbinom{n-1}{k} x^k y^{n-k} \end{aligned}$$

Letting $i = k + 1$ in the first sum and $i = k$ in the second sum, we find that

$$\begin{aligned} (x+y)^n &= \sum_{i=1}^{n} \tbinom{n-1}{i-1} x^i y^{n-i} + \sum_{i=0}^{n-1} \tbinom{n-1}{i} x^i y^{n-i} \\ &= x^n + \sum_{i=1}^{n-1} \left[\tbinom{n-1}{i-1} + \tbinom{n-1}{i} \right] x^i y^{n-i} + y^n \\ &= x^n + \sum_{i=1}^{n-1} \tbinom{n}{i} x^i y^{n-i} + y^n \\ &= \sum_{i=0}^{n} \tbinom{n}{i} x^i y^{n-i} \end{aligned}$$

Combinatorial Proof of the Binomial Theorem: Consider the product

$$(x_1 + y_1)(x_2 + y_2) \dots (x_n + y_n)$$

Its expansion consists of the sum of 2^n terms, each term being the product of n factors. Moreover, each of the 2^n terms in the sum will contain as a factor either x_i or y_i, for each $i = 1, 2, ., ., n$. For example,

$$(x_1 + y_1)(x_2 + y_2) = x_1 x_2 + x_1 y_2 + y_1 x_2 + y_1 y_2$$

Now, how many of the 2^n terms in the sum will have k of the x_i's and $(n - k)$ of the y_i's as factors? As each term consisting of k of the x_i's and $(n - k)$ of the y_i's corresponds to a choice of a group of k from the n values x_1, x_2, ..., x_n, there are $\binom{n}{k}$ such terms. Thus, letting $x_i = x$, $y_i = y$, with $i = 1, ..., n$, we see that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

where the next-to-last equality follows by 1. By induction, the theorem is proved.

□

1.4 Multinomial Coefficients

A set of n distinct items is to be divided into r distinct groups of respective sizes n_1, n_2, ..., n_r where $\sum_{i=1}^{r} n_i = n$. How many different divisions are possible? To answer this, we note that there are $\binom{n}{n_1}$ possible choices for the first group; for each choice of the first group, there are $\binom{n-n_1}{n_2}$ possible choices for the second group; for each choice of the first two groups, there are $\binom{n-n_1-n_2}{n_2}$ possible choices for the second group; for each choice of the first two groups, there are $\binom{n-n_1-n_2}{n_3}$ possible choices for the third group; and so on. It follows from the generalized version of the basic counting principle that htere are

$$\begin{aligned} & \binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-n_2-\cdots-n_{r-1}}{n_r} \\ = & \frac{n!}{(n-n_1)!n_1!}\frac{(n-n_1)!}{(n-n_1-n_2)!n_2!}\cdots\frac{(n-n_1-n_2-\cdots-n_{r-1})!}{0!n_r!} \\ = & \frac{n!}{n_1!n_2!\ldots n_r!} \end{aligned}$$

Notation If $n_1+n_2+\cdots+n_r=n$, we define $\binom{n}{n_1,n_2,\ldots,n_r}$ by

$$\binom{n}{n_1,n_2,\ldots,n_r}=\frac{n!}{n_1!n_2!\ldots n_r!}$$

Thus, $\binom{n}{n_1,n_2,\ldots,n_r}$ represents the number of possible divisions of n distinct objects into r distinct groups of respective sizes $n_1,n_2,...,n_r$.

Example 1.9. A police department in a small city consists of 10 officers. If the department policy is to have 5 of the officers patrolling the streets, 2 of the officers working full time at the station, and 3 of the officers on reserve at the station, how many different divisions of the 10 officers into the 3 groups are possible?

Solution There are $\frac{10!}{5!2!3!}=2520$ possible divisions.

□

Example 1.10. Ten children are to be divided into an A team and a B team of 5 each. The A team will play in one league and the B team in another. How many different divisionsa re possible?

Solution There are $\frac{10!}{5!5!}=252$ possible divisions.

□

Theorem 1.11. *The Multinomial Theorem.*

$$(x_1+x_2+\cdots+x_r)^n=\sum_{\substack{(n_1,\ldots,n_r):\\ n_1+\cdots+n_r=n}}\binom{n}{n_1,n_2,...,n_r}x_1^{n_1}x_2^{n_2}\ldots x_r^{n_r}$$

The number $\binom{n}{n_1,n_2,\ldots,n_r}$ are known as multinomial coefficients.

2 AXIOMS OF PROBABILITY

Go back to Table of Contents. Please click TOC

2.1 Sample Space and Events

Consider an experiment whose outcome is not predictable with certainty. However, although the outcome of the experiment will not be known in advance, let us suppose that the set of all possible outcomes is known. This set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted bythe following:

1. If the outcome of an experiment consists of the determination of the sex of a newborn child, then
$$S = \{g, b\}$$
where the outcome g means that the child is a girl and b that it is a boy.
2. If the outcome of an experiment is the order of finish in a race among the 7 horses having post positions 1, 2, 3, 4, 5, 6, and 7, then
$$S = \{\text{all 7! permutations of } (1, 2, 3, 4, 5, 6, 7)\}$$
The outcome $(2, 3, 1, 6, 5, 4, 7)$ means, for isntance, that the number 2 horse comes in first, then the number 3 horse, then the number 1 horse, and so on.
3. If the experiment consists of flipping two coins, then the sample space consists of the following four points:
$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$
4. If the experiment consists of tossing two dice, then the sample space consists of the 36 points
$$S = \{(i, j) : i, j = 1, 2, 3, 4, 5, 6\}$$
where the outcome (i, j) is said to occur if i appears on the leftmost die and j on the other die.
5. If the experiment consists of measuring (in hours) the lifetime of a transistor, then the sample space consists of all nonnegative real numbers; that is,
$$S = \{x : 0 \leq x < \infty\}$$
Any subset E of the sample space is known as an event. In other words, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in E, then we say that E has occurred.

The operations forming unions, intersections, and complements of events obey certain rules similar to the rules of algebra. We list a few of these rules:

Commutative laws $E \cup F = F \cup E$ and $EF = FE$

Associative laws $(E \cup F) \cup G = E \cup (F \cup G)$ and $(EF)G = E(FG)$

Distributive laws $(E \cup F)G = EG \cup FG$ and $EF \cup G = (E \cup G)(F \cup G)$

These relations are verified by showing that any outcome that is contained in the event on the left side of the equality sign is also contained in the event on the right side, and vice versa.

The following useful relationships among the three basic operations of forming unions, intersections, and complements are known as DeMorgan's laws:

$$\left(\bigcup_{i=1}^{n} E_i\right)^c = \bigcap_{i=1}^{n} E_i^c$$
$$\left(\bigcap_{i=1}^{n} E_i\right)^c = \bigcup_{i=1}^{n} E_i^c$$

For instance, for two events E and F, DeMorgan's laws state that

$$(E \cup F)^c = E^c F^c \text{ and } (EF)^c = E^c \cup F^c$$

which can be easily proved by Venn diagrams.

To prove DeMorgan's laws for general n, suppose first that x is an outcome of $\left(\bigcup_{i=1}^{n} E_i\right)^c$. Then x is not contained in $\bigcup_{i=1}^{n} E_i$, which means that x is not contained in any of the events E_i, $i = 1, 2, .., n$, implying that x is contained in E_i^c for all $i = 1, 2, ..., n$ and thus is contained in $\bigcap_{i=1}^{n} E_i^c$. To go the other way, suppose that x is an outcome of $\bigcap_{i=1}^{n} E_i^c$. Then x is contained in E_i^c for all $i = 1, 2, ..., n$ which means that x is not contained in E_i for any $i = 1, 2, ..., n$ implying that x is not contained in $\bigcup_{i}^{n} E_i$, in turn implying that x is contained in $\left(\bigcup_{1}^{n} E_i\right)^c$. This proves the first of DeMorgan's laws.

To prove the second of DeMorgan's laws, we use the first law to obtain

$$\left(\bigcup_{i=1}^{n} E_i^c\right)^c = \bigcap_{i=1}^{n} (E_i^c)^c$$

2.2 Axioms of Probability

One way of defining the probability of an event is in terms of its relative frequency. Such a definition usually goes as follows: we suppose that an experiment, whose sample space is S, is repeatedly performed under exactly the same conditions. For each event E of the sample space S, we define $n(E)$ to be the number of times in the first n repetitions of the experiment that the event E occurs. Then $P(E)$, the probability of the event E, is defined as

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

That is, $P(E)$ is defined as the (limiting) proportion of time that E occurs. It is thus the limiting relative frequency of E.

Proposition 2.1. *The three axioms of probability:*
Axiom 1

$$0 \leq P(E) \leq 1$$

Axiom 2

$$P(S) = 1$$

Axiom 3
For any sequence of mutually exclusive events $E_1, E_2, ...$(that is, events for which $E_i E_j = \emptyset$ when $i \neq j$),

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Remark 2.2. We have supposed that $P(E)$ is defined for all the events E of the sample space. Actually, when the sample space is an uncountably infinite set. $P(E)$ is defined only for a class of events called measurable. However, this restriction need not concern us, as all events of any practical interest are measurable.

Proposition 2.3.

$$P(E^c) = 1 - P(E)$$

In words, Proposition 4.1 states that the probability that an event does not occur is 1 minus the probability that it does occur.

Proposition 2.4. *If $E \subset F$, then $P(E) \leq P(F)$.*

Proof: Since $E \subset F$, it follows that we can express F as

$$F = E \cup E^c F$$

Hence, because E and $E^c F$ are mutually exclusive, we obtain, from Axiom 3.

$$P(F) = P(E) + P(E^c F)$$

which proves the result, since $P(E^c F) \geq 0$.

□

Proposition 2.5.

$$P(E \cup F) = P(E) + P(F) - P(EF)$$

Proof: To derive a formula for $P(E \cup F) = P(E \cup E^c F)$, we first note that $E \cup F$ can be written as the union of the two disjoint events D and $E^c F$. Thus, from Axiom 3, we obtain

$$\begin{aligned} P(E \cup F) &= P(E \cup E^c F) \\ &= P(E) + P(E^c F) \end{aligned}$$

Moreover, since $F = EF \cup E^c F$, we again obtain from Axiom 3

$$P(F) = P(EF) + P(E^c F)$$

or, equivalently,

$$P(E^c F) = P(F) - P(EF)$$

thereby completing the proof.

□

Example 2.6. J is taking two books along on her holiday vacation. With probability 0.5, she will like the first book; with probability 0.4, she will like the second book; and with probability 0.3, she will like both books. What is the probability that she likes neither book?

Solution Let B_i denote the event that J likes book i, $i = 1, 2$. Then the probability that she likes at least one of the books is

$$P(B_1 \cup B_2) = P(B_1) + P(B_2) - P(B_1 B_2) = 0.5 + 0.4 - 0.3 = 0.6$$

Proposition 2.7.

$$\begin{aligned} P(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i) - \sum_{i_1 < i_2} P(E_{i_1} E_{i_2}) \\ + (-1)^{r+1} \sum_{i_1 < i_2 < \cdots < i_r} P(E_{i_1} E_{i_2} \dots E_{i_r}) + \dots \\ + \cdots + (-1)^{n+1} P(E_1 E_2 \dots E_n) \end{aligned}$$

The summation $\sum_{i_1 < i_2 < \cdots < i_r} P(E_{i_1} E_{i_2} \dots E_{i_r})$ is taken over all of the $\binom{n}{r}$ possible subsets of size r of the set $\{1, 2, ..., n\}$.

Remark 2.8. 1. For a noninductive argument for Proposition 2.7, note first that if an outcome of the sample space is not a member of any of the sets E_i, then its probability does not contribute anything to either side of the equality. Now, suppose that an outcome is an exactly m of the events E_i, where $m > 0$. Then, since it is in $\bigcup_i E_i$, its probability is counted once $P\left(\bigcup_i E_i\right)$; also, as this outcome is contained in $\binom{m}{k}$ subsets of the type $E_{i_1} E_{i_2} \dots E_{i_k}$, its probability is counted

$$\binom{m}{1} - \binom{m}{2} + \binom{m}{3} - \cdots \pm \binom{m}{m}$$

times on the right of the equality sign in Proposition 2.7. Thus, for $m > 0$, we must show that

$$1 = \binom{m}{1} - \binom{m}{2} + \binom{m}{3} - \cdots \pm \binom{m}{m}$$

However, since $1 = \binom{m}{0}$, the preceding equation is equivalent to

$$\sum_{i=0}^{m} \binom{m}{i} (-1)^i = 0$$

and the latter equation follows from the binomial theorem, since

$$0 = (-1 + 1)^m = \sum_{i=0}^{m} \binom{m}{i} (-1)^i (1)^{m-i}$$

2. The following is a cucinct way of writing the inclusion-exclusion identity:

$$P(\cup_{i=1}^{n} E_i) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P(E_{i_1} \dots E_{i_r})$$

3. In the inclusion-exclusion identity, going out one term results in an upper bound on the probability of the union, going out two terms results in a lower bound on the probability, going out three terms results in an upper bound on the probability, going out four terms results in a lower bound, and so on. That is, for events $E_1, ..., E_n$. We have

$$P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i) \tag{2}$$

$$P(\cup_{i=1}^{n} E_i) \geq \sum_{i=1}^{n} P(E_i) - \sum_{j<i} P(E_i E_j) \tag{3}$$

$$P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i) - \sum_{j<i} P(E_i E_j) + \sum_{k<j<i} P(E_i E_j E_k) \tag{4}$$

and so on. To prove the validity of these bounds, note the identity

$$\cup_{i=1}^{n} E_i = E_1 \cup E_1^c E_2 \cup E_1^c E_2^c E_3 \cup \cdots \cup E_1^c \dots E_{n-1}^e E_n$$

That is, at least one of the events E_i occurs if E_1 occurs, or if E_1 does not occur but E_2 does, or if E_1 and E_2 do not occur but E_3 does, and so on. Because the right-hand side is the union of disjoint events, we obtain

$$\begin{array}{rcl} P(\cup_{i=1}^{n} E_i) & = & P(E_1) + P(E_1^c E_2) + P(E_1^c E_2^c E_3) + \cdots + P(E_1^c \dots E_{n-1}^c E_n) \\ & = & P(E_1) + \sum_{i=2}^{n} P(E_1^c \dots E_{i-1}^c E_i) \end{array} \tag{5}$$

Now let $B_i = E_1^c \dots E_{i=1}^c = (\cup_{j<i} E_j)^c$ be the event that none of the first $i-1$ events occurs. Applying the identity

$$P(E_i) = P(B_i E_i) + P(B_i^c E_i)$$

shows that

$$P(E_i) = P(E_1^c \dots E_{i-1}^c E_i) + P(E_i \cup_{j<i} E_j)$$

or, equivalently,

$$P(E_1^c \dots E_{i-1}^c E_i) = P(E_i) - P(\cup_{j<i} E_i E_j)$$

Substituting this equation into 5 yields

$$P(\cup_{i=1}^{n} E_i) = \sum_{i} P(E_i) - \sum_{i} P(\cup_{j<i} E_i E_j) \tag{6}$$

Because probabilities are always nonnegative, Inequality 2 follows directly from 3. Now, fixing i and applying Inequality 2 to $P(\cup_{j<i} E_i E_j)$ yields

$$P(\cup_{j<i} E_i E_j) \leq \sum_{j<i} P(E_i E_j)$$

which, by Equation 3, gives Inequality 3. Similarly, fixing i and appylying Inequality 3 to $P(\cup_{j<i} E_i E_j)$ yields

$$\begin{array}{rcl} P(\cup_{j<i} E_i E_j) & \leq & \sum_{j<i} P(E_i E_j) - \sum_{k<j<i} P(E_i E_j E_i E_k) \\ & = & \sum_{j<i} P(E_i E_j) - \sum_{k<j<i} P(E_i E_j E_k) \end{array}$$

which, by Equation, gives Inequality 5. The next inclusion-exclusion inequality is now obtained by fixing i and applying Inequality 4 to $P(\cup_{j<i} E_i F_j)$, and so on.

3 CONDITIONAL PROBABILITY AND INDEPENDENCE

Go back to Table of Contents. Please click TOC

3.1 Conditional Probabilities

Suppose that we toss 2 dice, and suppose that each of the 36 possible outcomes is equally likely to occur and hence has probability 1/36. Suppose that we observe that the first die is a 3. Then, given this information, what is the probability that the sum of the 2 dice equals 8? This is a calculation of conditional probability.

If we let E and F denote, respectively, the event that the sum of the dice is 8 and the event that the first die is a 3, then the probability just obtained is called the conditional probability that E occurs given that F has occurred and is denoted by

$$P(E|F)$$

A general formula for $P(E|F)$ that is valid for all events E and F is derived in the same manner: If the event F occurs, then, in order for E to occur, it is necessary that the actual occurrence be a point both in E and in F; that is, it must be in EF. now, since we know that F has occurred, it follows that F becomes our new, or reduced, sample space; hence, the probability that the event EF occurs will equal the probability of EF relative to the probability of F. That is, we have the following definition.

Definition 3.1. If $P(F) > 0$, then

$$P(E|F) = \frac{P(EF)}{P(F)}$$

Example 3.2. Joe is 80 percent certain that his missing key is in one of the two pockets of his hanging jacket, being 40- percent certain it is in the left-hand pocket and 40 percent certain it is in the right-hand pocket. If a search of the left-hand pocket does not find the key, what is the conditional probability that it is in the other pocket?

Solution If we let L be the event that the key is in the left-hand pocket of the jacket, and R be the event that it is in the riht-hand pocket, then the desired probability $P(R|L^c)$ can be obtained as follows:

$$\begin{aligned} P(R|L^c) &= \tfrac{P(RL^c)}{P(L^c)} \\ &= 2/3 \end{aligned}$$

□

Example 3.3. Suppose that an urn contains 8 red balls and 4 white balls. We draw 2 balls from the urn without replacement. (a) If we assume that at each draw, each ball in the urn is equally likely to be chosen, what is the probability that both balls drawn are red? (b) Now suppose that hte balls have different weights, with each red balls having weight r and each white ball having weight w. Suppose that the probability that a given ball in the urn is the next one selected is its weight divided by the sum of the weights of all balls currently in the urn. Now what is the probability that both balls are red?

Solution Let R_1 and R_2 denote, respectively, the events that the first and second balls drawn are red. Now, given that the first ball selected is red, there are 7 remaining red balls and 4 white balls, so $P(R_2|R_1) = 7/11$. As $P(R_1)$ is clearly 8/12, the desired probability is

$$\begin{aligned} P(R_1R_2) &= P(R_1)P(R_2|R_1) \\ &= \tfrac{21}{3}\tfrac{7}{11} = \tfrac{14}{33} \end{aligned}$$

This probability could have been by $P(R_1R_2) = \binom{8}{2}/\binom{12}{2}$.

For part 9b), we again let R_i be the event that the ith ball chosen is red and use

$$P(R_1R_2) = P(R_1)P(R_2|R_1)$$

Now, number the red balls, and let B_i, $i = 1, ..., 8$ be the event that the first ball drawn is red ball number i. Then

$$P(R_1) = P(\cup_{i=1}^{8} B_i) = \sum_{i=1}^{8} P(B_i) = 8\frac{r}{8r + 4w}$$

Moreover, given that the first ball is red, the urn then contains 7 red and 4 white balls. Thus, by an argument similar to the preceeding one,

$$P(R_2|R_1) = \frac{7r}{7r+4w}$$

Hence, the probability that both balls are red is

$$P(R_1R_2) = \frac{8r}{8r4w}\frac{7r}{7r+4w}$$

□

Proposition 3.4. *The multiplication rule says*

$$P(E_1E_2E_3\dots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2)\dots P(E_n|E_1\dots D_{n-1})$$

To prove the multiplication rule, just apply the definition of conditional probability to its right-hand side, giving

$$P(E_1)\frac{P(E_1E_2)}{P(E_1)}\frac{P(E_1E_2E_3)}{P(E_1E_2)}\dots\frac{P(E_1E_2\dots E_n)}{P(E_1E_2\dots E_{n-1}} = P(E_1E_2\dots E_n)$$

3.2 Bayes's Formula

Let E and F be events. We may express E as

$$E = EF \cup EF^c$$

for, in order for an outcome to be in E, it must either be in both E and F or be in E but not in F. As EF and EF^c are clearly mutually exclusive, we have, by Axiom 3,

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)[1-P(F)] \end{aligned}$$

Example 3.5. At a certain stage of a criminal investigation, the inspector in charge is 60 percent convinced of the guilt of a certain suspect. Suppose, however, that a new piece of evidence which shows that the criminal has a certain characteristic (such as left-handedness, baldness, or brown hair) is uncovered. If 20 percent of the population possesses this chracteristic, how certain of the guilt of the suspect should the inspector now be if it turns out that the suspect has the characteristic?

Solution Letting G denote the event that the suspect is guilty and C the event that he possesses the characteristic of the criminal, we have

$$\begin{aligned} P(G|C) &= \frac{P(GC)}{P(C)} \\ &= \frac{P(C|G)P(G)}{P(C|G)P(G)+P(C|G^c)P(G^c)} \\ &= \frac{1(0.6)}{1(0.6)+(0.2)(0.4)} \\ &\approx 0.882 \end{aligned}$$

□

Definition 3.6. The odds of an event A are defined by

$$\frac{P(A)}{P(A^c)} = \frac{P(A)}{1-P(A)}$$

That is, the odds of an event A tell how much more likely it is that the event A occurs that it is that it does not occur. For instance, if $P(A) = 2/3$, then $P(A) = 2P(A^c)$, so the odds are 2. If the odds are equal to α, then it is common to say that the odds are "α to 1" in favor of the hypothesis.

Consider now a hypothesis H that is true with probability $P(H)$, and suppose that new evidence E is introduced. Then, the conditional probabilities, given the evidence E, that H is true and that H is not true are respectively given by

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \quad P(H^c|E) = \frac{P(E|H^c)P(H^c)}{P(E)}$$

Therefore, the new odds after the evidence E has been introduced are

$$\frac{P(H|E)}{P(E)} = \frac{P(H)}{PH^c)} \frac{P(E|H)}{P(E|H^c)} \tag{7}$$

That is, the new value of the odds of H is the old value multiplied by the ratio of the conditional probability of the new evidence given that H is true to the conditional probability given that H is not true. Thus, equation 7 verifies the result of example 3.5.

Definition 3.7. Suppose that $F_1, F_2, ..., F_n$ are mutually exclusive events such that

$$\bigcup_{i=1}^{n} F_i = S$$

In other words, exactly one of the events $F_1, F_2, ..., F_n$ must occur. By writing

$$E = \bigcup_{i=1}^{n} EF_i$$

and using the fact that the events EF_i, $i = 1, ..., n$ mutually exclusive, we obtain

$$\begin{aligned} P(E) &= \sum_{i=1}^{n} P(EF_i) \\ &= \sum_{i=1}^{n} P(E|F_i)P(F_i) \end{aligned}$$

Proposition 3.8.

$$\begin{aligned} P(E_j|E) &= \frac{P(EF_j)}{P(E)} \\ &= \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)} \end{aligned} \tag{8}$$

Equation 8 is known as Bayes's formula.

3.3 Independent Events

The previous examples in this chapter show that $P(E|F)$, the conditional probability of E given F, is not generally equal to $P(E)$, the unconditional probability of E. In other words, knowing that F has occurred generally changes the chances of E's occurrence. In this special cases where $P(E|F)$ does in fact equal $P(E)$, we say that E is independent of F. That is, E is independent of F if knowledge that F has occurred does not change the probability that E occurs.

Since $P(E|F) = P(EF)/P(F)$, it follows that E is independent of F if

$$P(EF) = P(E)P(F) \tag{9}$$

Definition 3.9. Two events E and F are said to be independent if Equation 9 holds. Two events E and F that are not independent are said to be dependent.

Example 3.10. A card is selected at random from an ordinary deck of 52 playing cards. If E is the event that the selected card is an ace and F is the event that it is a spade, then E and F are independent. This follows because $P(EF) = 1/52$, whereas $P(E) = 4/52$ and $P(F) = 13/52$.

□

Proposition 3.11. *If E and F are independent, then so are E and F^c.*

Proof Assume that E and F are independent. Since $E = EF \cup EF^c$ and EF and ER^c are obviously mutually exclusive, we have

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E)P(F) + P(EF^c) \end{aligned}$$

or, equivalently,

$$\begin{aligned} P(EF^c) &= P(E)[1 - P(F)] \\ &= P(E)P(F^c) \end{aligned}$$

and the result is proved.

□

Definition 3.12. Three events E, F, and G are said to be independent if

$$P(EFG) = P(E)P(F)P(G)$$

$$P(EF) = P(E)P(F)$$

$$P(EG) = P(E)P(G)$$

$$P(FG) = P(F)P(G)$$

Note that if E, F, and G are independent, then E will be independent of any event formed from F and G. For instance, E is independent of $F \cup G$, since

$$\begin{aligned} P(E(F \cup G)] &= P(EF \cup EG) \\ &= P(EF) + P(EG) - P(EFG) \\ &= P(E)P(F) + P(E)P(G) - P(E)P(FG) \\ &= P(E)[P(F) + P(G) - P(FG)] \\ &= P(E)P(F \cup G) \end{aligned}$$

Of course, we may also extend the definition of independence to more than three events. The events E_1, E_2, ..., E_n are said to be independent if for every subset $E_{1'}$, $E_{2'}$, ..., $E_{r'}$, $r \leq n$ of these events,

$$P(E_{1'}E_{2'} \dots E_{r'}) = P(E_{1'})P(E_{2'}) \dots P(E_{r'})$$

Finally, we define an infinite set of events to be independent if every finite subset of those events is independent.

Example 3.13. An infinite sequence of independent trials is to be performed. Each trial results in a success with probability p and a failure with probabilitiy $1 - p$. What is the probability that

(a) at least 1 success occurs in the first n trials;

(b) exactly k successes occur in the first n trials;

(c) all trials result in successes?

Solution In order to determine the probability of at least 1 success in the first n trials, it is easiest to compute first the probability of the complementary event: that of no successes in the first n trials. If we let E_i denote the event of a failure on the ith trial, then the probability of no successes is, by independence,

$$P(E_1E_2 \dots E_n) = P(E_1)P(E_2) \dots P(E_n) = (1-p)^n$$

Hence, the answer to part (a) is $1 - (1-p)^n$.

To compute the answer to part (b), consider any particular sequence of the first n outcomes containing k successes and $n - k$ failures. Each one of these sequences will, by the assumed independence of trials, occur with probability $p^k(1-p)^{n-k}$. Since there are $\binom{n}{k}$ such sequences [there are $n!/k!(n-k)!$ permutations of k successes and $n - k$ failures], the desired probability in part (b) is

$$P\{\text{exactly } k \text{ successes}\} = \binom{n}{k} p^k (1-p)^{n-k}$$

To answer part (c), we note that, by part (a), the probability of the first n trials all resulting in success is given by

$$P(E_1^c E_2^c \dots E_n^c) = p^n$$

Thus, using the continuity property of probability, we see that the desired probability is given by

$$\begin{aligned} p\left(\bigcap_{i=1}^{\infty} E_i^c\right) &= P\left(\lim_{n\to\infty} \bigcap_{i=1}^{n} E_i^c\right) \\ &= \lim_{n\to\infty} P\left(\bigcap_{i=1}^{n} E_i^c\right) \\ &= \lim_n p^n = \begin{cases} 0 & \text{if } p < 1 \\ 1 & \text{if } p = 1 \end{cases} \end{aligned}$$

Example 3.14. Independent trials consisting of rolling a pair of fair are performed. What is the probability that an outcome of 5 appears before an outcome of 7 when the outcome of a roll is the sum of the dice?

Solution If we let E_n denote the event that no 5 or 7 appears on the first $n-1$ trials and a 5 appears on the nth trial, then the desired probability is

$$P\Big(\bigcap_{n=1}^{\infty} E_n\Big) = \sum_{n=1}^{\infty} P(E_n)$$

Now, since $P\{5 \text{ on any trial}\} = 4/36$ and $P\{7 \text{ on any trial}\} = 6/36$, we obtain, by the independence of trials,

$$P(E_n) = (1 - \frac{10}{36})^{n-1} \frac{4}{36}$$

Thus,

$$\begin{aligned} P\Big(\bigcup_{n=1}^{\infty} E_n\Big) &= \tfrac{1}{9} \sum_{n=1}^{\infty} \Big(\tfrac{13}{18}\Big)^{n-1} \\ &= \tfrac{1}{9} \tfrac{1}{1} - \tfrac{13}{18} \\ &= \tfrac{2}{5} \end{aligned}$$

This result could also have been obtained by the use of conditional probabilities. If we let E be the event that a 5 occurs before a 7, then we can obtain the desired probability, $P(E)$, by conditioning on the outcome of the first trial, as follows: let F be the event that the first trial results in a 5, let G be the event that it results in a 7, and H be the event that the first trial results in neither a 5 nor a 7. Then, conditioning on which one of these events occurs gives

$$P(E) = P(E|F)P(F) + P(E|G)P(G) + P(E|H)P(H)$$

However,

$$P(E|F) = 1$$

$$P(E|G) = 0$$

$$P(E|H) = P(E)$$

The first two equalities are obvious. The third follows because if the first outcome results in neither a 5 nor a 7, then at that point the situation is exactly as it was when the problem first started — namely, the experimenter will continually roll a pair of fair dice until either a 5 or 7 appears. Moreover, the trials are independent; therefore the outcome of the first trial will have no effect on subsequent rolls of the dice. Since $P(F) = 4/36$, $P(G) = 6/36$, and $P(H) = 26/36$, it follows that

$$P(E) = 1/9 + P(E)13/18$$

or

$$P(E) = 2/5$$

The reader should note that the answer is quite intuitive. That is, because a 5 occurs on any roll with probability 4/36 and a 7 with probability 6/36, it seems intuitive that the odds that a 5 appears before a 7 should be 6 to 4 against. The probability should then be 4/10, as indeed it is.

The same argument shows that if E and F are mutually exclusive events of an experiment, then, when independent trials of the experiment are performed, the event E will occur before the event F with probability

$$\frac{P(E)}{P(E) + P(F)}$$

□

4 RANDOM VARIABLES

Go back to Table of Contents. Please click TOC

4.1 Random Variables

When an experiment is performed, we are frequently interested mainly in some function of the outcome as opposed to the actual outcome itself.

Example 4.1. Suppose that there are N distinct types of coupons and that each time one obtains a coupon, it is, independently of previous selections, equally likely to be any one of the N types. One random variable of itnerest is T, the number of coupons that need to be collected until one obtains a complete set of at least one of each type. Rather than derive $P\{T = n\}$ directly, let us start by considering the probability that T is greater than n. To do so, fix n and define the events A_1, A_2, ..., A_N as follows: A_j is the event that no type j coupon is contained among the first n coupons collected, $j = 1, ..., N$. Hence,

$$\begin{aligned} P\{T > n\} &= P\left(\bigcup_{j=1}^{N} A_j\right) \\ &= \sum_j P(A_j) - \sum_{j-1<j_2} \sum P(A_{j1}A_{j2}) + \\ &\quad +(-1)^{k+1} \sum_{j_1<j_2<\ldots j_k} \sum\sum P(A_{j1}A_{j2}\ldots A_{jk})\ldots \\ &\quad (-1)^{N+1} P(A_1A_2\ldots A_N) \end{aligned}$$

Now, A_j will occur if each of the n coupons collected is not of type j. Since each of the coupons will not be of type j with probability $(N-1)/N$, we have, by the assumed independence of the types of successive coupons,

$$P(A_j) = \left(\frac{N-1}{N}\right)^n$$

Also, the event $A_{j1}A_{j2}$ will occur if none of the first n coupons collected is of either type j_1 or type j_2. Thus, again using independence, we see that

$$P(A_{j1}A_{j2}) = \left(\frac{N-2}{N}\right)^n$$

The same reasoning gives

$$P(A_{j1}A_{j2}\ldots A_{jk}) = \left(\frac{N-k}{N}\right)^n$$

and we see that for $n > 0$,

$$\begin{aligned} P\{T > n\} &= N\left(\tfrac{N-1}{N}\right)^n - \binom{N}{2}\left(\tfrac{N-2}{N}\right)^n + \binom{N}{3}\left(\tfrac{N-3}{N}\right)^n - \ldots \\ &\quad +(-1)^N \binom{N}{N-1}\left(\tfrac{1}{N}\right)^n \\ &= \sum_{j=1}^{N-1} \binom{N}{i}\left(\tfrac{N-i}{N}\right)^n (-1)^{i+1} \end{aligned}$$

The probability that T equals n can now be obtained from the preceeding formula by the use of

$$P\{T > n-1\} = P\{T = n\} + P\{T > n\}$$

or, equivalently,

$$P\{T = n\} = P\{T > n-1\} - P\{T > n\}$$

Another random variable of interest is the number of distinct types of coupons that are contained in the first n selections – call this random variable D_n. To compute $P\{D_n = k\}$, let us start by fixing attention on a particular set of k distinct types, and let us then determine the probability that this set constitutes the set of distinct types obtained in the first n selections. Now, in order for this to be the situation, it is necessary and sufficient that of the first n coupons obtained,

$$A: \text{ each is one of these } k \text{ types}$$

$$B: \text{ each of these } k \text{ types is represented}$$

Now, each coupon selected will be one of the k types with probability k/N, so the probability that A will be valid is $(k/N)^n$. Also, given that a coupon is of one of the k types under consideration, it is easy to see that it is equally likely to be of any one of these k types. Hence, the conditional probability of B given that A occurs is the same as the probability that a set of n coupons, each equally likely to be any of k possible types, contains a complete set of all k types. But this is just the probability that the number needed to amass a complete set, when choosing among k types, is less than or equal to n and is thus obtainable with k replacing N. Thus, we have

$$P(A) = \left(\frac{k}{N}\right)^n$$

$$P(B|A) = 1 - \sum_{i=1}^{k-1} \sum \binom{k}{i} \left(\frac{k-i}{k}\right)^n (-1)^{i+1}$$

Finally, as there are $\binom{N}{k}$ possible choices for the set of k types, we arrive at

$$\begin{aligned} P\{D_n = k\} &= \tbinom{N}{k} P(AB) \\ &= \tbinom{N}{k} \left(\tfrac{k}{N}\right)^n \left[1 - \textstyle\sum_{i=1}^{k-1} \tbinom{k}{i} \left(\tfrac{k-i}{k}\right)^n (-1)^{i+1}\right] \end{aligned}$$

□

Remark 4.2. Since one must collect at least N coupons to obtain a complete set, it follows that $P\{T > n\} = 1$ if $n < N$. Therefore, we obtain the interesting combinatorial identity that for integers $1 \leq n < N$,

$$\sum_{i=1}^{N-1} \binom{N}{i} \left(\frac{N-i}{N}\right)^n (-1)^{i+1} = 1$$

which can be written as

$$\sum_{i=0}^{N-1} \binom{N}{i} \left(\frac{N-i}{N}\right)^n (-1)^{i+1} = 0$$

or, upon multiplying by $(-1)^N N^n$ and letting $j = N - i$,

$$\sum_{j=1}^{N} \binom{N}{j} j^n (-1)^{j-1} = 0 \; 1 \leq n < N$$

□

4.2 Discrete Random Variables

A random variable that can take on at most a countable number of possible values is said to be discrete. For a discrete random variable X, we define the probability mass function $p(a)$ of X by

$$p(a) = P\{X = a\}$$

The probability mass function $p(a)$ is positive for at most a countable number of values of a. That is, if X must assume one of the values x_1, x_2, ..., then

$$p(x_i) \geq 0 \text{ for } i = 1, 2, ...$$

$$p(x) = 0 \text{ for all other values of } x$$

Since X must take on one of the values x_i, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

It is often instructive to present the probability mass function in a graphical format by plotting $p(x_i)$ on the y-axis against x_i on the x-axis.

Example 4.3. The probability mass function of a random variable X is given by $p(i) = c\lambda^i/i!$, $i = 0, 1, 2, ...$, where λ is some positive value. Find (a) $P\{X = 0\}$ and (b) $P\{X > 2\}$.

Solution Since $\sum_{i=0}^{\infty} p(i) = 1$, we have

$$c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = 1$$

which, because $e^x = \sum_{i=0}^{\infty} x^i/i!$, implies that

$$ce^{\lambda} = 1 \text{ or } c = e^{-\lambda}$$

Hence,

(a) $P\{X = 0\} = e^{-\lambda}\lambda^0/0! = e^{-\lambda}$

(b)

$$\begin{aligned} P\{X > 2\} &= 1 - P\{X \leq 2\} = 1 - P\{X = 0\} - P\{X = 1\} - P\{X = 2\} \\ &= 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2 e^{-\lambda}}{2} \end{aligned}$$

□

4.3 Expected Value

One of the most important concepts in probability theory is that of the expectation of a random variable. If X is a discrete random variable having a probability mass function $p(x)$, then the expectation, or the expected value, of X, denoted by $E[X]$, is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

In words, the expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it.

4.4 Expectation of a Function of a Random Variable

Suppose that we are given a discrete random variable along with its probability mass function and that we want to compute the expected value of some function of X, say $g(X)$. How can we do this? One way is as follows: since $g(X)$ is itself a discrete random variable, it has a probability mass function, which can be determined from the probability mass function of X. Once we have determined the probability mass function of $g(X)$, we can compute $E[g(X)]$ by using the definition of expected value.

Proposition 4.4. *If X is a discrete random variable that takes on one of the values x_i, $i \geq 1$, with respective probabilities $p(x_i)$, then, for any real-valued function g,*

$$E[g(X)] = \sum_i g(x_i)p(x_i)$$

Proof Suppose that y_j, for $j \geq 1$, represent the different values of $g(x_i)$, $i \geq 1$. Then, grouping all the $g(x_i)$ having the same value gives

$$\begin{aligned} \sum_i g(x_i)p(x_i) &= \sum_j \sum_{i:g(x_i)=y_j} g(x_i)p(x_i) \\ &= \sum_j \sum_{i:g(x_i)=y_j} y_j p(x_i) \\ &= \sum_j y_j \sum_{i:g(x_i)=y_j} p(x_i) \\ &= \sum_j y_j P\{g(X) = y_j\} \\ &= E[g(X)] \end{aligned}$$

□

4.5 Variance

Definition 4.5. If X is a random variable with mean μ, then the variance of X, denoted by $\text{Var}(X)$, is defined by

$$\text{Var}(X) = E[(X - \mu)^2]$$

An alternative formula for $\text{Var}(X)$ is derived as follows:

$$\begin{aligned}\text{Var}(X) &= E[(X-\mu)^2] \\ &= \sum(x-\mu)^2 p(x) \\ &= \sum_x (x^2 - 2\mu x + \mu^2)p(x) \\ &= \sum_x x^2 p(x) - 2\mu \sum_x xp(x) + \mu^2 \sum_x p(x) \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2\end{aligned}$$

That is,

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

In words, the variance of X is equal to the expected value of X^2 minus the square of its expected value.

Proposition 4.6. *A useful identify is that for any constants a and b,*

$$Var(aX+b) = a^2\, Var(X)$$

To prove this equality, let $\mu = E[X]$ and note that $E[aX+b] = a\mu + b$. Therefore,

$$\begin{aligned}Var(aX+b) &= E[(aX+b-a\mu-b)^2] \\ &= E[a^2(X-\mu)^2] \\ &= a^2 E[(X-\mu)^2] \\ &= a^2 E[(X-\mu)^2] \\ &= a^2\, Var(X)\end{aligned}$$

Remark 4.7. Analogous to the means being the center of gravity of a distribution of mass, the variance represents, in the terminology of mechanics, the moment of inertia.

Remark 4.8. The square root of the $\text{Var}(X)$ is called the standard deviation of X, and we denote it by $\text{SD}(X)$. That is,

$$SD(X) = \sqrt{\text{Var}(X)}$$

4.6 Bernoulli and Binomial Random Variables

Suppose that a trial, or an experiment, whose outcome can be classified as either a success or a failure is performed. If we let $X = 1$ when the outcome is a success and $X = 0$ when it is a failure, then the probability mass function of X is given by

$$p(0) = P\{X=0\} = 1-p;\ P(1) = P\{X=1\} = p \tag{10}$$

A random variable X is said to be a Bernoulli random variable if its probability mass function is given by 10 for some $p \in (0,1)$.

Suppose now that n independent trials, each of which results in a success with probability p or in a failure with probability $1-p$, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a binomial random variable with parameters (n,p). Thus, a Bernoulli random variable is just a binomial random variable with parameters $(1,p)$.

The probability mass function of a binomial random variable having parameters (n,p) is given by

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}\ \ i = 0,1,...,n \tag{11}$$

We will now examine the properties of a binomial random variable with parameters n and p. To begin, let us compute its expected value and variance. To begin, note that

$$\begin{aligned}E[X^k] &= \textstyle\sum_{i=0}^n i^k \binom{n}{i} p^i (1-p)^{n-i} \\ &= \textstyle\sum_{i=1}^n i^k \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Using the identity

$$i\binom{n}{i} = n\binom{n-1}{i-1}$$

gives that

$$\begin{aligned} E[X^k] &= np\sum_{i=1}^{n} i^{k-1}\binom{n-1}{i-1}p^{i-1}(1-p)^{n-i} \\ &= np\sum_{j=0}^{n-1}(j+1)^{k-1}\binom{n-1}{j}p^{j}(1-p)^{n-1-j} \text{ by letting } j=i-1 \\ &= npE[(Y+1)^{k-1}] \end{aligned}$$

where Y is a binomial random variable with parameters $n-1$, p. Setting $k=1$ in the preceding equation and using the preceding formula for the expected value of a binomial random variable yields

$$\begin{aligned} \text{Var}(X) &= E[X^2]-(E[X])^2 \\ &= np[(n-1)p+1]-(np)^2 \\ &= np(1-p) \end{aligned}$$

Proposition 4.9. *If X is a binomial random variable with parameters (n,p), where $0<p<1$, then as k goes from 0 to n, $P\{X=k\}$ first increases monotonically and then decreases monotonically, reaching its largest value when k is the largest integer less than or equal to $(n+1)p$.*

Proof We prove the proposition by considering $P\{X=k\}/P\{X=k-1\}$ and determining for what values of k it is greater or less than 1. Now,

$$\begin{aligned} \frac{P\{X=k\}}{P\{X=k-1\}} &= \frac{\frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}}{\frac{n!}{(n-k+1)!(k-1)!}p^{k-1}(1-p)^{n-k-1}} \\ &= \frac{(n-k+1)p}{k(1-p)} \end{aligned}$$

Hence, $P\{X=k\}\geq P(X=k-1)$ if and only if

$$(n-k+1)p\geq k(1-p)$$

or, equivalently, if and only if

$$k\leq (n+1)p$$

and the proposition is proved.

□

Suppose that X is binomial with parameters (n,p). They key to computing its distribution function

$$P\{X\leq i\}=\sum_{k=0}^{i}\binom{n}{k}p^k(1-p)^{n-k}\ i=0,1,...,n$$

is to utilize the following relationship between $P\{X=k+1\}$ and $P\{X=k\}$, which was established in the proof above:

$$P\{X=k+1\}=\frac{p}{1-p}\frac{n-k}{k+1}P\{X=k\} \tag{12}$$

4.7 Geometric Random Variable

Suppose that independent trials, each having a probability p, $0<p<1$, of being a success, are performed until a success occurs. If we let X equal the number of trials required, then

$$P\{X=n\}=(1-p)^{n-1}p\ n=1,2,... \tag{13}$$

Example 4.10. An urn contains N white and M black balls. Balls are randomly selected, one at a time, until a black one is obtained. If we assume that each ball selected is replaced before the next one is drawn, what is the probability that

(a) exactly n draws are needed?

(b) at least k draws are needed?

Solution If we let X denote the number of draws needed to select ta black ball, then X satisfies Equation 13 with $p=M/(M+N)$. Hence,

(a)

$$\begin{aligned} P\{X=n\} &= \left(\frac{N}{M+N}\right)^{n-1}\frac{M}{M+N} \\ &= \frac{MN^{n-1}}{(M+N)^n} \end{aligned}$$

(b)

$$\begin{aligned} P\{X \geq k\} &= \frac{M}{M+N} \sum_{n=k}^{\infty} \left(\frac{N}{M+N}\right)^{n-1} \\ &= \left(\frac{M}{M+N}\right)\left(\frac{M}{M+N}\right)^{k-1} / \left[1 - \frac{N}{M+N}\right] \end{aligned}$$

Of course, part (b) could have been obtained directly, since the probability that at least k trials are necessary to obtain a success is equal to the probability that the first $k-1$ trials are all failures. That is, for a geometric random variable,

$$P\{X \geq k\} = (1-p)^{k-1}$$

4.8 Cumulative Distribution Function

Recall that for the distribution function F of X, $F(b)$ denotes the probability that the random variable X takes on a value that is less than or equal to b. Following are some properties of the cumulative distribution function (c.d.f.) F:

1. F is a nondecreasing function; that is, if $a < b$, then $F(a) \leq F(b)$.
2. $\lim_{b\to\infty} F(b) = 1$
3. $\lim_{b\to-\infty} F(b) = 0$
4. F is right continuous. That is, for any b and any decreasing sequence b_n, $n \leq 1$, that converges to b, $\lim_{b\to\infty} F(b_n) = F(b)$.

Property 1 follows, as was noted before, because for $a < b$, the event whose union is the event $\{X < \infty\}$. hence, by the continuity property of probabilities,

$$\lim_{n\to\infty} P\{X \leq b_n\} = P\{X < \infty\} = 1$$

which proves property 2.

The proof of property 3 is similar and is left as an exercise. To prove property 4, we note that if b_n decreases to b, then $\{X \leq b_n\}$, $n \geq 1$, are decreasing events whose intersections is $\{X \leq b\}$. The continuity property then, yields

$$\lim_{n\to\infty} P\{X \leq b_n\} = P\{X \leq b\}$$

5 CONTINUOUS RANDOM VARIABLES

Go back to Table of Contents. Please click TOC

We say that X is a continuous random variable if there exists a nonnegative function f, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers,

$$P\{X \in B\} = \int_B f(x)dx \tag{14}$$

The function f is called the probability density function of the random variable X. In words, Equation 14 states that the probability that X will be in B may be obtained by integrating the probability density function over the set B. Since X must assume some value, f must satisfy

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)dx$$

All probability statements about X can be answered in terms of f. For instance, letting $B = [a, b]$, we obtain

$$P\{a \leq X \leq b\} \int_a^b f(x)dx \tag{15}$$

Example 5.1. Suppose that X is a continuous random variable whose probability density function is

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) What is the value of C?

(b) Find $P\{X > 1\}$.

Solution (a) Since f is a probability density function, we must have $\int_{\infty}^{\infty} f(x)dx = 1$, implying that

$$C \int_0^2 (4x - 2x^2) = 1$$

or

$$C\left[2x^2 - \frac{2}{3}x^3\right]\Big|_{x=0}^{x=2} = 1$$

or

$$C = \frac{3}{8}$$

Hence,
(b) $P\{X > 1\} \int_1^{\infty} f(x)dx = \frac{3}{8} \int_1^2 (4x - 2x^2)dx = \frac{1}{2}$

□

Proposition 5.2. *If X is a continuous random variable with probability density function $f(x)$, then, for any real-valued function g,*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

An application of this proposition is

$$\begin{aligned} E[e^X] &= \int_0^1 e^x dx \text{ since } f(x) = 1, 0 < x < 1 \\ &= e - 1 \end{aligned}$$

Lemma 5.3. *For a nonnegative random variable Y,*

$$E[Y] = \int_0^{\infty} P\{Y > y\}dy$$

Proof We present a proof when Y is a continuous random variable with probability density function f_Y. we have

$$\int_0^\infty P\{Y > y\}dy = \int_0^\infty \int_y^\infty f_Y(x)dxdy$$

where we have used the fact that $P\{Y > y\} = \int_y^\infty f_Y(x)dx$. Interchanging the order of integration in the preceding equation yields

$$\begin{array}{rcl} \int_0^\infty P\{Y > y\}dy & = & \int_0^\infty \left(\int_0^x dy \right) f_Y(x)dx \\ & = & \int_0^\infty x f_Y(x)dx \\ & = & E[Y] \end{array}$$

Proof From Lemma 5.3, for any function g for which $g(x) \geq 0$,

$$\begin{array}{rcl} E[g(X)] & = & \int_0^\infty P\{g(X) > y\}dy \\ & = & \int_0^\infty \int_{x:g(x)>y} f(x)dxdy \\ & = & \int_{x:g(x)>0} g(x)f(x)dx \end{array}$$

which completes the proof.

□

Corollary 5.4. *If a and b are constants, then*

$$E[aX + b] = aE[X] + b$$

5.1 Uniform Random Variable

A random variable is said to be uniformly distributed over the interval $(0, 1)$ if its probability density function is given by

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

Note that Equation 16 is a density function, since $f(x) \geq 0$ and $\int_{-\infty}^\infty f(x)dx = \int_0^1 dx = 1$. Because $f(x) > 0$ only when $x \in (0, 1)$, it follows that X must assume a value in interval $(0, 1)$. Also since $f(x)$ is constant for $x \in (0, 1)$, X is just as likely to be near any value in $(0, 1)$ as it is to be near any other value. To verify this statement, note that for any $0 < a < b < 1$,

$$P\{a \leq X \leq b\} = \int_a^b f(x)dx = b - a$$

In other words, the probability that X is in any particular subinterval of $90, 1)$ equals the length of that subinterval.

In general, we say that X is a uniform random variable on the interval (α, β) if the probability density function of X is given by

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Since $F(a) = \int_{-\infty}^a f(x)dx$, it follows from Equation **??** that the distribution function of a uniform random variable on the interval (α, β) is given by

$$F(a) = \begin{cases} 0 & a \leq \alpha \\ \frac{a - \alpha}{\beta - \alpha} & \alpha < a < \beta \\ 1 & \alpha \geq \beta \end{cases}$$

5.2 Normal Random Variables

We say that X is a normal random variable, or simply that X is normally distributed, with parameters μ and σ^2 if the density of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$$

The density function is a bell-shaped curve that is symmetric about μ.

To prove that $f(x)$ is indeed a probability density function, we need to show that

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2}dx = 1$$

Making the substitution $y = (x-\mu)/\sigma$, we see that

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}dy$$

Hence, we must show that

$$\int_{-\infty}^{\infty} e^{-y^2/2}dy = \sqrt{2\pi}$$

Toward this end, let $I = \int_{-\infty}^{\infty} e^{-y^2/2}dy$. Then

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-y^2/2}dy \int_{-\infty}^{\infty} e^{-x^2/2}dx \\ &= \int_{-\infty}^{-\infty}\int_{-\infty}^{\infty} e^{-(y^2+x^2)/2}dydx \end{aligned}$$

we now evaluate the double integral by means of a change of variables to polar coordinates. Let $x = r\cos\theta$, $y = r\sin\theta$, and $dydx = rd\theta dr$. Thus,

$$\begin{aligned} I^2 &= \int_0^{\infty}\int_0^{2\pi} e^{-r^2/2}rd\theta dr \\ &= 2\pi\int_0^{\infty} re^{-r^2/2}dr \\ &= -2\pi e^{-r^2/2}|_0^{\infty} \\ &= 2\pi \end{aligned}$$

Hence, $I = \sqrt{2\pi}$, and the result is proved.

An important fact about normal random variables is that if X is normally distributed with parameters μ and σ^2, then $Y = aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$. To prove this statement, suppose that $a > 0$. The proof when $a < 0$ is similar. Let F_Y denote the cumulative distribution function of Y. Then

$$\begin{aligned} F_Y(x) &= P\{Y \le x\} \\ &= P\{aX + b \le x\} \\ &= P\left\{X \le \tfrac{x-b}{a}\right\} \\ &= F_X\left(\tfrac{x-b}{a}\right) \end{aligned}$$

where F_X is the cumulative distribution function of X. By differentiation, the density function of Y is then

$$\begin{aligned} f_Y(x) &= \tfrac{1}{a}f_X\left(\tfrac{x-b}{a}\right) \\ &= \tfrac{1}{\sqrt{2\pi}a\sigma}\exp\left\{-\left(\tfrac{x-b}{a}-\mu\right)^2/2\sigma^2\right\} \\ &= \tfrac{1}{\sqrt{2\pi}a\sigma}\exp -(x-b-a\mu)^2/2(a\sigma)^2 \end{aligned}$$

which shows that Y is normal with parameters $a\mu + b$ and $a^2\sigma^2$.

Example 5.5. Find $E[X]$ and $\mathrm{Var}(X)$ when X is a normal random variable with parameters μ and σ^2.

Solution Let us start by finding the mean and variance of the standard normal random variable $Z = (X-\mu)/\sigma$. We have

$$\begin{aligned} E[Z] &= \int_{-\infty}^{\infty} xf_Z(x)dx \\ &= \tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\exp(-x^2/2)dx \\ &= -\tfrac{1}{\sqrt{2\pi}}\exp(-x^2/2)|_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Thus,

$$\begin{aligned}\text{Var}(Z) &= E[Z^2] \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 \exp(-x^2/2)dx\end{aligned}$$

Integration by parts (with $\mu = x$ and $dv = xe^{-x^2/2}$) now gives

$$\begin{aligned}\text{Var}(Z) &= \frac{1}{\sqrt{2\pi}}\Big(-xe^{-x^2/2}|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2dx}\Big) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}dx \\ &= 1\end{aligned}$$

Because $X = \mu + \sigma Z$, the preceding yields the results

$$E[X] = \mu + \sigma E[Z] = \mu$$

and

$$\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2$$

□

It is customary to denote the cumulative distribution function of a standard normal random variable by $\Phi(x)$. That is,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}dy$$

Theorem 5.6. *The DeMoivre-Laplace Limit Theorem. If S_n denotes the number of successes that occur when n independent trials, each resulting in a success with probability p, are performed, then, for any $a < b$,*

$$P\left\{a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right\} \to \Phi(a) - \Phi(b)$$

as $n \to \infty$.

5.3 Gamma Distribution

A random variable is said to have a gamma distribution with parameters (α, λ), $\lambda > 0$, $\alpha > 0$, if its density function is given by

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

Integration of $\Gamma(\alpha)$ by parts yields

$$\begin{aligned}\Gamma(\alpha) &= -e^{-y}y^{\alpha-1}\Big|_0^{\infty} + \int_0^{\infty} e^{-y}(\alpha-1)y^{\alpha-2}dy \\ &= (\alpha-1)\int_0^{\infty} e^{-y}y^{\alpha-2}dy \\ &= (\alpha-1)\Gamma(\alpha-1)\end{aligned} \tag{18}$$

For integral values of α, say, $\alpha \equiv n$, we obtain, by applying Equation 18 repeatedly,

$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1) \\ &= (n-1)(n-2)\Gamma(n-2) \\ &= \dots \\ &= (n-1)(n-2)\dots 3, 2\Gamma(1)\end{aligned}$$

Since $\Gamma(1) = \int_0^{\infty} e^{-x}dx = 1$, it follows that, for integral values of n,

$$\Gamma(n) = (n-1)!$$

When α is a positive integer, say, $\alpha = n$, the gamma distribution with parameters (α, λ) often arises, in practice as the distribution of the amount of time one has to wait until a total of n events has occurred. Let T_n denote the time at which the nth event occurs, and note that T_n is less than or equal to t if and only if the number of events that have occurred by time t is at least n. That is, with $N(t)$ equal to the number of events in $[0, t]$.

$$\begin{aligned} P\{T_n \leq t\} &= P\{N(t) \geq n\} \\ &= \sum_{j=n}^{\infty} P\{N(t) = j\} \\ &= \sum_{j=n}^{\infty} \frac{e^{-\lambda t}(\lambda t)^j}{j!} \end{aligned}$$

Hence, T_n has the gamma distribution with parameters (n, λ). (This distribution is often referred to in the literature as the n-Erlang distribution.) Note that when $n = 1$, this distribution reduces to the exponential distribution.

The gamma distribution with $\lambda = 1/2$ and $\alpha = n/2$, n a positive integer, is called χ^2_n (read "chi-squared") distribution with n degrees of freedom.

Example 5.7. Let X be a gamma random variable with parameters α and λ. Calculate (a) $E[X]$ and (b) $\text{Var}(X)$.

Solution (a)

$$\begin{aligned} E[X] &= \frac{1}{\Gamma(\alpha)} \int_0^{\infty} \lambda x e^{-\lambda x} (\lambda x)^{\alpha - 1} dx \\ &= \frac{1}{\lambda \Gamma(\alpha)} \int_0^{\infty} \lambda e^{-\lambda x} (\lambda x)^{\alpha} dx \\ &= \frac{\Gamma(\alpha + 1)}{\lambda \Gamma(\alpha)} \\ &= \frac{\alpha}{\lambda} \quad \text{by Equation 18} \end{aligned}$$

6 JOINTLY DISTRIBUTED RANDOM VARIABLES

Go back to Table of Contents. Please click TOC

6.1 JOint Distribution Functions

We have concerned ourselves only with probability distributions for single random variables. However, we are often interested in probability statements concerning two or more random variables. In order to deal with such probabilities, we define, for any two random variables X and Y, the joint cumulative probability distribution function of X and Y by

$$F(a,b) = P\{X \leq a, Y \leq b\} \quad -\infty < a, b < \infty$$

The distribution of X can be obtained from the joint distribution of X and Y as follows:

$$\begin{aligned} F_X(a) &= P\{X \leq a\} \\ &= P\{X \leq a, Y < \infty\} \\ &= P\left(\lim_{b\to\infty}\{X \leq a, Y \leq b\}\right) \\ &= \lim_{b\to\infty} P\{X \leq a, Y \leq b\} \\ &= \lim_{b\to\infty} F(a,b) \\ &= F(a,\infty) \end{aligned}$$

The distribution functions F_X and F_Y are sometimes referred to as the marginal distributions of X and Y.

All joint probability statements about X and Y can, in theory, be answered in terms of their joint distribution function.

Example 6.1. The joint density function of X and Y is given by

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Compute (a) $P\{X > 1, Y < 1\}$, (b) $P\{X < Y\}$, and (c) $P\{X < a\}$.

Solution (a)

$$\begin{aligned} P\{X > 1, Y < 1\} &= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y}dxdy \\ &= \int_0^1 2e^{-2y}(e^{-x}|_1^\infty)dy \\ &= e^{-1}\int_0^1 2e^{-2y}dy \\ &= e^{-1}(1-e^{-2}) \end{aligned}$$

(b)

$$\begin{aligned} P\{X < Y\} &= \iint_{(x,y):x<y} 2e^{-x}e^{-2y}dxdy \\ &= \int_0^\infty \int_0^y 2e^{-x}e^{-2y}dxdy \\ &= \int_0^\infty 2e^{-2y}(1-e^{-y})dy \\ &= \int_0^\infty 2e^{-2y}dy - \int_0^\infty 2e^{-3y}dy \\ &= 1 - \frac{2}{3} \\ &= \frac{1}{3} \end{aligned}$$

(c)

$$\begin{aligned} P\{X < a\} &= \int_0^a \int_0^\infty 2e^{-2y}e^{-x}dydx \\ &= \int_0^a e^{-x}dx \\ &= 1 - e^{-a} \end{aligned}$$

6.2 Independent Random Variables

The random variables X and Y are said to be independent if, for any two sets of real numbers A and B,

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \tag{19}$$

In other words, X and Y are independent if, for all A and B, the events $E_A = \{X \in A\}$ and $F_B = \{Y \in B\}$ are independent.

Example 6.2. A man and a women decide to meet at a certain location. If each of them independently arrives at a time uniformly distributed between 12 noon and 1 P.M., find the probability that the first to arrive has to wait longer than 10 minutes.

Solution If we let X and Y denote, respectively, the time past 12 that the man and the woman arrive, then X and Y are independent random variables, each of which is uniformly distributed over (0, 60). The desired probability $P(X+10<Y)+P(Y+10<X)$, which, by symmetry, equals $2P(X+10<Y)$, is obtained as follows:

$$\begin{aligned} 2P\{X+10<Y\} &= 2\iint_{X+10<y} f(x,y)dxdy \\ &= 2\iint_{x+10<y} f_X(x)f_Y(y)dxdy \\ &= 2\int_{10}^{60}\int_{0}^{y-10}\left(\frac{1}{60}\right)^2 dxdy \\ &= \frac{2}{(60)^2}\int_{10}^{60}(y-10)dy \\ &= \frac{25}{36} \end{aligned}$$

□

7 PROPERTIES OF EXPECTATION

Go back to Table of Contents. Please click TOC

We develop and exploit additional properties of expected values. To begin, recall that the expected value of the random variable X is defined by

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

when X is a continuous random variable with probability density function $f(x)$.

Since $E[X]$ is a weighted average of the possible values of X, it follows that if X must lie between a and b, then so must its expected value. That is, if

$$P\{a \leq X \leq b\} = 1$$

then $a \leq E[X] \leq b$.

To verify the preceding statement, suppose that X is a discrete random variable for which $P\{a \leq X \leq b\} = 1$. Since this implies that $p(x) = 0$ for all x outside of the interval $[a, b]$, it follows that

$$\begin{aligned} E[X] &= \textstyle\sum_{x:p(x)>0} xp(x) \\ &\geq \textstyle\sum_{x:p(x)>0} ap(x) \\ &= a\textstyle\sum_{x:p(x)>0} ap(x) \\ &= a\textstyle\sum_{x:p(x)>0} p(x) \\ &= a \end{aligned}$$

7.1 Expectation of Sums of Random Variables

Proposition 7.1. *If X and Y have a joint probability mass function $p(x, y)$, then*

$$E[g(X, Y)] = \sum_y \sum_x g(x, y)p(x, y)$$

If X and Y have a joint probability density function $f(x, y)$, then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy$$

Let us prove it when the random variables X and Y are jointly continuous with joint density function $f(x, y)$ and when $g(X, Y)$ is a nonnegative continuous with joint density function $f(x, y)$ and when $g(X, Y)$ is a nonnegative random variable. Because $g(X, Y) \geq 0$, we have that

$$E[g(X, Y)] = \int_0^{\infty} P\{g(X, Y) > t\}dt$$

Writing

$$P\{g(X, Y) > t\} = \iint_{(x,y):g(x,y)>t} f(x, y)dydx$$

shows that

$$E[g(X, Y)] = \int_0^{\infty} \int \int_{(x,y):g(x,y)>t} f(x, y)dydxdt$$

Interchanging the order of integration gives

$$\begin{aligned} E[g(X, Y)] &= \textstyle\int_x \int_y \int_{t=0}^{g(x,y)} f(x, y)dtdydx \\ &= \textstyle\int_x \int_y g(x, y)f(x, y)dydx \end{aligned}$$

Example 7.2. The sample mean

Let $X_1, ..., X_n$ be independent and identically distributed random variables having distribution function F and expected value μ. Such a sequence of random variables is said to constitute a sample from the distribution F. The quantity

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

is called the sample mean. Compute $E[\bar{X}]$.
Solution

$$\begin{aligned} E[\bar{X}] &= E\left[\sum_{i=1}^n \frac{X_i}{n}\right] \\ &= \frac{1}{n}E\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^n E[X_i] \\ &= \frac{1}{n}\cdot n\cdot \mu \\ &= \mu \end{aligned}$$

That is, the expected value of the sample mean is μ, the mean of the distribution. When the distribution mean μ is unknown, the sample mean is often used in statistics to estimate it.

□

Example 7.3. Boole's inquality
Let $A_1, .., A-n$ denote events, and define the indicator variables $X_i, i = 1, ..., n$, by

$$X_i = \begin{cases} 1 & \text{if } A_i \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Let

$$X = \sum_{i=1}^n X_i$$

so X denotes the number of the events A_i that occur. Finally, let

$$Y = \begin{cases} 1 & \text{if } X \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

so Y is equal to 1 if at least one of the A_i occurs and is 0 otherwise. Now, it is immediate that $X \geq Y$ so $E[X] \geq E[Y]$. But since

$$E[X] = \sum_{i=1}^n E[X_i] = \sum -i = 1^n P(A_i)$$

and

$$E[Y] = P\{\text{at least one of the } A_i \text{ occur}\} = P\left(\bigcup_{i=1}^n A_i\right)$$

we obtain Boole's inequality, namely,

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$$

□

7.2 Covariance, Variance of Sums, and Correlations

Proposition 7.4. *If X and Y are independent, then, for any functions h and g,*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Proof Suppose that X and Y are jointly continuous with joint density $f(x, y)$. Then

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx \\ &= E[h(Y)]E[g(X)] \end{aligned}$$

□

Definition 7.5. The covariance between X and Y, denoted by $\text{Cov}(X, Y)$, is defined by

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Upon expanding the right side of the preceding definition, we see that

$$\begin{aligned}\text{Cov}(X,Y) &= E[XY - E[X]Y - XE[Y] + E[Y]E[X]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

Proposition 7.6. *(i)* $Cov(X,Y) = Cov(Y,X)$

(ii) $Cov(X,X) = Cov(X)$

(iii) $Cov(aX,Y) = aCov(X,Y)$

Proof Let $\mu_i = E[X_i]$ and $v_j = E[Y_i]$. Then

$$E\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mu_i,\ E\Big[\sum_{j=1}^{m} Y_j\Big] = \sum_{j=1}^{m} v_j$$

and

$$\begin{aligned}\text{Cov}\Big(\textstyle\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\Big) &= E\Big[\Big(\textstyle\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\Big)\Big(\sum_{j=1}^{m} Y_j - \sum_{j=1}^{m} v_j\Big)\Big] \\ &= E\Big[\textstyle\sum_{i=1}^{n}(X_i - \mu_i)\sum_{j=1}^{m}(Y_j - v_j)\Big] \\ &= E\Big[\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m}(X_i - \mu_i)(Y_j - v_j)\Big] \\ &= \textstyle\sum_{i=1}^{n}\sum_{j=1}^{m} E[(X_j - \mu_i)(Y_j - v_j)]\end{aligned}$$

where the last equality follows because the expected value of a sum of random variables is equal to the sum of the expected values.

□

It follows that, upon taking $Y_j = X_j$, $j = 1, ..., n$, that

$$\begin{aligned}\text{Var}\Big(\textstyle\sum_{i=1}^{n} X_i\Big) &= \text{Cov}\Big(\textstyle\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\Big) \\ &= \textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\ &= \textstyle\sum_{i=1}^{n} \text{Var}(X_i) + \sum\sum_{i\neq j} \text{Cov}(X_i, X_j)\end{aligned}$$

Since each pair of indices i, j, $i \neq j$, appears twice in the double summation, the preceding formula is equivalent to

$$\text{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum\sum_{i<j} \text{Cov}(X_i, X_j) \tag{20}$$

If $X_1, ..., X_n$ are pairwise independent, in that X_i and X_j are independent for $i \neq j$, then Equation 20 reduces to

$$\text{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \text{Var}(X_i)$$

Example 7.7. Let $X_1, ..., X_n$ be independent and identically distributed random variables having expected value μ and variance σ^2, and let $\bar{X} = \sum_{i=1}^{n} X_i/n$ be the sample mean. The quantities $X_i - \bar{X}$, $i = 1, ..., n$, are called deviations, as they equal the differences between the individual data and the sample mean. The random variable

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

is called the sample variance. Find (a) $\text{Var}(\bar{X})$ and (b) $E[S^2]$.

Solution (a)

$$\begin{aligned}\text{Var}(\bar{X}) &= \Big(\tfrac{1}{n}\Big)^2 \text{Var}\Big(\textstyle\sum_{i=1}^{n} X_i\Big) \\ &= \Big(\tfrac{1}{n}\Big)^2 \textstyle\sum_{i=1}^{n} \text{Var}(X_i) \ \text{ by independence} \\ &= \tfrac{\sigma^2}{n}\end{aligned}$$

(b)

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^n (X_i-\mu+\mu-\bar{X})^2 \\ &= \sum_{i=1}^n (X_i-\mu)^2 + \sum_{i=1}^n (\bar{X}-\mu)^2 - 2(\bar{X}-\mu)\sum_{i=1}^n (X_i-\mu) \\ &= \sum_{i=1}^n (X_i-\mu)^2 + n(\bar{X}-\mu)^2 - 2(\bar{X}-\mu)n(\bar{X}-\mu) \\ &= \sum_{i=1}^n (X_i-\mu)^2 - n(\bar{X}-\mu)^2\end{aligned}$$

Taking expectations of the proceeding

$$\begin{aligned}(n-1)E[S^2] &= \sum_{i=1}^n E[(X_i-\mu)^2] - nE[(\bar{X}-\mu)^2] \\ &= n\sigma^2 - n\text{Var}(\bar{X}) \\ &= (n-1)\sigma^2\end{aligned}$$

Given $\mu_x = E[X]$, $\sigma_x^2 = \text{Var}(X)$, and $\mu_y = E[Y]$, $\sigma_y^2 = \text{Var}(Y)$. Consequently,

$$\begin{aligned}Corr(X,Y) &= \frac{\text{Cov}(X,Y)}{\sigma_x\sigma_y} \\ &= \frac{E[XY]-\mu_x\mu_y}{\sigma_x\sigma_y}\end{aligned}$$

To determine $E[XY]$, we condition on Y. That is, we use the identity

$$E[XY] = E[E[XY|Y]]$$

Consequently,

$$E[XY|Y] = Y\mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y^2 - \mu_y Y)$$

implying that

$$\begin{aligned}E[XY] &= E\left[Y\mu_x + \rho\tfrac{\sigma_x}{\sigma_y}(Y^2-\mu_y Y)\right] \\ &= \mu_x\mu_y + \rho\tfrac{\sigma_x}{\sigma_y}E[Y_2 - \mu_y Y] \\ &= \mu_x\mu_y + \rho\sigma_x\sigma_y\end{aligned}$$

Therefore,

$$\text{Corr}(X,Y) = \frac{\rho\sigma_x\sigma_y}{\sigma_x\sigma_y} = \rho$$

□

7.3 Moment Generating Functions

The moment generating function $M(t)$ of the random variable X is defined for all real values of g by

$$\begin{aligned}M(t) &= E[e^{tX}] \\ &= \begin{cases} \sum_x e^{tx}p(x) & \text{if } X \text{ is discrete with mass function } p(x) \\ \int_{-\infty}^{\infty} e^{tx}f(x)dx & \text{if } X \text{ is continuous with density } f(x)\end{cases}\end{aligned}$$

We call $M(t)$ the moment generating function because all of the moments of X can be obtained by successively differentiating $M(t)$ and then evaluating the result at $t = 0$. For example,

$$\begin{aligned}M'(t) &= \tfrac{d}{dt}E[e^{tx}] \\ &= E[\tfrac{d}{dt}(e^{tX})] \\ &= E[Xe^{tX}]\end{aligned} \tag{21}$$

Similarly,

$$\begin{aligned}M''(t) &= \tfrac{d}{dt}M'(t) \\ &= \tfrac{d}{dt}E[Xe^{tX}] \\ &= E[\tfrac{d}{dt}(Xe^{tX}] \\ &= E[X^2e^{tX}]\end{aligned}$$

Thus,

$$M''(0) = E[X^2]$$

Example 7.8. Binomial distribution with parameters n and p. If X is a binomial random variable with parameters n and p, then

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= (pe^t + 1 - p)^n \end{aligned}$$

where the last equality follows from the binomial theorem. Differentiation yields

$$M'(t) = n(pe^t + 1 - p)^{n-1} pe^t$$

Thus,

$$E[X] = M'(0) = np$$

Differentiating a second time yields

$$M''(t) = n(n-1)(pe^t + 1 - p)^{n-2}(pe^t)^2 + n(pe^t + 1 - p)^{n-1} pe^t$$

so

$$E[X^2] = M''(0) = n(n-1)p^2 + np$$

The variance of X is given by

$$\begin{aligned} \text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= np(1-p) \end{aligned}$$

□

Example 7.9. Poisson distribution with mean λ If X is a Poisson random variable with parameter λ, then

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \sum_{n=0}^{\infty} \frac{e^{tn} e^{-\lambda} \lambda^n}{n!} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\ &= e^{-\lambda} e^{\lambda e^t} \end{aligned}$$

Differentiating yields

$$\begin{aligned} M'(t) &= \lambda e^t \exp(\lambda(e^t - 1)) \\ M''(t) &= (\lambda e^t)^2 \exp(\lambda(e^t - 1)) + \lambda e^t \exp(\lambda(e^t - 1)) \end{aligned}$$

Thus

$$\begin{aligned} E[X] &= M'(0) = \lambda \\ E[X^2] &= M''(0) = \lambda^2 + \lambda \\ \text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= \lambda \end{aligned}$$

Example 7.10. Exponential distribution with parameter λ

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda} dx \\ &= \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx \\ &= \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda \end{aligned}$$

Differentiation of $M(t)$ yields

$$M'(t) = \frac{\lambda}{(\lambda - t)^2} \; M''(t) = \frac{2\lambda}{(\lambda - t)^3}$$

Hence,

$$E[X] = M'(0) = \frac{1}{\lambda}, \; E[X^2] = M''(0) = \frac{2}{\lambda^2}$$

The variance of X is given by

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{1}{\lambda^2}$$

□

Example 7.11. Normal distribution

We first compute the moment generating function of a unit normal random variable with parameters 0 and 1. Letting Z be such a random variable, we have

$$\begin{aligned} M_Z(t) &= E[e^{tZ}] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x^2-2tx)}{2} \right\} dx \\ &= e^{t^2/2} \end{aligned}$$

Hence, the moment generating function is

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= E[e^{t(\mu+\sigma Z}] \\ &= e^{t\mu} E[e^{t\sigma Z}] \\ &= \exp\left\{ \frac{\sigma^2 t^2}{2} + \mu t \right\} \end{aligned}$$

By differentiating, we have

$$M_X'(t) = (\mu + t\sigma^2) \exp\left\{ \frac{\sigma^2 t^2}{2} + \mu t \right\}$$

$$M_X''(t) = (\mu + t\sigma^2)^2 \exp\left\{ \frac{\sigma^2 t^2}{2} + \mu t \right\} + \sigma^2 \exp\left\{ \frac{\sigma^2 t^2}{2} + \mu t \right\}$$

Thus,

$$E[X] = M'(0) = \mu$$

$$E[X^2] = M''(0) = \mu^2 + \sigma^2$$

implying that

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \sigma^2$$

□

7.4 Multivariate Normal Distribution

Let $Z_1, ..., Z_n$ be a set of n independent unit normal random variables. If, for some constants a_{ij}, $1 \leq j \leq m$, $1 \leq j \leq n$, and μ_i, $1 \leq i \leq m$,

$$X_1 = a_{11} Z_1 + \cdots + a_{1n} Z_n + \mu_1$$

$$X_2 = a_{21} Z_1 + \cdots + a_{2n} Z_n + \mu_2$$

$$...$$

$$X_i = a_{i1} Z_1 + \cdots + a_{in} Z_n + \mu_i$$

$$X_m = a_{m1} Z_1 + \cdots + a_{mn} Z_n + \mu_m$$

then the random variables $X_1, ..., X_m$ are said to have a multivariate normal distribution.

From the fact that the sum of independent normal random variables is a normal random variable, it follows that each X_i is a normal random variable with mean and variance given, respectively, by

$$E[X_i] = \mu_i$$

$$\text{Var}(X_i) = \sum_{j=1}^{n} a_{ij}^2$$

Let us now consider

$$M(t_1, ..., t_m) = E[\exp(t_1 X_1 + \cdots + t_m X_m)]$$

the joint moment generating function of $X_1, ..., X_m$. The first thing to ntoe is that since $\sum_{i=1}^{m} t_i X_i$ is itself a linear combination of the independent normal random variables $Z_1, ..., Z_n$, it is also normally distributed. Its mean and variance are

$$E\left[\sum_{i=1}^{m} t_i X_i\right] = \sum_{i=1}^{m} t_i \mu_i$$

and

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^{m} t_i X_i\right) &= \text{Cov}\left(\sum_{i=1}^{m} t_i X_i, \sum_{j=1}^{m} t_j X_j\right) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} t_i t_j \text{Cov}(X_i, X_j) \end{aligned}$$

Now, if Y is a normal random variable with mean μ and variance σ^2, then

$$E[e^Y] = M_Y(t)|_{t=1} = e^{\mu + \sigma^2/2}$$

Thus,

$$M_{(t_1,...,t_m)} = \exp\left\{\sum_{i=1}^{m} t_i \mu_i + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} t_i t_j \text{Cov}(X_i, X_j)\right\}$$

which shows that the joint distribution of $X_1, ..., X_m$ is completely determined from a knowledge of the values of $E[X_i]$ and $\text{Cov}(X_i, X_j)$, $i, j = 1, 2, ..., m$.

References

[1] Ross, Sheldon M., *Introduction to Probability and Statistics for Engineers and Scientists*, 9*th* Edition.