Notes for A First Course in Probability

CHAPTER 1: COMBINATORIAL ANALYSIS

Inclusion–exclusion identity.

$$E[aX+b] = aE[X] + b.$$
$$\mathrm{Var}(X) = E[X^2] - (E[X])^2.$$

Bernoulli random variable $(1,p)$ → Binomial random variable (n,p)

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

Poisson random variable (λ)

Geometric random variable $(1,p)$ → Negative binomial random variable (r,p)

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]. \quad (S)$$

Exponential random variable λ → Gamma distribution (α,λ)

chi-squared distribution.

t-distribution.

$$E[X+Y] = E[X] + E[Y].$$

Sample mean:

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

If X and Y are independent, then, for any functions h and g,

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$
$$\mathrm{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])].$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

Correlation:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$
$$E[X] = E\big[E[X|Y]\big]$$

1.1. Introduction

In fact, many problems in probability theory can be solved simply by counting the number of different ways that a certain event can occur. The mathematical theory of counting is formally known as *combinatorial analysis*.

1.2. The basic principle of counting

The basic principle of **counting** will be fundamental to all our work. Loosely put, it 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**
> 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.

$$
\begin{array}{llll}
(1,1), & (1,2), & \ldots, & (1,n) \\
(2,1), & (2,2), & \ldots, & (2,n) \\
\vdots & & & \\
(m,1), & (m,2), & \ldots, & (m,n)
\end{array}
$$

Figure 1. 1 Proof of the Basic Principle

Example 2a: A small community consists of 10 women, each of 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 the basic principle that there are $10 \times 3 = 30$ possible choices.

> **The generalized basic principle of counting**
> 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, there are n_3 possible outcomes of the third experiment; and if ⋯, then there is a total of $n_1 \cdot n_2 \cdots n_r$ possible outcomes of the r experiments.

1.3. Permutation

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. This result could also have been obtained from the basic principle, since the first object in the permutation can be any of the 3, the second object in the permutation can then be chosen from any of the remaining 2, and the third object in the permutation is then the remaining 1. Thus, there are $3 \cdot 2 \cdot 1 = 6$ possible permutations. (has the order of selection)

> Suppose now that we have n objects. Reasoning similar to that we have just used for the 3 letters then shows that there are
> $$n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1 = n!$$
> different permutations of the n objects.

Whereas $n!$ (read as "n factorial") is defined to equal $1 \cdot 2 \ldots n$ when n is a positive integer, it is convenient to define $0!$ to equal 1.

Example 3c: Ms. Jones has 10 books that she is going to put on her bookshelf. Of these, 4 are mathematics books, 3 are chemistry books, 2 are history books, and 1 is a language book. Ms. Jones wants to arrange her books so that all the books dealing with the same subject are together on the shelf. How many different arrangements are possible?

Solution There are $4!\,3!\,2!\,1!$ arrangements such that the mathematics books are first in line, then the chemistry books, then the history books, and then the language book. Similarly, for

each possible ordering of the subjects, there are $4!\,3!\,2!\,1!$ Possible arrangements. Hence, as there are $4!$ possible orderings of the subjects, the desired answer is $4!\,4!\,3!\,2!\,1! = 6912$.

We shall now determine the number of permutations of a set of n objects when certain of the objects are indistinguishable from one another. To set this situation straight in our minds, consider the following example.

Example 3d: How many different letter arrangements can be formed from the letters $PEPPER$?

Solution We first note that there are $6!$ permutations of the letters $P_1E_1P_2P_3E_2R$ when the $3P$'s and the $2E$'s are distinguished from one another. However, consider any one of these permutations—for instance, $P_1P_2E_1P_3E_2R$. If we now permute the P's among themselves and the E's among themselves, then the resultant arrangement would still be of the form $PPEPER$. That is, all $3!\,2!$ Permutations

$$\begin{array}{ll} P_1P_2E_1P_3E_2R & P_1P_2E_2P_3E_1R \\ P_1P_3E_1P_2E_2R & P_1P_3E_2P_2E_1R \\ P_2P_1E_1P_3E_2R & P_2P_1E_2P_3E_1R \\ P_2P_3E_1P_1E_2R & P_2P_3E_2P_1E_1R \\ P_3P_1E_1P_2E_2R & P_3P_1E_2P_2E_1R \\ P_3P_2E_1P_1E_2R & P_3P_2E_2P_1E_1R \end{array}$$

are of the form $PPEPER$. Hence, there are $6!/(3!\,2!) = 60$ possible letter arrangements from the letters $PEPPER$.

> In general, the same reasoning as that used in Example 3d (Page 17) shows that there are
>
> $$\frac{n!}{n_1!\,n_2!\cdots n_r!}$$
>
> different permutations of n objects, of which n_1 are alike, n_2 are alike, $\cdots$, n_r are alike.

1.4. Combination

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

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

In general, as $n(n-1)\cdots(n-r+1)$ represents the number of different ways that a group of r items could be selected from n items when the order of selection is relevant, and as each group of r items will be counted $r!$ times in this count, it follows that the number of different groups of r items that could be formed from a set of n items is

$$\frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{(n-r)!\,r!}$$

Notation and terminology

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 is consistent with the preceding interpretation because in a set of size n there is exactly 1 subset of size n (namely, the entire set), and exactly one subset of size 0 (namely the empty set). A useful convention is to define $\binom{n}{r}$ equal to 0 when either $r > n$ or $r < 0$.

Example 4c: Consider a set of n antennas of which m are defective and $n-m$ are functional and assume that all of the defectives and all of the functionals are considered indistinguishable. How many linear orderings are there in which no two defectives are consecutive?

Solution Imagine that the $n-m$ functional antennas are lined up among themselves. Now, if no two defectives are to be consecutive, then the spaces between the functional antennas must each contain at most one defective antenna. That is, in the $n-m+1$ possible positions—represented in Figure 1.2 by carets—between the $n-m$ functional antennas, we must select m of these in which to put the defective antennas. Hence, there are $\binom{n-m+1}{m}$ possible orderings in which there is at least one functional antenna between any two defective ones.

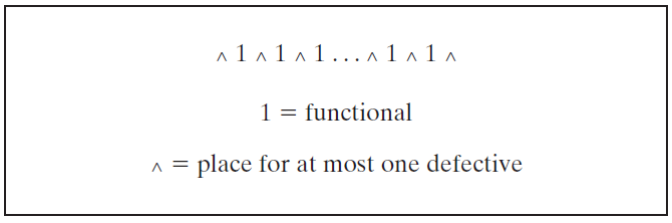


Figure 1. 2 No consecutive defectives.

A useful combinatorial identity (Page 7)

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

The values $\binom{n}{r}$ are often referred to as *binomial coefficients* because of their prominence in the binomial theorem.

The binomial theorem

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

like $(x+y)^3 = y^3 + 3xy^2 + 3x^2y + x^3$

1.5. Multinomial Coefficients

In this section, we consider the following problem: A set of n distinct items is to be **divided** into r distinct groups of respective sizes $n_1, n_2, \cdots, n_r$, where $\sum_{i=1}^{r} n_i = n$. How many different divisions are possible? To answer this question, 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 $\binom{n-n_1-n_2}{n_3}$ possible choices for the third group; and so on. It then follows from the generalized version of the basic counting principle that there are

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-n_2-\cdots-n_{r-1}}{n_r} = \frac{n!}{n_1!\,n_2!\cdots n_r!}$$

possible divisions.

Another way to see this result is to consider the n values $1,1,\cdots,1,2,\cdots,2,\cdots,r,\cdots,r$, where i appears n_i times, for $i = 1,\cdots,r$. Every permutation of these values corresponds to a division of the n items into the r groups. (Page 9)

Notation

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

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

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

The multinomial theorem

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

That is, the sum is over all nonnegative integer-valued vectors $(n_1, n_2, \cdots, n_r)$ such that $n_1 + n_2 + \cdots + n_r = n$.

The numbers $\binom{n}{n_1, n_2, \cdots, n_r}$ are known as *multinomial coefficients*.

1.6. The Number of Integer Solutions of Equations

More generally, if we supposed there were r types of fish and that a total of n were caught then the number of possible outcomes would be the number of nonnegative integer-valued vectors $x_1, \cdots, x_r$ such that

$$x_1 + x_2 + \cdots + x_r = n$$

$$0 \wedge 0 \wedge 0 \wedge \cdots \wedge 0 \wedge 0$$

n objects 0

Choose $r-1$ of the spaces $\wedge$.

Figure 1.1 Number of **positive** solutions

Note that any selection of $r-1$ of the $n-1$ spaces between adjacent zeroes corresponds to a positive solution of $x_1 + x_2 + \cdots + x_r = n$.

Proposition 6.1: There are $\binom{n-1}{r-1}$ distinct positive integer-valued vectors $(x_1, x_2, \cdots, x_r)$ satisfying the equation

$$x_1 + x_2 + \cdots + x_r = n \quad x_i > 0, i = 1, \cdots, r$$

To obtain the number of **nonnegative** (as opposed to positive) solutions, note that the number of nonnegative solutions of $x_1 + x_2 + \cdots + x_r = n$ is the same as the number of positive solutions of $y_1 + \cdots + y_r = n + r$ (seen by letting $y_i = x_i + 1, i = 1, \cdots, r$).

Proposition 6.2: There are $\binom{n+r-1}{r-1}$ distinct nonnegative integer-valued vectors $(x_1, x_2, \cdots, x_r)$ satisfying the equation

$$x_1 + x_2 + \cdots + x_r = n$$

Theoretical Exercises

2. Two experiments are to be performed. The first can result in any one of m possible outcomes. If the first experiment results in outcome i, then the second experiment can result in any of n_i possible outcomes, $i = 1,2, \cdots, m$. What is the number of possible outcomes of the two experiments?

$$\sum_{i=1}^{m} n_i.$$

3. In how many ways can r objects be selected from a set of n objects if the order of selection is considered relevant?

$$n(n-1)\cdots(n-r+1) = n!/(n-r)!$$

7. Give an analytic proof of equation $\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}$.

$$\binom{n-1}{r} + \binom{n-1}{r-1} = \frac{(n-1)!}{r!\,(n-1-r)!} + \frac{(n-1)!}{(n-r)!\,(r-1)!}$$

$$= \frac{n!}{r!\,(n-r)!}\left[\frac{n-r}{n} + \frac{r}{n}\right] = \binom{n}{r}$$

8. Prove that

$$\binom{n+m}{r} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0}$$

Hint: Consider a group of n men and m women. How many groups of size r are possible?

There are $\binom{n+m}{r}$ groups of size r. As there are $\binom{n}{i}\binom{m}{r-i}$ groups of size r that consist of i men and $r-i$ women, we see that

$$\binom{n+m}{r} = \sum_{i=0}^{r} \binom{n}{i}\binom{m}{r-i}.$$

11. The following identity is known as Fermat's combinatorial identity:

$$\binom{n}{k} = \sum_{i=k}^{n} \binom{i-1}{k-1} \quad n \geq k$$

Give a combinatorial argument (no computations are needed) to establish this identity. *Hint*: Consider the set of numbers 1 through n. How many subsets of size k have i as their highest numbered member?

The number of subsets of size k that have i as their highest numbered member is equal to $\binom{i-1}{k-1}$, the number of ways of choosing $k-1$ of the numbers $1, \cdots, i-1$. Summing over i yields the number of subsets of size k.

13. Show that, for $n > 0$,

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

Hint: Use the binomial theorem.

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

17. Present a combinatorial explanation of why $\binom{n}{r} = \binom{n}{r, n-r}$.

A choice of r elements from a set of n elements is equivalent to breaking these elements into two subsets, one of size r (equal to the elements selected) and the other of size $n-r$ (equal to the elements not selected).

CHAPTER 2: AXIOMS OF PROBABILITY

2.2. 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 by S**.**

Following are some examples: (Page 21)

$$\mathrm{S} = \{\text{girl}, \text{boy}\},$$
$$S = \{\text{all } 7! \text{ permutations of } (1,2,3,4,5,6,7)\},$$
$$S = \{(H,H),(H,T),(T,H),(T,T)\}$$

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.

For any two events E and F of a sample space S, we define:

- $E \cup F$ is called the ***union*** of the event E and the event F, which consists of all outcomes that are either in E or in F or in both E and F.

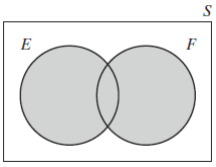


Figure 2. 1 Venn diagram: $E \cup F$

- $E \cap F$ or EF is called the ***intersection*** of E and F, which consists of all outcomes that are both in E and in F.

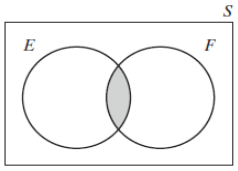


Figure 2. 2 Venn diagram: $E \cap F$ *or* EF

- $\emptyset$ is the null event which refers to the event consisting of no outcome. If $EF = \emptyset$, then E and F are said to be ***mutually exclusive***.
- $\bigcup_{n=1}^{\infty} E_n$ is the union of events-$E_1, E_1, \cdots$ of $n = 1,2,\cdots$.
- $\bigcap_{n=1}^{\infty} E_n$ is the event consisting of those outcomes that are in all of the events $E_n, n = 1,2,\cdots$
- E^c, referred to as the complement of E, to consist of all outcomes in the sample space S that are not in E. ($S^c = \emptyset$)

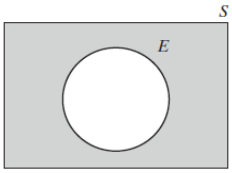


Figure 2. 3 Venn diagram: E^c

- $E \subset F, F \supset E$: we say that E is contained in F, or E is a subset of F, which we sometimes say as F is a superset of E.

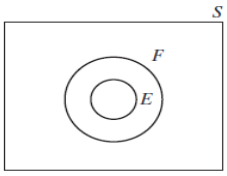
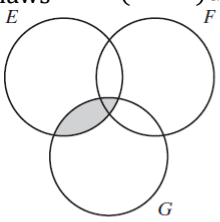


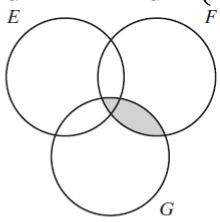
Figure 2. 4 Venn diagram: $E \subset F, F \supset E$

- $E = F$: E and F are equal ($E \subset F$ and $F \subset E$).
- Rules:

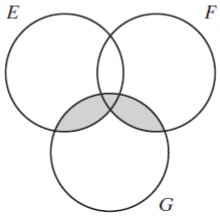
Commutative laws	$E \cup F = F \cup E$	$EF = FE$
Associative laws	$(E \cup F) \cup G = E \cup (F \cup G)$	$(EF)G = E(FG)$
Distributive laws	$(E \cup F)G = EG \cup FG$	$EF \cup G = (E \cup G)(F \cup G)$



(a) Shaded region: EG.



(b) Shaded region: FG.



(c) Shaded region: $(E \cup F)G$.

Figure 2. 5 $(E \cup F)G = EG \cup FG$.

- *DeMorgan's* laws:

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

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

2.3. 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 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 the limiting relative frequency of E.

Although the preceding definition is certainly intuitively pleasing and should always be kept in mind by the reader, it possesses a serious drawback: How do we know that $\mathrm{n}(E)/n$ will converge to some constant limiting value that will be the same for each possible sequence of repetitions of the experiment?

Consider an experiment whose sample space is S. For each event E of the sample space S, we assume that a number $P(E)$ is defined and satisfies the following three axioms:

The three axioms of probability

Axiom 1

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

Axiom 2

$$P(S) = 1$$

Axiom 3

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

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

We refer to $P(E)$ as the ***probability*** of the event E.

The null event has probability 0 of occurring as $P(\emptyset) = 0$.

Technical Remark. 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.

2.4. Some simple propositions

Proposition 4.1: $P(E^c) = P(S) - P(E) = 1 - P(E)$

Proposition 4.2: If $E \subset F$, then $P(E) \le P(F)$.

Proposition 4.3: $P(E \cup F) = P(E) + P(F) - P(EF)$

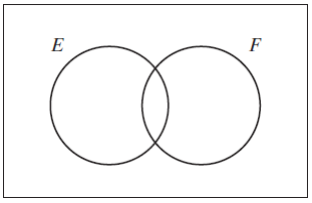


Figure 2. 6 Venn diagram.

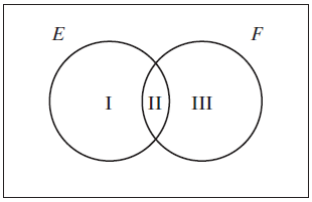


Figure 2. 7 Venn diagram in sections.

We may also calculate the probability that any one of the three events E, F, and G occurs, namely,

$$P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(EF) - P(EG) - P(FG) + P(EFG)$$

In fact, the following proposition, known as the ***inclusion–exclusion identity***, can be proved by **mathematical induction**:

Proposition 4.4:

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i) - \sum_{i_1 < i_2} P\left(E_{i_1} E_{i_2}\right) +$$

$$+(-1)^{r+1} \sum_{i_1 < i_2 < \cdots < i_r} P\left(E_{i_1} E_{i_2} \cdots E_{i_r}\right)$$

$$+ \cdots + (-1)^{n+1} P(E_1 E_2 \cdots E_n)$$

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

In words, Proposition 4.4 states that the probability of the union of n events equals the sum of the probabilities of these events taken one at a time, minus the sum of the probabilities of these events taken two at a time, plus the sum of the probabilities of these events taken three at a time, and so on.

2.5. Sample Spaces Having Equally Likely Outcomes

In many experiments, it is natural to assume that all outcomes in the sample space are **equally** likely to occur. That is, consider an experiment whose sample space S is a finite set, say, $S = \{1,2,\cdots,N\}$. Then, it is often natural to assume that

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{N\})$$

which implies, from Axioms 2 and 3, that

$$P(\{i\}) = \frac{1}{N} \quad i = 1,2,\cdots,N$$

From this equation, it follows from Axiom 3 that, for any event E,

$$P(E) = \frac{\text{number of outcomes in } E}{\text{number of outcomes in } S}$$

In words, if we assume that all outcomes of an experiment are equally likely to occur, then the probability of any event E equals the proportion of outcomes in the sample space that are contained in E.

Example 5b: If 3 balls are "randomly drawn" from a bowl containing 6 white and 5 black balls, what is the probability that one of the balls is white and the other two black?

$$\frac{\binom{6}{1}\binom{5}{2}}{\binom{11}{3}} = \frac{4}{11}$$

When the experiment consists of a random selection of k items from a set of n items, we have the **flexibility** of either letting the outcome of the experiment be the ordered selection of the k items or letting it be the unordered set of items selected. In the former case, we would assume that each new selection is equally likely to be any of the so far unselected items of the set, and in the latter case, we would assume that all $\binom{n}{k}$ possible subsets of k items are equally likely to be the set selected.

Example 5n: Compute the probability that if 10 married couples are seated at random at a round table, then no wife sits next to her husband. (Page 40)

2.6. Probability as a continuous set function

A sequence of events $\{E_n, n \geq 1\}$ is said to be an increasing sequence if

$$E_1 \subset E_2 \subset \cdots \subset E_n \subset E_{n+1} \subset \cdots$$

whereas it is said to be a decreasing sequence if

$$E_1 \supset E_2 \supset \cdots \supset E_n \supset E_{n+1} \supset \cdots$$

If $\{E_n, n \geq 1\}$ is an increasing sequence of events, then we define a new event, denoted by $\lim_{n\to\infty} E_n$, by

$$\lim_{n\to\infty} E_n = \bigcup_{i=1}^{\infty} E_i$$

Similarly, if $\{E_n, n \geq 1\}$ is a decreasing sequence of events, we define $\lim_{n\to\infty} E_n$, by

$$\lim_{n\to\infty} E_n = \bigcap_{i=1}^{\infty} E_i$$

Proposition 6.1: If $\{E_n, n \geq 1\}$ is either an increasing or a decreasing sequence of events, then

$$\lim_{n\to\infty} P(E_n) = P\left(\lim_{n\to\infty} E_n\right)$$

Example 6a: Probability and a "paradox" (Page 44)

2.7. Probability as a Measure of Belief

The most simple and natural interpretation is that the probabilities referred to are measures of the individual's degree of belief in the statements that he or she is making.

$P(A)$ can be interpreted either as a long-run relative frequency or as a measure of one's degree of belief.

Theoretical Exercises

6. Let E, F, and G be three events. Find expressions for the events so that, of E, F, and G,

(a) only E occurs;

$$EF^cG^c$$

(b) both E and G, but not F, occur;

$$EF^cG$$

(c) at least one of the events occurs;

$$E \cup F \cup G$$

(d) at least two of the events occur;

$$EF \cup EG \cup FG$$

(e) all three events occur;

$$EFG$$

(f) none of the events occurs;

$$E^cF^cG^c$$

(g) at most one of the events occurs;

$$E^cF^cG^c \cup EF^cG^c \cup E^cFG^c \cup E^cF^cG$$

(h) at most two of the events occur;

$$(EFG)^c$$

(i) exactly two of the events occur;

$$EFG^c \cup EF^cG \cup E^cFG$$

(j) at most three of the events occur.

$$S$$

11. If $P(E) = .9$ and $P(F) = .8$, show that $P(EF) \geq .7$. In general, prove *Bonferroni's inequality*, namely,

$$P(EF) \geq P(E) + P(F) - 1.$$

$$1 \geq P(E \cup F) = P(E) + P(F) - P(EF)$$

12. Show that the probability that exactly one of the events E or F occurs equals $P(E) + P(F) - 2P(EF)$.

$$P(EF^c \cup E^cF) = P(EF^c) + P(E^cF)$$

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

CHAPTER 3: CONDITIONAL PROBABILITY AND INDEPENDENCE

3.1 Introduction

The importance of conditional probability is twofold. In the first place, we are often interested in calculating probabilities when some partial information concerning the result of an experiment is available; in such a situation, the desired probabilities are conditional. Second, even when no partial information is available, conditional probabilities can often be used to compute the desired probabilities **more easily**.

3.2 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 $\frac{1}{36}$. Suppose further 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? To calculate this probability, we reason as follows: Given that the initial die is a 3, there can be at most 6 possible outcomes of our experiment, namely, $(3,1)$, $(3,2)$, $(3,3)$, $(3,4)$, $(3,5)$, and $(3,6)$. Since each of these outcomes originally had the same probability of occurring, the outcomes should still have equal probabilities. That is, given that the first die is a 3, the (conditional) probability of each of the outcomes $(3,1)$, $(3,2)$, $(3,3)$, $(3,4)$, $(3,5)$, and $(3,6)$ is $\frac{1}{6}$, whereas the (conditional) probability of the other 30 points in the sample space is 0. Hence, the desired probability will be $\frac{1}{6}$.

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

If $P(F) > 0$, then

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

Multiplying both sides of the Equation above by $P(F)$, we obtain

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

In words, it states that the probability that both E and F occur is equal to the probability that F occurs multiplied by the conditional probability of E given that F occurred. It is often quite useful in computing the probability of the intersection of events.

A generalization of the Equation above, which provides an expression for the probability of the intersection of an arbitrary number of events, is sometimes referred to as the *multiplication rule*.

The multiplication rule

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

3.3 Bayes's formula

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

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

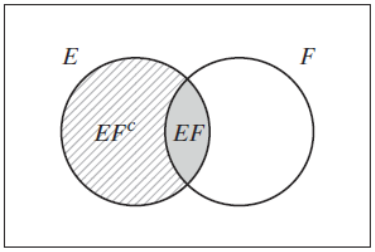


Figure 3. 1 $E = EF \cup EF^c$

As EF and EF^c are clearly mutually exclusive, we have, by Axiom 3,

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

which can be used to compute $P(E)$ by "conditioning" on whether F occurs. This is an extremely useful formula, because its use often enables us to determine the probability of an event by first "conditioning" upon whether or not some second event has occurred.

The change in the probability of a hypothesis when new evidence is introduced can be expressed compactly in terms of the change in the ***odds*** of that hypothesis, where the concept of odds is defined as follows. (Page 68)

Definition

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 than it is that it does not occur. For instance, if $P(A) = \frac{2}{3}$, 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

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

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.

Equation $P(E) = P(EF) + P(EF^c)$ may be generalized as follows: Suppose that $F_1, F_2, \cdots, F_N$ are mutually exclusive events such that

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

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

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

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

The law of total probability

$$P(E) = \sum_{i=1}^{n} P(EF_i)$$

$$= \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

Thus, it, often referred to as the ***law of total probability***, shows how, for given events $F_1, F_2, \cdots, F_N$, of which one and only one must occur, we can compute $P(E)$ by first conditioning on which one of the F_i occurs. That is, Equation above states that $P(E)$ is equal to a weighted average of $P(E|F_i)$, each term being weighted by the probability of the event on which it is conditioned.

Suppose now that E has occurred and we are interested in determining which one of the F_j **also** occurred. Then, by the law of total probability, we have the following proposition.

Proposition 3.1:

$$P(F_j|E) = \frac{P(EF_j)}{P(E)}$$

$$= \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$$

Equation above is known as ***Bayes's formula***, after the English philosopher Thomas Bayes. If we think of the events F_j as being possible "hypotheses" about some subject matter, then Bayes's formula may be interpreted as showing us how opinions about these hypotheses held before the experiment was carried out [that is, the $P(F_j)$] should be modified by the evidence produced by the experiment.

3.4 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 the **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

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

Definition

Two events E and F are said to be ***independent*** if the Equation $P(EF) = P(E)P(F)$ holds. Two events E and F that are not independent are said to be ***dependent***.

Example 4b: Two coins are flipped, and all 4 outcomes are assumed to be equally likely. If E is the event that the first coin lands on heads and F the event that the second lands on tails, then E and F are independent, since $P(EF) = P(\{(H,T)\}) = \frac{1}{4}$, whereas $P(E) = P(\{(H,H),(H,T)\}) = \frac{1}{2}$ and $P(F) = P(\{(H,T),(T,T)\}) = \frac{1}{2}$.

Proposition 4.1: If E and F are independent, then so are E and F^c. ($E = EF \cup EF^C$)

Thus, if E is independent of F, then the probability of E's occurrence is unchanged by information as to whether or not F has occurred.

Definition

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

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

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

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

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

Sometimes, a probability experiment under consideration consists of performing a sequence of subexperiments. For instance, if the experiment consists of continually tossing a coin, we may think of each toss as being a subexperiment. In many cases, it is reasonable to assume that the outcomes of any group of the subexperiments have no effect on the probabilities of the outcomes of the other subexperiments. If such is the case, we say that the subexperiments are independent. More formally, we say that the subexperiments are independent if $E_1, E_2, \cdots, E_n, \cdots$ is necessarily an independent sequence of events whenever E_i is an event whose occurrence is completely determined by the outcome of the ith subexperiment.

If each subexperiment has the same set of possible outcomes, then the subexperiments are often called *trials*.

Example 4j: The problem of the points (Page 81)

Fermat argued that in order for n successes to occur before m failures, it is necessary and sufficient that there be at least n successes in the first $m + n - 1$ trials.

Example 4k: Service protocol in a serve and rally game (Page 82)

Example 4n: The method of introducing probability into a problem whose statement is purely deterministic has been called the *probabilistic method* (Page 89)

3.5 $P(\cdot\,|F)$ is a probability

Conditional probabilities satisfy all of the properties of ordinary probabilities, which shows that

$P(E|F)$ satisfies the three axioms of a probability.

(a) $0 \le P(E|F) \le 1$.

(b) $P(S|F) = 1$.

(c) If $E_i, i = 1,2,\cdots,$ are mutually exclusive events, then

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

If we define $Q(E) = P(E|F)$, then, from Proposition 5.1, $Q(E)$ may be regarded as a probability function on the events of S. Hence, all of the propositions previously proved for probabilities apply to $Q(E)$. For instance, we have

$$Q(E_1 \cup E_2) = Q(E_1) + Q(E_2) - Q(E_1E_2)$$

or, equivalently,

$$P(E_1 \cup E_2|F) = P(E_1|F) + P(E_2|F) - P(E_1E_2|F)$$

Also, if we define the conditional probability $Q(E_1|E_2)$ by $Q(E_1|E_2) = Q(E_1E_2)/Q(E_2)$, then, we have

$$Q(E_1) = Q(E_1|E_2)Q(E_2) + Q(E_1|E_2^c)Q(E_2^c)$$

which is equivalent to

$$P(E_1|F) = P(E_1|E_2F)P(E_2|F) + P(E_1|E_2^cF)P(E_2^c|F)$$

An important concept in probability theory is that of the ***conditional independence*** of events. We say that the events E_1 and E_2 are conditionally independent given F if given that F occurs, the conditional probability that E_1 occurs is unchanged by information as to whether or not E_2 occurs. More formally, E_1 and E_2 are said to be conditionally independent given F if

$$P(E_1|E_2F) = P(E_1|F)$$

or, equivalently,

$$P(E_1E_2|F) = P(E_1|F)P(E_2|F)$$

Example 5e: Laplace's rule of succession (Page 95)

Example 5f: Updating information sequentially (Page 96)

Suppose there are n mutually exclusive and exhaustive possible hypotheses, with initial (sometimes referred to as *prior*) probabilities $P(H_i), \sum_{i=1}^{n} P(H_i) = 1$. Now, if information that the event E has occurred is received, then the conditional probability that H_i is the true hypothesis (sometimes referred to as the *updated* or *posterior* probability of H_i) is

$$P(H_i|E) = \frac{P(E|H_i)P(H_i)}{\sum_j P(E|H_j)P(H_j)}$$

Theoretical Exercises

1. Show that if $P(A) > 0$, then

$$P(AB|A) \ge P(AB|A \cup B).$$

$$P(AB|A) = \frac{P(AB)}{P(A)} \ge \frac{P(AB)}{P(A \cup B)} = P(AB|A \cup B)$$

2. Let $A \in B$. Express the following probabilities as simply as possible:

$$P(A|B), P(A|B^c), P(B|A), P(B|A^c).$$

$$P(A|B) = \frac{P(A)}{P(B)}, \quad P(A|B^c) = 0, \quad P(B|A) = 1, \quad P(B|A^c) = \frac{P(BA^c)}{P(A^c)}$$

5. (a) Prove that if E and F are mutually exclusive, then

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

(b) Prove that if E_i, $i \ge 1$ are mutually exclusive, then

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

6. Prove that if E_1, E_2, $\cdots$, E_n are independent events, then

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = 1 - \prod_{i=1}^{n} [1 - P(E_i)].$$

$$P\left(\bigcup_1^n E_i\right) = 1 - P\left(\bigcap_1^n E_i^c\right) = 1 - \prod_1^n [1 - P(E_i)]$$

25. Prove directly that

$$P(E|F) = P(E|FG)P(G|F) + P(E|FG^c)P(G^c|F).$$

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

$$P(E|FG)P(G|F) = \frac{P(EFG)}{P(FG)}\frac{P(FG)}{P(F)} = \frac{P(EFG)}{P(F)}$$

$$P(E|FG^c)P(G^c|F) = \frac{P(EFG^c)}{P(F)}.$$

The result now follows since

$$P(EF) = P(EFG) + P(EFG^c)$$

27. Extend the definition of conditional independence to more than 2 events.

$E_1, E_2, \cdots, E_n$ are conditionally independent given F if for all subsets $i_1, \cdots, i_r$ of $1,2,\cdots,n$

$$P\left(E_{i_1} \cdots E_{i_r}|F\right) = \prod_{j=1}^{r} P\left(E_{i_j}|F\right)$$

CHAPTER 4: RANDOM VARIABLES

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. For instance, in tossing dice, we are often interested in the sum of the two dice and are not really concerned about the separate values of each die. That is, we may be interested in knowing that the sum is 7 and may not be concerned over whether the actual outcome was $(1,6), (2,5), (3,4), (4,3), (5,2)$, or $(6,1)$. These real-valued **functions** defined on the sample space, are known as ***random variables***.

Because the value of a random variable is determined by the outcome of the experiment, we may assign probabilities to the possible values of the random variable.

Addition: A random variable translates the outcome of an experiment to an outcome value. As shown in Figure 4.1, Top: flipping a coin yields a head, which is mapped by the random variable X to the outcome value $X(\text{head}) = 1$, usually written as $X = 1$. Bottom: flipping a coin yields a tail, which is mapped to the outcome value $X(\text{tail}) = 0$, usually written as $X = 0$.

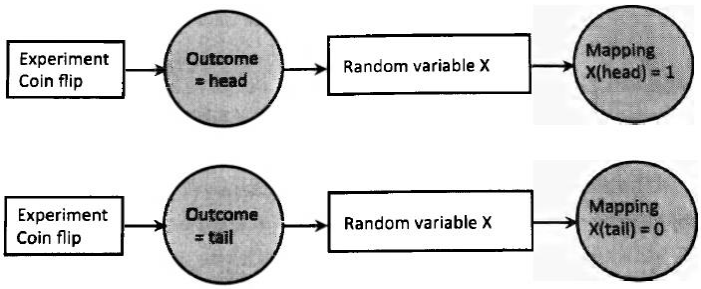


Figure 4. 1 Random variables.

The crucial point is that these outcomes are subject to a degree of randomness.

Random variables are not the same as the variables used in algebra.

The value of the random variable is a **mapping** from the experiment outcome to a numerical *outcome value*.

In our experiment, this function maps the coin flip outcome to the number of heads observed:

$$X(x_h) = 1,$$
$$X(x_t) = 0.$$

Thus, a random variable (function) takes an *argument* (e.g. x_h or x_t), and returns an outcome value (e.g. 0 or 1). An equivalent, and more conventional, notation for defining a random variable is

$$X = \begin{cases} 0, & \text{if the outcome is a tail,} \\ 1, & \text{if the outcome is a head.} \end{cases}$$

The subtle distinction between an outcome x and an outcome value $X(x)$ is sometimes vital, but in practice we only need to distinguish between them if the numbers of outcomes and outcome values are not equal (e.g. the two-dice example in Section 3.5). For example, suppose we roll a die, and we define the random variable X to be 0 when the outcome is an odd number and 1 when the outcome is an even number, so that

$$X = \begin{cases} 0, & \text{if the outcome } x \text{ is } 1,3 \text{ or } 5, \\ 1, & \text{if the outcome } x \text{ is } 2,4 \text{ or } 6. \end{cases}$$

In this case, the number of outcomes is six, but the number of outcome values is just two.

Key point. A random variable X is a function that maps each outcome x of an experiment (e.g. a coin flip) to a number $X(x)$, which is the outcome value of x. If the outcome value of x is 1 then this may be written as $X = 1$.

Example 1a: Suppose that our experiment consists of tossing 3 fair coins. If we let Y denote the number of heads that appear, then Y is a random variable taking on one of the values 0, 1, 2, and 3 with respective probabilities

$$P\{Y = 0\} = P\{(T,T,T)\} = \frac{1}{8}$$
$$P\{Y = 1\} = P\{(T,T,H),(T,H,T),(H,T,T)\} = \frac{3}{8}$$
$$P\{Y = 2\} = P\{(T,H,H),(H,T,H),(H,H,T)\} = \frac{3}{8}$$
$$P\{Y = 3\} = P\{(H,H,H)\} = \frac{1}{8}$$

For a random variable X, the function F defined by (here I think x is outcome value)

$$F(x) = P\{X \le x\} \quad -\infty < x < \infty$$

is called the ***cumulative*** *distribution function* or, more simply, the ***distribution function*** of X. Thus, the distribution function specifies, for all real values x, the probability that the random variable is less than or equal to x.

Now, suppose that $a \le b$. Then, because the event $\{X \le a\}$ is contained in the event $\{X \le b\}$, it follows that $F(a)$, the probability of the former, is less than or equal to $F(b)$, the probability of the latter. In other words, $F(x)$ is a nondecreasing function of x.

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** (A random variable whose set of possible values is either finite or countably infinite).

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 $x_1, x_2, \cdots$, then

$$p(x_i) \ge 0 \quad \text{for i} = 1,2,\cdots$$
$$p(x) = 0 \quad \text{for all other values of } x$$

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. For instance, if the probability mass function of X is

$$p(0) = \frac{1}{4}, \quad p(1) = \frac{1}{2}, \quad p(2) = \frac{1}{4}$$

we can represent this function graphically as shown in Figure 4.2. Similarly, a graph of the probability mass function of the random variable representing the sum when two dice are rolled looks like Figure 4.3.

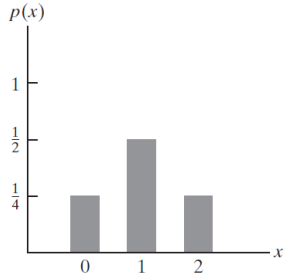


Figure 4. 2

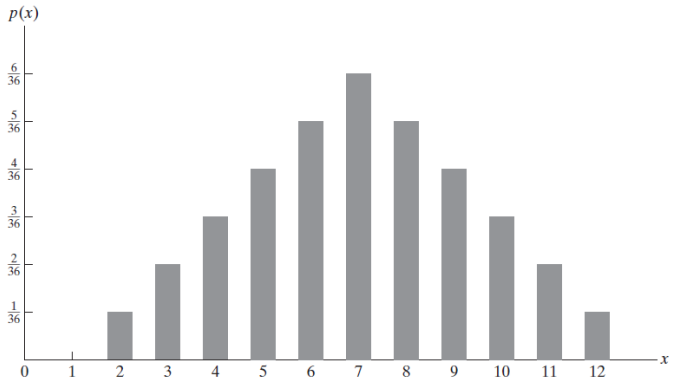


Figure 4. 3

Example 2a: The probability mass function of a random variable X is given by $p(i) = c\lambda^i/i!$, $i = 0,1,2,\cdots$, 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

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

Hence,

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

(b) $P\{X > 2\} = 1 - P\{X \le 2\} = 1 - P\{X = 0\} - P\{X = 1\} - P\{X = 2\}$

$$= 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2 e^{-\lambda}}{2}.$$

The cumulative distribution function F can be expressed in terms of $p(a)$ by

$$F(a) = \sum_{\text{all } x \le a} p(x)$$

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

If X is a discrete random variable whose possible values are $x_1, x_2, x_3, \cdots$, where $x_1 < x_2 < x_3 < \cdots$, then the distribution function F of X is a step function. That is, the value of F is constant in the intervals (x_{i-1}, x_i) and then takes a step (or jump) of size $p(x_i)$ at x_i. For instance, if X has a probability mass function given by

$$p(1) = \frac{1}{4} \quad p(2) = \frac{1}{2} \quad p(3) = \frac{1}{8} \quad p(4) = \frac{1}{8}$$

then its cumulative distribution function is

$$F(a) = \begin{cases} 0 & a < 1 \\ 1/4 & 1 \le a < 2 \\ 3/4 & 2 \le a < 3 \\ 7/8 & 3 \le a < 4 \\ 1 & 4 \le a \end{cases}$$

This function is depicted graphically in Figure 4.4.

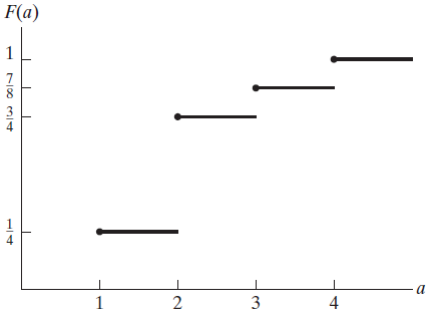


Figure 4. 4

4.3 Expected values

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.

Another motivation of the definition of expectation is provided by the frequency interpretation of probabilities. This interpretation (partially justified by the strong law of large numbers, to be presented in Chapter 8) assumes that if an infinite sequence of independent replications of an experiment is performed, then, for any event E, the proportion of time that E occurs will be $P(E)$. Now, consider a random variable X that must take on one of the values $x_1, x_2, \cdots x_n$ with respective probabilities $p(x_1), p(x_2), \cdots, p(x_n)$, and think of X as representing our winnings in a single game of chance. That is, with probability $p(x_i)$, we shall win x_i units $i = 1,2, \cdots, n$. By the frequency interpretation, if we play this game continually, then the proportion of time that we win x_i will be $p(x_i)$. Since this is true for all i, $i = 1,2, \cdots, n$, it follows that our average winnings per game will be

$$\sum_{i=1}^{n} x_i p(x_i) = E[X]$$

Example 3b: We say that I is an indicator variable for the event A if

$$I = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A^c \text{ occurs} \end{cases}$$

Find $E[I]$.

Solution Since $p(1) = P(A), p(0) = 1 - P(A)$, we have

$$E[I] = P(A)$$

That is, the expected value of the indicator variable for the event A is equal to the probability that A occurs.

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 accomplish 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 $\mathrm{E}[g(X)]$ by using the definition of expected value.

Example 4a: Let X denote a random variable that takes on any of the values -1, 0, and 1 with respective probabilities

$$P\{X = -1\} = .2 \quad P\{X = 0\} = .5 \quad P\{X = 1\} = .3$$

Compute $E[X^2]$.

Solution

Let $Y = X^2$. Then the probability mass function of Y is given by

$$P\{Y = 1\} = P\{X = -1\} + P\{X = 1\} = .5$$
$$P\{Y = 0\} = P\{X = 0\} = .5$$

Hence,

$$E[X^2] = E[Y] = 1(.5) + 0(.5) = .5$$

Note that

$$.5 = E[X^2] \neq (E[X])^2 = .01$$

Although the preceding procedure will always enable us to compute the expected value of any function of X from a knowledge of the probability mass function of X, there is **another** way of thinking about $E[g(X)]$: Since $g(X)$ will equal $g(x)$ whenever X is equal to x, it seems reasonable that $E[g(X)]$ should just be a weighted average of the values $g(x)$, with $g(x)$ being weighted by the probability that X is equal to x. That is, the following result is quite intuitive.

Proposition 4.1: 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,

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

In other words, the worth of an action can be measured by the expected value of the utility of its consequence, and the action with the largest expected utility is the most preferable. (Page 125)

Corollary 4.1: If a and b are constants, then

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

Proof:

$$E[aX + b] = \sum_{x:p(x)>0} (ax + b)p(x)$$

$$= a \sum_{x:p(x)>0} xp(x) + b \sum_{x:p(x)>0} p(x)$$

$$= aE[X] + b.$$

The expected value of a random variable X, $E[X]$ is also referred to as the ***mean*** or the *first moment of* X. The quantity $E[X^n], n \geq 1$, is called the *nth moment* of X.

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

4.5 Variance

Given a random variable X along with its distribution function F, it would be extremely useful if we were able to summarize the essential properties of F by certain suitably defined measures. One such measure would be $E[X]$, the expected value of X.

Because we expect X to take on values around its mean $E[X]$, a reasonable way of measuring the possible variation of X would be to look at how far apart X would be from its mean, on the average. One possible way to measure this variation would be to consider the quantity $E[|X - \mu|]$, where $\mu = E[X]$. However, it turns out to be mathematically inconvenient to deal with this quantity, so a more tractable quantity is usually considered—namely, the expectation of the square of the difference between X and its mean.

Definition

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

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

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

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

$$= \sum_x (x - \mu)^2 p(x)$$

$$= \sum_x (x^2 - 2\mu x + \mu^2)p(x)$$

$$= \sum_x x^2 p(x) - 2\mu \sum_x xp(x) + \mu^2 \sum_x p(x)$$

$$= E[X^2] - 2\mu^2 + \mu^2$$

$$= E[X^2] - \mu^2$$

That is,

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

In words, the variance of X is equal to the expected value of X^2 minus the square of its expected value. In practice, this formula frequently offers the easiest way to compute $\text{Var}(X)$.

A useful identity is that for any constants a and b,

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

Proof:

Let $\mu = E[X]$ and note from Corollary 4.1 that $E[aX + b] = a\mu + b$. Therefore,

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

Remarks (a) 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.

(b) 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,

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

4.6 The 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

$$\begin{aligned}p(0) &= P\{X = 0\} = 1 - p \\ p(1) &= P\{X = 1\} = p\end{aligned}$$

where $p, 0 \le p \le 1$, is the probability that the trial is a success.

A random variable X is said to be a ***Bernoulli*** *random variable* (after the Swiss mathematician James Bernoulli) if its probability mass function is given by above equation 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

$$p(i) = \binom{n}{i} p^i(1 - p)^{n-i} \quad i = 0,1,\cdots,n$$

Note that, by the binomial theorem, the probabilities sum to 1; that is,

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{n} \binom{n}{i} p^i(1 - p)^{n-i} = [p + (1 - p)]^n = 1$$

Example 6a: Five fair coins are flipped. If the outcomes are assumed independent, find the probability mass function of the number of heads obtained.

Solution

If we let X equal the number of heads (successes) that appear, then X is a binomial random variable with parameters $(n = 5,\ p = \frac{1}{2})$. Hence,

$$P\{X=0\}=\binom{5}{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{5}=\frac{1}{32}$$

$$P\{X=1\}=\binom{5}{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{4}=\frac{5}{32}$$

$$P\{X=2\}=\binom{5}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{3}=\frac{10}{32}$$

$$P\{X=3\}=\binom{5}{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{2}=\frac{10}{32}$$

$$P\{X=4\}=\binom{5}{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{1}=\frac{5}{32}$$

$$P\{X=5\}=\binom{5}{5}\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{0}=\frac{1}{32}$$

● Properties of binomial random variables

If X is a binomial random variable with parameters n and p, then

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

$$=\sum_{i=1}^{n} i^k\binom{n}{i}p^i(1-p)^{n-i}$$

Using the identity

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

gives

$$E[X^k]=np\sum_{i=1}^{n} i^{k-1}\binom{n-1}{i-1}p^{i-1}(1-p)^{n-i}$$

$$=np\sum_{j=0}^{n-1}(j+1)^{k-1}\binom{n-1}{j}p^j(1-p)^{n-1-j} \quad \begin{matrix}\text{by letting}\\ j=i-1\end{matrix}$$

$$=np\,E[(Y+1)^{k-1}]$$

where Y is a binomial random variable with parameters $n-1$, p. Setting $k=1$ in the preceding equation yields

$$E[X]=np$$

That is, the expected number of successes that occur in n independent trials when each is a success with probability p is equal to np. Setting $k=2$ in the preceding equation and using the preceding formula for the expected value of a binomial random variable yields

$$E[X^2]=np\,E[Y+1]$$
$$=np[(n-1)p+1]$$
$$\mathrm{Var}[X]=E[X^2]-(E[X])^2$$
$$=np(1-p)$$

The following proposition details how the binomial probability mass function first increases and then decreases.

Proposition 6.1: 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 $(n+1)p$.

$$\because \frac{P\{X=k\}}{P\{X=k-1\}}=\frac{\frac{n!}{(n-k)!\,k!}p^k(1-p)^{n-k}}{\frac{n!}{(n-k+1)!\,(k-1)!}p^{k-1}(1-p)^{n-k+1}}=\frac{(n-k+1)p}{k(1-p)}$$

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

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

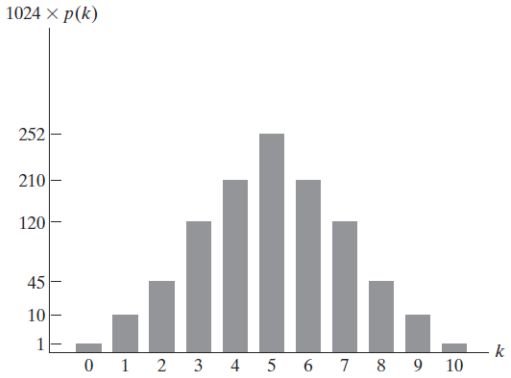


Figure 4. 5 Graph of $p(k)=\binom{10}{k}\left(\frac{1}{2}\right)^{10}$

● Computing the Binomial Distribution Function

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

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

is to utilize the following relationship between $P\{X=k+1\}$ and $P\{X=k\}$, which was established in the proof of Proposition 6.1:

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

4.7 The Poisson random variable

A random variable X that takes on one of the values $0,1,2,\cdots$ is said to be a ***Poisson*** random variable with parameter λ if, for some $\lambda>0$,

$$p(i)=P\{X=i\}=e^{-\lambda}\frac{\lambda^i}{i!}\quad i=0,1,2,\cdots$$

which defines a probability mass function. Since

$$\sum_{i=0}^{\infty}p(i)=e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}=e^{-\lambda}e^{\lambda}=1$$

The Poisson random variable has a tremendous range of applications in diverse areas because it may be used as an approximation for a binomial random variable with parameters (n,p) when n is large and p is small enough so that np is of moderate size. To see this, suppose that X is a binomial random variable with parameters (n,p), and let $\lambda=np$. Then

$$\begin{aligned}P\{X=i\}&=\frac{n!}{(n-i)!\,i!}p^i(1-p)^{n-i}\\&=\frac{n!}{(n-i)!\,i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i}\\&=\frac{n(n-1)\cdots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}\end{aligned}$$

Now, for n large and λ moderate,

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \quad \frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1, \quad \left(1-\frac{\lambda}{n}\right)^i \approx 1$$

Hence, for n large and λ moderate,

$$P\{X=i\} \approx e^{-\lambda}\frac{\lambda^i}{i!}$$

In other words, if n independent trials, each of which results in a success with probability p, are performed, then when n is large and p is small enough to make np moderate, the number of successes occurring is approximately a Poisson random variable with parameter $\lambda = np$. This value λ (which will later be shown to equal the expected number of successes) will usually be determined empirically.

Computing the expected value and variance of the Poisson random variable with parameter λ,

$$\begin{aligned}
E[X] &= \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!} \\
&= \lambda \sum_{i=1}^{\infty} \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \quad \text{by letting } j = i-1 \\
&= \lambda \quad \text{since } \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{\lambda} \\
E[X^2] &= \sum_{i=0}^{\infty} \frac{i^2 e^{-\lambda}\lambda^i}{i!} \\
&= \lambda \sum_{i=1}^{\infty} \frac{ie^{-\lambda}\lambda^{i-1}}{(i-1)!} \\
&= \lambda \sum_{j=0}^{\infty} \frac{(j+1)e^{-\lambda}\lambda^j}{j!} \quad \text{by letting } j = i-1 \\
&= \lambda \left[\sum_{j=0}^{\infty} \frac{je^{-\lambda}\lambda^j}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} \right] \\
&= \lambda[\lambda+1]
\end{aligned}$$

where the final equality follows because the first sum is the expected value of a Poisson random variable with parameter λ and the second is the sum of the probabilities of this random variable.

$$\begin{aligned}
\mathrm{Var}[X] &= E[X^2] - (E[X])^2 \\
&= \lambda
\end{aligned}$$

Hence, the expected value and variance of a Poisson random variable are both equal to its parameter λ.

In fact, it remains a good approximation even when the trials are not independent, provided that their dependence is weak.

For a second illustration of the strength of the Poisson approximation when the trials are weakly dependent, (Page 138)

For the number of events to occur to approximately have a Poisson distribution, it is not essential that all the events have the same probability of occurrence, but only that all these probabilities be small. The following is referred to as the *Poisson paradigm*.

Example 7d: Length of the longest run (inclusion–exclusion identity) (Page 140)

Another use of the Poisson probability distribution arises in situations where "events" occur at certain points in time. (Page 144)

- Computing the Poisson Distribution Function

If X is Poisson with parameter λ, then

$$\frac{P\{X=i+1\}}{P\{X=i\}}=\frac{e^{-\lambda}\lambda^{i+1}/(i+1)!}{e^{-\lambda}\lambda^{i}/i!}=\frac{\lambda}{i+1}$$

Starting with $P\{X=0\}=e^{-\lambda}$, we can use it to compute successively

$$P\{X=1\}=\lambda P\{X=0\}$$
$$P\{X=2\}=\frac{\lambda}{2}P\{X=1\}$$
$$\vdots$$
$$P\{X=i+1\}=\frac{\lambda}{i+1}P\{X=i\}$$

4.8 Other discrete probability distributions

● The 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

$$P\{X=n\}=(1-p)^{n-1}p,\quad n=1,2,\cdots$$

The above equation follows because, in order for X to equal n, it is necessary and sufficient that the first $n-1$ trials are failures and the nth trial is a success. The above equation then follows, since the outcomes of the successive trials are assumed to be independent.

Since

$$\sum_{n=1}^{\infty}P\{X=n\}=p\sum_{n=1}^{\infty}(1-p)^{n-1}=\frac{p}{1-(1-p)}=1$$

it follows that with probability 1, a success will eventually occur. Any random variable X whose probability mass function is given by above equation $P\{X=n\}$ is said to be a ***geometric*** random variable with parameter p.

Example 8b: Find the expected value of a geometric random variable.

Solution

with $q=1-p$, we have

$$E[X]=\sum_{i=1}^{\infty}iq^{i-1}p$$
$$=\sum_{i=1}^{\infty}(i-1+1)q^{i-1}p$$
$$=\sum_{i=1}^{\infty}(i-1)q^{i-1}p+\sum_{i=1}^{\infty}q^{i-1}p$$
$$=\sum_{j=0}^{\infty}jq^{j}p+1$$
$$=q\sum_{j=1}^{\infty}jq^{j-1}p+1$$
$$=qE[X]+1$$

Hence,

$$p\,E[X]=1$$

yielding the result

$$E[X]=\frac{1}{p}$$

In other words, if independent trials having a common probability p of being successful are performed until the first success occurs, then the expected number of required trials equals $1/p$. For instance, the expected number of rolls of a fair die that it takes to obtain the value 1 is 6.

Example 8c: Find the variance of a geometric random variable.

Solution

With $q=1-p$, we have

$$E[X^2]=\sum_{i=1}^{\infty}i^2q^{i-1}p$$

$$= \sum_{i=1}^{\infty} (i-1+1)^2 q^{i-1} p$$

$$= \sum_{i=1}^{\infty} (i-1)^2 q^{i-1} p + \sum_{i=1}^{\infty} 2(i-1) q^{i-1} p + \sum_{i=1}^{\infty} q^{i-1} p$$

$$= \sum_{j=0}^{\infty} j^2 q^j p + 2\sum_{i=1}^{\infty} i q^{i-1} p - 2\sum_{i=1}^{\infty} q^{i-1} p + \sum_{i=1}^{\infty} q^{i-1} p$$

$$= qE[X^2] + 2E[X] - 1$$

Using $E[X] = 1/p$,

$$E[X^2] = \frac{q+1}{p^2}$$

giving the result

$$\mathrm{Var}(X) = \frac{q+1}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

● The Negative Binomial Random Variable

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

$$P\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r} \quad n = r, r+1, \cdots$$

which follows because, in order for the rth success to occur at the nth trial, there must be $r-1$ successes in the first $n-1$ trials and the nth trial must be a success. The probability of the first event is

$$\binom{n-1}{r-1} p^{r-1} (1-p)^{n-r}$$

and the probability of the second is p; thus, by independence, the equation $P\{X = n\}$ is established. To verify that a total of r successes must eventually be accumulated, either we can prove analytically that

$$\sum_{n=r}^{\infty} P\{X = n\} = \sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (1-p)^{n-r} = 1$$

or we can give a probabilistic argument. (Page 149)

Any random variable X whose probability mass function is given by above equation $P\{X = n\}$ is said to be a ***negative binomial*** random variable with parameters (r, p). Note that a geometric random variable is just a negative binomial with parameter $(1, p)$.

Example 8e: The Banach match problem

Example 8f: Compute the expected value and the variance of a negative binomial random variable with parameters r and p.

$$E[X] = \frac{r}{p}$$

$$E[X^2] = \frac{r}{p}\left(\frac{r+1}{p} - 1\right)$$

$$\mathrm{Var}(X) = \frac{r}{p}\left(\frac{r+1}{p} - 1\right) - \left(\frac{r}{p}\right)^2$$

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

Thus, if independent trials, each of which is a success with probability p, are performed, then the expected value and variance of the number of trials that it takes to amass r successes is r/p and $r(1-p)/p^2$, respectively.

Since a geometric random variable is just a negative binomial with parameter $r = 1$, it follows from the preceding example that the variance of a geometric random variable with parameter p is equal to $(1-p)/p^2$.

● The Hypergeometric Random Variable

Suppose that a sample of size n is to be chosen randomly (without replacement) from an urn

containing N balls, of which m are white and $N-m$ are black. If we let X denote the number of white balls selected, then

$$P\{X=i\}=\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \quad i=0,1,\cdots,n$$

A random variable X whose probability mass function is given by above equation for some values of n,N,m is said to be a ***hypergeometric*** random variable.

Remark Although we have written the hypergeometric probability mass function with i going from 0 to n, $P\{X=i\}$ will actually be 0, unless i satisfies the inequalities $n-(N-m)\le i\le \min(n,m)$, However, above equation $P\{X=i\}$ is always valid because of our convention that $\binom{r}{k}$ is equal to 0 when either $k<0$ or $r<k$.

Example 8h: Such an estimate is called a *maximum likelihood* estimate.

If n balls are randomly chosen without replacement from a set of N balls of which the fraction $p=m/N$ is white, then the number of white balls selected is hypergeometric. Now, it would seem that when m and N are large in relation to n, it shouldn't make much difference whether the selection is being done with or without replacement, because, no matter which balls have previously been selected, when m and N are large, each additional selection will be white with a probability approximately equal to p. In other words, it seems intuitive that when m and N are large in relation to n, the probability mass function of X should approximately be that of a binomial random variable with parameters n and p.

Example 8j: Determine the expected value and the variance of X, a hypergeometric random variable with parameters n, N, and m.

$$E[X]=\frac{nm}{N}$$

$$E[X^2]=\frac{nm}{N}\left[\frac{(n-1)(m-1)}{N-1}+1\right]$$

$$\mathrm{Var}(X)=\frac{nm}{N}\left[\frac{(n-1)(m-1)}{N-1}+1-\frac{nm}{N}\right]$$

Letting $p=m/N$

$$\mathrm{Var}(X)=np(1-p)\left(1-\frac{n-1}{N-1}\right)$$

Remark We have shown that if n balls are randomly selected without replacement from a set of N balls, of which the fraction p are white, then the expected number of white balls chosen is np. In addition, if N is large in relation to n [so that $(N-n)/(N-1)$ is approximately equal to 1], then

$$\mathrm{Var}(X)\approx np(1-p)$$

In other words, $E[X]$ is the same as when the selection of the balls is done with replacement (so that the number of white balls is binomial with parameters n and p), and if the total collection of balls is large, then $\mathrm{Var}(X)$ is approximately equal to what it would be if the selection were done with replacement. This is, of course, exactly what we would have guessed, given our earlier result that when the number of balls in the urn is large, the number of white balls chosen approximately has the mass function of a binomial random variable.

● The Zeta (or Zipf) Distribution

A random variable is said to have a zeta (sometimes called the Zipf) distribution if its probability mass function is given by

$$P\{X=k\}=\frac{C}{k^{\alpha+1}} \quad k=1,2,\cdots$$

for some value of $\alpha>0$. Since the sum of the foregoing probabilities must equal 1, it follows that

$$C=\left[\sum_{k=1}^{\infty}\left(\frac{1}{k}\right)^{\alpha+1}\right]^{-1}$$

The zeta distribution owes its name to the fact that the function

$$\zeta(s) = 1 + \left(\frac{1}{2}\right)^s + \left(\frac{1}{3}\right)^s + \cdots + \left(\frac{1}{k}\right)^s + \cdots$$

is known in mathematical disciplines as the Riemann zeta function.

4.9 Expected value of sums of random variables

A very important property of expectations is that the expected value of a sum of random variables is equal to the sum of their expectations. Under the assumption that the set of possible values of the probability experiment—that is, the sample space S—is either finite or countably infinite. Although the result is true without this assumption.

For a random variable X, let $X(s)$ denote the value of X when $s \in S$ is the outcome of the experiment. Now, if X and Y are both random variables, then so is their sum. That is, $Z = X + Y$ is also a random variable. Moreover, $Z(s) = X(s) + Y(s)$.

Let $p(s) = P(\{s\})$ be the probability that s is the outcome of the experiment. Because we can write any event A as the finite or countably infinite union of the mutually exclusive events $\{s\}$, $s \in A$, it follows by the axioms of probability that

$$P(A) = \sum_{s \in A} p(s)$$

When $A = S$, the preceding equation gives

$$1 = \sum_{s \in S} p(s)$$

Now, let X be a random variable, and consider $E[X]$. Because $X(s)$ is the value of X when s is the outcome of the experiment, it seems intuitive that $E[X]$ - the weighted average of the possible values of X, with each value weighted by the probability that X assumes that value—should equal a weighted average of the values $X(s)$, $s \in S$, with $X(s)$ weighted by the probability that s is the outcome of the experiment.

Proposition 9.1:

$$E[X] = \sum_{s \in S} X(s)p(s)$$

Corollary 9.2: For random variables $X_1, X_2, \cdots, X_n$,

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Proof:

Let $Z = \sum_{i=1}^{n} X_i$,

$$E[Z] = \sum_{s \in S} Z(s)p(s)$$

$$= \sum_{s \in S} \big(X_1(s) + X_2(s) + \cdots + X_n(s)\big)p(s)$$

$$= \sum_{s \in S} X_1(s)p(s) + \sum_{s \in S} X_2(s)p(s) + \cdots + \sum_{s \in S} X_n(s)p(s)$$

$$= E[X_1] + E[X_2] + \cdots + E[X_n].$$

Example 9d: Find the expected **total** number of successes that result from n trials when trial i is a success with probability p_i, $i = 1, \cdots, n$.

Solution Letting

$$X_i = \begin{cases} 1, & \text{if trial } i \text{ is a success} \\ 0, & \text{if trial } i \text{ is a failure} \end{cases}$$

we have the representation

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

Consequently,

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p_i$$

Note that this result does **not** require that the trials be independent. It includes as a special case the expected value of a binomial random variable, which assumes independent trials and all $p_i = p$, and thus has mean np. It also gives the expected value of a hypergeometric random variable representing the number of white balls selected when n balls are randomly selected, without replacement, from an urn of N balls of which m are white. We can interpret the hypergeometric as representing the number of successes in n trials, where trial i is said to be a success if the ith ball selected is white. Because the ith ball selected is equally likely to be any of the N balls and thus has probability m/N of being white, it follows that the hypergeometric is the number of successes in n trials in which each trial is a success with probability $p = m/N$. Hence, even though these hypergeometric trials are dependent, it follows from the result of Example 9d that the expected value of the hypergeometric is $np = nm/N$.

Example 9e: Derive an expression for the variance of the number of successful trials in Example 9d, and apply it to obtain the variance of a binomial random variable with parameters n and p, and of a hypergeometric random variable equal to the number of white balls chosen when n balls are randomly chosen from an urn containing N balls of which m are white.

$$E[X^2] = E\left[\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} X_j\right)\right]$$
$$= \sum_i p_i + \sum_{i=1}^{n}\sum_{j\neq i} E[X_i X_j]$$

4.10 Properties of the 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, $n \geq 1$, that converges to $b, \lim_{n\to\infty} F(b_n) = F(b)$

All probability questions about X can be answered in terms of the c.d.f., F. For example,

$$P\{a < X \leq b\} = F(b) - F(a) \quad \text{for all } a < b$$

If we want to compute the probability that X is strictly less than b, we can again apply the continuity property to obtain

$$P\{X < b\} = P\left(\lim_{n\to\infty}\left\{X \leq b - \frac{1}{n}\right\}\right)$$
$$= \lim_{n\to\infty} P\left(X \leq b - \frac{1}{n}\right)$$
$$= \lim_{n\to\infty} F\left(b - \frac{1}{n}\right)$$

Example 10a: (Page 161)

Theoretical Exercises

4.2. If X has distribution function F, what is the distribution function of e^X?

Take any $a \in e^X$, we have that

$$F_{e^X}(a) = P(e^X \leq a) = P(X \leq \log(a)) = F_X(\log(a))$$

So, the distribution of e^X can be written as composition of logarithm function and the distribution function of X.

4.3. If X has distribution function F, what is the distribution function of the random variable $\alpha X + \beta$, where α and β are constants, $\alpha \neq 0$?

Take any $a \in \alpha X + \beta$. Suppose that $\alpha > 0$, we have that

$$F_{\alpha X+\beta}(a) = P(\alpha X + \beta \leq a) = P(\alpha X \leq a - \beta) = P\left(X \leq \frac{a-\beta}{\alpha}\right) = F_X\left(\frac{a-\beta}{\alpha}\right)$$

Similarly, if $\alpha < 0$, we get that

$$F_{\alpha X+\beta}(a) = P(\alpha X + \beta \leq a) = P(\alpha X \leq a - \beta) = P\left(X \geq \frac{a-\beta}{\alpha}\right)$$
$$= 1 - P\left(X \leq \frac{a-\beta}{\alpha}\right) = 1 - F_X\left(\frac{a-\beta}{\alpha}\right)$$

So, the distribution of $\alpha X + \beta$ can be written as composition of some functions that include distribution function of X.

CHAPTER 5: CONTINUOUS RANDOM VARIABLES

5.1 Introduction

In Chapter 4, we considered discrete random variables—that is, random variables whose set of possible values is either finite or countably infinite. However, there also exist random variables whose set of possible values is uncountable. Two examples are the time that a train arrives at a specified stop and the lifetime of a transistor. Let X be such a random variable. 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,

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

The function f is called the ***probability density function*** (p. d. f) of the random variable X. In words, the equation above 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

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

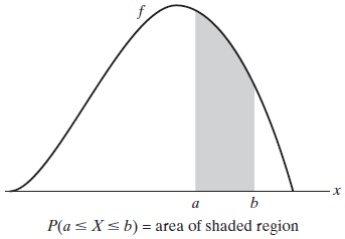


Figure 5. 1 Probability density function f.

If we let $a = b$ in the equation above, we get

$$P\{X = a\} = \int_a^a f(x)dx = 0$$

In words, this equation states that the probability that a continuous random variable will assume any fixed value is zero. Hence, for a continuous random variable,

$$P\{X < a\} = P\{X \le a\} = F(a) = \int_{-\infty}^{a} f(x)\,dx$$

The relationship between the cumulative distribution F and the probability density f is expressed by

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

Differentiating both sides of the preceding equation yields

$$\frac{d}{da}F(a) = f(a)$$

That is, the density is the **derivative** of the cumulative distribution function. A somewhat more intuitive interpretation of the density function may be obtained from $P\{a \le X \le b\} = \int_a^b f(x)\,dx$ as follows:

$$P\left\{a - \frac{\varepsilon}{2} \le X \le a + \frac{\varepsilon}{2}\right\} = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)\,dx \approx \varepsilon f(a)$$

when ε is small and when $f(\cdot)$ is continuous at $x = a$. In other words, the probability that X

will be contained in an interval of length ε around the point a is approximately $\varepsilon f(a)$. From this result, we see that $f(a)$ is a measure of how likely it is that the random variable will be near a.

Example 1d: If X is continuous with distribution function F_X and density function f_X, find the density function of $Y = 2X$.

Solution We will determine f_Y in two ways. The first way is to derive, and then differentiate, the distribution function of Y:

$$\begin{aligned} F_Y(a) &= P\{Y \le a\} \\ &= P\{2X \le a\} \\ &= P\{X \le a/2\} \\ &= F_X(a/2) \end{aligned}$$

Differentiation gives

$$f_Y(a) = \frac{1}{2} f_X(a/2)$$

Another way to determine f_Y is to note that

$$\begin{aligned} \varepsilon f_Y(a) &\approx P\left\{a - \frac{\varepsilon}{2} \le Y \le a + \frac{\varepsilon}{2}\right\} \\ &= P\left\{a - \frac{\varepsilon}{2} \le 2X \le a + \frac{\varepsilon}{2}\right\} \\ &= P\left\{\frac{a}{2} - \frac{\varepsilon}{4} \le X \le \frac{a}{2} + \frac{\varepsilon}{4}\right\} \\ &\approx \frac{\varepsilon}{2} f_X(a/2) \end{aligned}$$

Dividing through by ε gives the same result as before.

5.2 Expectation and variance of continuous random variables

If X is a continuous random variable having probability density function $f(x)$, then, because

$$f(x)\,dx \approx P\{x \le X \le x + dx\} \quad \text{for } dx \text{ small}$$

it is easy to see that the analogous definition is to define the expected value of X by

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

Example 2b: The density function of X is given by

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

Find $E[e^X]$.

Solution Let $Y = e^X$. We start by determining F_Y, the cumulative distribution function of Y. Now, for $1 \le y \le e$ $(0 \le \log(y) \le 1)$,

$$\begin{aligned} F_Y(y) &= P\{Y \le y\} \\ &= P\{e^X \le y\} \\ &= P\{X \le \log(y)\} \\ &= \int_{-\infty}^{\log(y)} f(x)dx = \int_{0}^{\log(y)} 1dx \\ &= \log(y) \end{aligned}$$

By differentiating $F_Y(y)$, we can conclude that the probability density function of Y is given by

$$f_Y(y) = \frac{1}{y} \quad 1 \le y \le e$$

Hence,

$$\begin{aligned} E[e^X] = E[Y] &= \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= \int_{1}^{e} dy \\ &= e - 1 \end{aligned}$$

Although the method employed in Example 2b to compute the expected value of a function of X is always applicable, there is, as in the discrete case, an alternative way of proceeding. The following is a direct analog of Proposition 4.1 of Chapter 4.

Proposition 2.1: If X is a continuous random variable with probability density function $f(x)$,

then, for any real-valued function g,

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

Corollary 2.1: If a and b are constants, then

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

for a continuous random variable X.

The variance of a continuous random variable is defined exactly as it is for a discrete random variable, namely, if X is a random variable with expected value μ, then the variance of X is defined (for any type of random variable) by

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

The alternative formula,

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

is established in a manner similar to its counterpart in the discrete case.

It can be shown that, for constants a and b,

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

The proof mimics the one given for discrete random variables.

There are several important classes of continuous random variables that appear frequently in applications of probability; the next few sections are devoted to a study of some of them.

5.3 The 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

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

Note that the above equation is a density function, since $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 dx = 1$. 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 $(0,1)$ 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

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

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

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

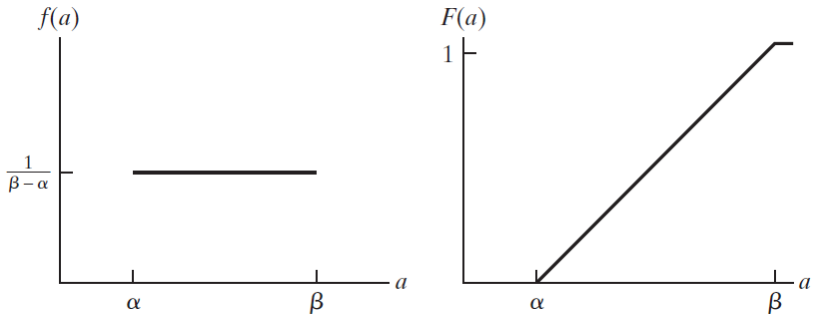


Figure 5. 2 Graph of $f(a)$ and $F(a)$ for a uniform (α, β) random variable.

Example 3a: Let X be uniformly distributed over (α, β). Find (a) $E[X]$ and (b) $\text{Var}(X)$.

Solution

(a)

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

$$= \frac{\beta + \alpha}{2}$$

In words, the expected value of a random variable that is uniformly distributed over some interval is equal to the midpoint of that interval.

(b)

$$E[X^2] = \int_{\alpha}^{\beta} \frac{1}{\beta - \alpha} x^2\,dx$$

$$= \frac{\beta^2 + \alpha\beta + \alpha^2}{3}$$

Hence,

$$\text{Var}(X) = \frac{\beta^2 + \alpha\beta + \alpha^2}{3} - \frac{(\alpha + \beta)^2}{4}$$

$$= \frac{(\beta - \alpha)^2}{12}$$

Therefore, the variance of a random variable that is uniformly distributed over some interval is the square of the length of that interval divided by 12.

Example 3d: Bertrand's paradox. It represents our initial introduction to a subject commonly referred to as geometrical probability. (Page 186)

5.4 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

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

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

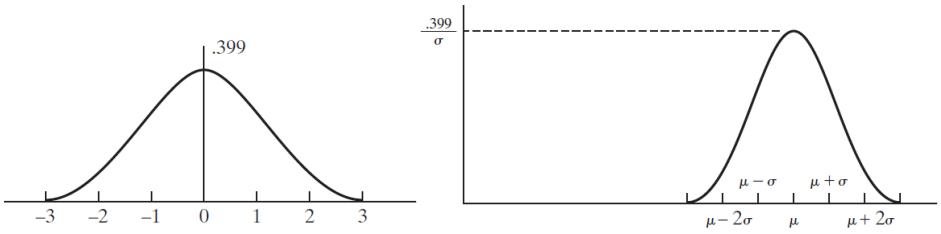


Figure 5. 3 Normal density function: (left) $\mu = 0, \sigma = 1$; (right) arbitrary μ, σ^2

The normal distribution was introduced by the French mathematician Abraham DeMoivre in 1733, who used it to approximate probabilities associated with binomial random variables when the binomial parameter n is large. This result was later extended by Laplace and others and is now encompassed in a probability theorem known as the central limit theorem, which is discussed in Chapter 8. The *central limit theorem*, one of the two most important results in probability theory, gives a theoretical base to the often-noted empirical observation that, in practice, many random phenomena obey, at least approximately, a normal probability distribution. Some examples of random phenomena obeying this behavior are the height of a man or woman, the velocity in any direction of a molecule in gas, and the error made in measuring a physical quantity.

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

$$I^2 = \int_{-\infty}^{\infty} e^{-y^2/2}dy \int_{-\infty}^{\infty} e^{-x^2/2}dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(y^2+x^2)/2}dydx$$

We now evaluate the double integral by means of a change of variables to polar coordinates. (That is, let $x = r\cos\theta$, $y = r\sin\theta$, and $dy\,dx = rd\theta dr$.) Thus,

$$I^2 = \int_0^{\infty}\int_0^{2\pi} e^{-r^2/2}rd\theta dr$$

$$= 2\pi\int_0^{\infty} re^{-r^2/2}dr$$

$$= -2\pi e^{-r^2/2}|_0^{\infty}$$

$$= 2\pi$$

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 $a^2\sigma^2$. (Page 188)

An important implication of the preceding result is that if X is normally distributed with parameters μ and σ^2, then $Z = (X-\mu)/\sigma$ is normally distributed with parameters 0 and 1. Such a random variable is said to be a ***standard***, or a ***unit***, normal random variable.

Example 4a: Find $E[X]$ and $\mathrm{Var}(X)$ when X is a normal random variable with parameters μ and σ^2,

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

$$E[Z] = \int_{-\infty}^{\infty} xf_Z(x)\,dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-x^2/2}\,dx$$

$$= 0$$

$$\mathrm{Var}(Z) = E[Z^2]$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2e^{-x^2/2}\,dx$$

$$= 1$$

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 (Page 190: A table for area $\Phi(x)$ under the standard normal curve to the left of X. $\Phi(1.96) - \Phi(-1.96) = 95\%$)

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

Table 5.1 Area $\Phi(x)$ Under the Standard Normal Curve to the Left of X.

X	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

The values of $\Phi(x)$ for nonnegative x are given in Table 5.1. For negative values of x, $\Phi(x)$ can be obtained from the relationship (following from the symmetry of the standard normal density)

$$\Phi(-x) = 1 - \Phi(x) \quad -\infty < x < \infty$$

This equation states that if Z is a standard normal random variable, then

$$P\{Z \le -x\} = P\{Z > x\} \quad -\infty < x < \infty$$

Since $Z = (X - \mu)/\sigma$ is a standard normal random variable whenever X is normally distributed with parameters μ and σ^2, it follows that the distribution function of X can be expressed as

$$F_X(a) = P\{X \le a\} = P\left(\frac{X-\mu}{\sigma} \le \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

● The Normal Approximation to the Binomial Distribution

An important result in probability theory known as the DeMoivre–Laplace limit theorem states that when n is large, a binomial random variable with parameters n and p will have approximately the same distribution as a normal random variable with the same **mean** and variance as the binomial. It formally states that if we "**standardize**" the binomial by first subtracting its mean np and then dividing the result by its standard deviation $\sqrt{np(1-p)}$,

then the distribution function of this standardized random variable (which has mean 0 and variance 1) will converge to the standard normal distribution function as $n \to \infty$.

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 $a < b$,

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

As $n \to \infty$.

Because the preceding theorem is only a special case of the central limit theorem, which is presented in Chapter 8, we shall not present a proof.

Note that we now have **two** possible approximations to binomial probabilities: the Poisson approximation, which is good when n is large and p is small, and the normal approximation, which can be shown to be quite good when $np(1-p)$ is large. (See Figure 5.4.) [The normal approximation will, in general, be quite good for values of n satisfying $np(1-p) \ge 10$.]

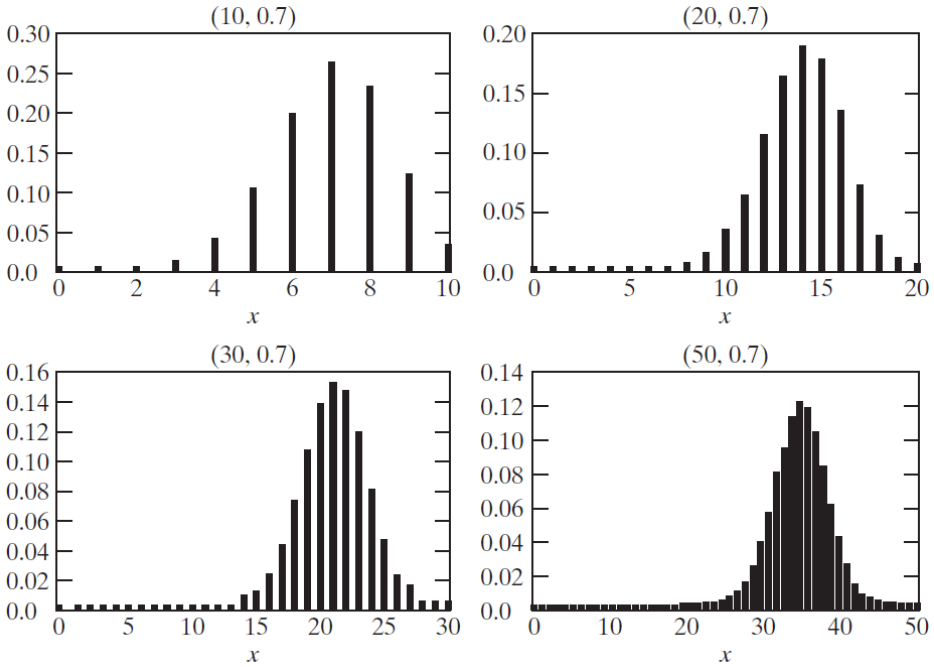


Figure 5. 4 The probability mass function of a binomial (n, p) random variable becomes more and more "normal" as n becomes larger and larger

Example 4g: Let X be the number of times that a fair coin that is flipped 40 times lands on heads. Find the probability that $X = 20$. Use the normal approximation and then compare it with the exact solution.

Solution To employ the normal approximation, note that because the binomial is a discrete integer-valued random variable, whereas the normal is a continuous random variable, it is best to write $P\{X = i\}$ as $P\{i - 1/2 < X < i + 1/2\}$ before applying the normal approximation (this is called the *continuity correction*). Doing so gives

$$P\{X = 20\} = P\{19.5 < X < 20.5\}$$

$$= P\left\{\frac{19.5 - 20}{\sqrt{10}} < \frac{X - 20}{\sqrt{10}} < \frac{20.5 - 20}{\sqrt{10}}\right\}$$

$$\approx P\left\{-0.16 < \frac{X-20}{\sqrt{10}} < 0.16\right\}$$
$$\approx \Phi(0.16) - \Phi(-0.16) \approx 0.1272$$

The exact result is

$$P\{X = 20\} = \binom{40}{20}\left(\frac{1}{2}\right)^{40} \approx 0.1254.$$

5.5 Exponential Random Variables

A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

is said to be an ***exponential*** random variable (or, more simply, is said to be exponentially distributed) with parameter λ. The cumulative distribution function $F(a)$ of an exponential random variable is given by

$$F(a) = P\{X \leq a\}$$
$$= 1 - e^{-\lambda a} \quad a \geq 0$$

Note that $F(\infty) = \int_0^\infty \lambda e^{-\lambda x}\, dx = 1$, as, of course, it must.

Example 5a: calculate (a) $E[X]$ and (b) $\mathrm{Var}(X)$:

$$E[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x}\, dx$$

Integrating by parts (with $\lambda e^{-\lambda x} = dv$ and $u = x^n$) yields

$$E[X^n] = \frac{n}{\lambda} E[X^{n-1}]$$

Letting $n = 1$ and then $n = 2$ gives

$$E[X] = \frac{1}{\lambda}$$

$$E[X^2] = \frac{2}{\lambda} E[X] = \frac{2}{\lambda^2}$$

Hence,

$$\mathrm{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Thus, the mean of the exponential is the reciprocal of its parameter λ, and the variance is the mean squared.

In practice, the exponential distribution often arises as the distribution of the amount of time until some specific event occurs. For instance, the amount of time (starting from now) until an earthquake occurs, or until a new war breaks out, or until a telephone call you receive turns out to be a wrong number are all random variables that tend in practice to have exponential distributions.

We say that a nonnegative random variable X is ***memoryless*** if

$$P\{X > s + t | X > t\} = P\{X > s\} \quad \text{for all } s, t \geq 0$$

If we think of X as being the lifetime of some instrument, the above equation states that the probability that the instrument survives for at least $s + t$ hours, given that it **has survived** t hours, is the same as the initial probability that it survives for at least s hours. In other words, if the instrument is alive at age t, the distribution of the remaining amount of time that it survives is the same as the original lifetime distribution. (That is, it is as if the instrument does

not "remember" that it has already been in use for a time t.)

The above equation is equivalent to

$$\frac{P\{X > s+t, X > t\}}{P\{X > t\}} = P\{X > s\}$$

or

$$P\{X > s+t\} = P\{X > s\}P\{X > t\}$$

Since the above equation is satisfied when X is exponentially distributed (for $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}$), it follows that exponentially distributed random variables are memoryless.

It turns out that not only is the exponential distribution memoryless, but it is also the unique distribution possessing this property. (Page 199)

A variation of the exponential distribution is the distribution of a random variable that is equally likely to be either positive or negative and whose absolute value is exponentially distributed with parameter λ, $\lambda \geq 0$. Such a random variable is said to have a ***Laplace*** distribution, and its density is given by

$$f(x) = \frac{1}{2}\lambda e^{-\lambda|x|} \quad -\infty < x < \infty$$

Its distribution function is given by

$$F(x) = \begin{cases} \frac{1}{2}\int_{-\infty}^{x} \lambda e^{\lambda y}\, dy & x < 0 \\ \frac{1}{2}\int_{-\infty}^{0} \lambda e^{\lambda y}\, dy + \frac{1}{2}\int_{0}^{x} \lambda e^{-\lambda y}\, dy & x > 0 \end{cases}$$

$$= \begin{cases} \frac{1}{2}e^{\lambda x} & x < 0 \\ 1 - \frac{1}{2}e^{-\lambda x} & x > 0 \end{cases}$$

● Hazard Rate Functions

Consider a positive continuous random variable X that we interpret as being the lifetime of some item. Let X have distribution function F and density f. The *hazard rate* (sometimes called the *failure rate*) function $\lambda(t)$ of F is defined by

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)}, \quad \text{where } \bar{F} = 1 - F$$

To interpret $\lambda(t)$, suppose that the item has survived for a time t and we desire the probability that it will not survive for an additional time dt. That is, consider $P\{X \in (t, t+dt) | X > t\}$. Now,

$$P\{X \in (t, t+dt) | X > t\} = \frac{P\{X \in (t, t+dt), X > t\}}{P\{X > t\}}$$

$$= \frac{P\{X \in (t, t+dt)\}}{P\{X > t\}}$$

$$\approx \frac{f(t)}{\bar{F}(t)}dt$$

Thus, $\lambda(t)$ represents the conditional probability intensity that a t-unit-old item will fail.

Suppose now that the lifetime distribution is exponential. Then, by the memoryless property, it follows that the distribution of remaining life for a t-year-old item is the same as that for a new item. Hence, $\lambda(t)$ should be constant. In fact, this checks out, since

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)}$$

$$= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}$$

$$= \lambda$$

Thus, the failure rate function for the exponential distribution is constant. The parameter λ is often referred to as the *rate* of the distribution.

It turns out that the failure rate function $\lambda(s), s \geq 0$, uniquely determines the distribution function F. To prove this, we integrate $\lambda(s)$ from 0 to t to obtain

$$\int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1-F(s)}ds$$

$$= -\log\big(1-F(s)\big)\Big|_0^t$$

$$= -\log\big(1-F(t)\big)$$

Solving the preceding equation for $F(t)$ gives

$$F(t) = 1 - \exp\left\{-\int_0^t \lambda(s)ds\right\}$$

Hence, a distribution function of a positive continuous random variable can be specified by giving its hazard rate function. For instance, if a random variable has a linear hazard rate function—that is, if

$$\lambda(t) = a + bt$$

then its distribution function is given by

$$F(t) = 1 - e^{-at-bt^2/2}$$

and differentiation yields its density, namely,

$$f(t) = (a+bt)e^{-(at+bt^2/2)} \quad t \geq 0$$

When $a = 0$, the preceding equation is known as the *Rayleigh density function*.

Example 5f: One often hears that the death rate of a person who smokes is, at each age, twice that of a nonsmoker. What does this mean? (Page 202)

5.6 Other Continuous Distributions

● The 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

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

where $\Gamma(\alpha)$, called the *gamma function*, is defined as

$$\Gamma(\alpha) = \int_0^\infty e^{-y}y^{\alpha-1}\,dy$$

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

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$$

For integral values of α, say, $\alpha = n$, we obtain, by applying the above equation repeatedly,

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

$$= (n-1)(n-2)\cdots 3\cdot 2\Gamma(1)$$

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 have occurred. More specifically, if events are occurring randomly and in accordance with the three axioms of Section 4.7, then it turns out that the amount of time one has to wait until a total of n events has occurred will be a gamma random variable with parameters (n, λ). Note that when $n = 1$, this distribution reduces to the exponential distribution.

The gamma distribution with $\lambda = \frac{1}{2}$ and $\alpha = n/2$, n a positive integer, is called the χ^2_n (read "chi-squared") distribution with n degrees of freedom. The chi-squared distribution often arises in practice as the distribution of the error involved in attempting to hit a target in n-dimensional space when each coordinate error is normally distributed.

Example 6a: Calculate (a) $E[X]$ and (b) $\mathrm{Var}(X)$.

$$E[X] = \frac{1}{\Gamma(\alpha)} \int_0^\infty \lambda x e^{-\lambda x} (\lambda x)^{\alpha-1}\, dx$$

$$= \frac{1}{\lambda \Gamma(\alpha)} \int_0^\infty \lambda e^{-\lambda x} (\lambda x)^{\alpha}\, dx$$

$$= \frac{\Gamma(\alpha+1)}{\lambda \Gamma(\alpha)}$$

$$= \frac{\alpha}{\lambda}$$

$$E[X^2] = \frac{(\alpha+1)\alpha}{\lambda^2}$$

$$\mathrm{Var}(X) = \frac{\alpha}{\lambda^2}$$

● The Weibull Distribution

The Weibull distribution is widely used in engineering practice due to its versatility. It was originally proposed for the interpretation of fatigue data, but now its use has been extended to many other engineering problems. In particular, it is widely used in the field of life phenomena as the distribution of the lifetime of some object, especially when the "weakest link" model is appropriate for the object. That is, consider an object consisting of many parts, and suppose that the object experiences death (failure) when any of its parts fails. It has been shown (both theoretically and empirically) that under these conditions, a Weibull distribution provides a close approximation to the distribution of the lifetime of the item.

The Weibull distribution function has the form

$$F(x) = \begin{cases} 0 & x \le v \\ 1 - \exp\left\{-\left(\frac{x-v}{\alpha}\right)^\beta\right\} & x > v \end{cases}$$

A random variable whose cumulative distribution function is given by the above equation is said to be *a Weibull random variable* with parameters v, α, and β. Differentiation yields the density:

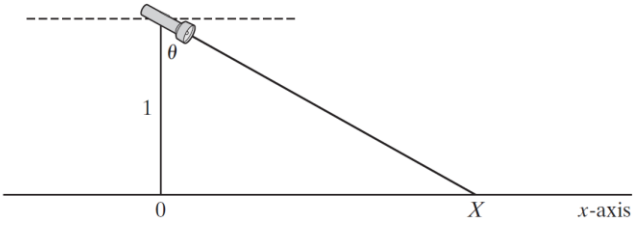
$$f(x) = \begin{cases} 0 & x \le v \\ \frac{\beta}{\alpha}\left(\frac{x-v}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{x-v}{\alpha}\right)^\beta\right\} & x > v \end{cases}$$

● The Cauchy Distribution

A random variable is said to have a Cauchy distribution with parameter θ, $-\infty < \theta < \infty$, if its density is given by

$$f(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2} \quad -\infty < x < \infty$$

Example 6b: Suppose that a narrow-beam flashlight is spun around its center, which is located a unit distance from the x-axis. (Page 206)



● The Beta Distribution

A random variable is said to have a beta distribution if its density is given by

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

The beta distribution can be used to model a random phenomenon whose set of possible values is some finite interval $[c,d]$-which, by letting c denote the origin and taking $d-c$ as a unit measurement, can be transformed into the interval $[0,1]$.

When $a = b$, the beta density is symmetric about $\frac{1}{2}$, giving more and more weight to regions about $\frac{1}{2}$ as the common value a increases. When $a = b = 1$, the beta distribution reduces to the uniform $(0,1)$ distribution. When $b > a$, the density is skewed to the left (in the sense that smaller values become more likely), and it is skewed to the right when $a > b$.

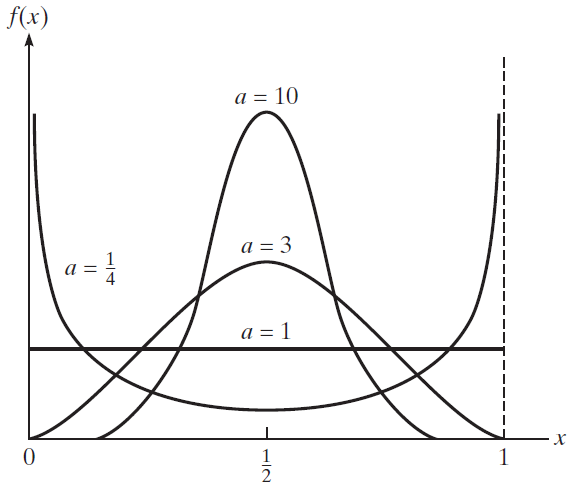


Figure 5. 5 Beta densities with parameters (a,b) when $a = b$

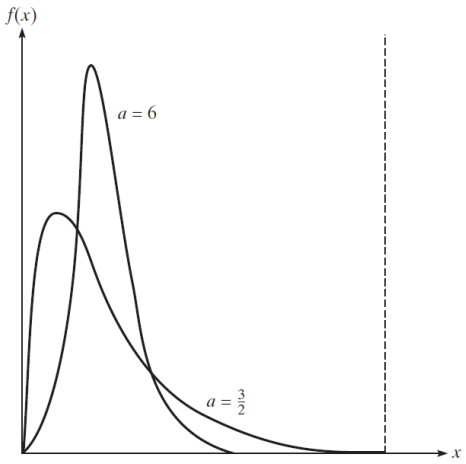


Figure 5. 6 Beta densities with parameters (a,b) when $a/(a+b) = 1/20$

The relationship

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

will be verified in Example 7c of Chapter 6.

it is an easy matter to show that if X is a beta random variable with parameters a and b, then

$$E[X] = \frac{a}{a+b}$$

$$\mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

5.7 The Distribution of a Function of a Random Variable

Often, we know the probability distribution of a random variable and are interested in determining the **distribution** of some function of it. For instance, suppose that we know the distribution of X and want to find the distribution of $g(X)$. To do so, it is necessary to express the event that $g(X) \le y$ in terms of X being in some set.

Addition:

In discrete case, for $y = g(x)$, then

$$p(y) = \sum_{x:y=g(x)} p(x).$$

Example 7a: Let X be uniformly distributed over $(0,1)$. We obtain the distribution of the random variable Y, defined by $Y = X^n$, as follows: For $0 \le y \le 1$,

$$\begin{aligned} F_Y(y) &= P\{Y \le y\} \\ &= P\{X^n \le y\} \\ &= P\{X \le y^{1/n}\} \\ &= F_X(y^{1/n}) \\ &= y^{1/n} \end{aligned}$$

For instance, the density function of Y is given by

$$f_Y(y) = \begin{cases} \frac{1}{n} y^{1/n-1} & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Example 7b: If X is a continuous random variable with probability density f_X, then the distribution of $Y = X^2$ is obtained as follows: For $y \ge 0$,

$$
\begin{aligned}
F_Y(y) &= P\{Y \le y\} \\
&= P\{X^2 \le y\} \\
&= P\{-\sqrt{y} \le X \le \sqrt{y}\} \\
&= F_X(\sqrt{y}) - F_X(-\sqrt{y})
\end{aligned}
$$

Differentiation yields

$$
f_Y(y) = \frac{1}{2\sqrt{y}}[f_X(\sqrt{y}) + f_X(-\sqrt{y})].
$$

Theorem 7.1: Let X be a continuous random variable having probability density function f_X. Suppose that $g(x)$ is a strictly monotonic (increasing or decreasing), differentiable (and thus continuous) function of x. Then the random variable Y defined by $Y = g(X)$ has a probability density function given by

$$
f_Y(y) = \begin{cases} f_X[g^{-1}(y)]\left|\frac{d}{dy} g^{-1}(y)\right| & \text{if } y = g(x) \text{ for some } x \\ 0 & \text{if } y \neq g(x) \text{ for all } x \end{cases}
$$

where $g^{-1}(y)$ is defined to equal that value of x such that $g(x) = y$.

Example 7e: **The Lognormal Distribution** If X is a normal random variable with mean μ and variance σ^2, then the random variable

$$
Y = e^X
$$

is said to be a *lognormal* random variable with parameters μ and σ^2. Thus, a random variable Y is lognormal if $\log(Y)$ is a normal random variable. The lognormal is often used as the distribution of the ratio of the price of a security at the end of one day to its price at the end of the prior day. (Page 210)

Theoretical Exercises

5.2. Show that

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

Hint: Show that

$$
\int_0^\infty P\{Y < -y\}dy = -\int_{-\infty}^0 x f_Y(x)dx
$$

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

$$
\int_0^\infty P\{Y < -y\}dy = \int_0^\infty \int_{-\infty}^{-y} f_Y(x)dxdy
$$

$$
= \int_{-\infty}^0 \int_0^{-x} f_Y(x)dydx = -\int_{-\infty}^0 x f_Y(x)dx
$$

Similarly,

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

Subtracting these equalities gives the result.

5.3. Show that if X has density function f, then

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

Hint: Using Theoretical Exercise 5.2, start with

$$E[g(X)] = \int_0^\infty P\{g(X) > y\}dy - \int_0^\infty P\{g(X) < -y\}dy$$

and then proceed as in the proof given in the text when $g(X) \geq 0$.

$$E[g(X)] = \int_0^\infty P\{g(X) > y\}dy - \int_0^\infty P\{g(X) < -y\}dy$$

$$= \int_0^\infty \int_{x:g(x)>y} f(x)dxdy - \int_0^\infty \int_{x:g(x)<-y} f(x)dxdy$$

by changing the integration order, we reach

$$E[g(X)] = \int_{x:g(x)>0} \int_0^{g(x)} dyf(x)dx - \int_{x:g(x)<0} \int_0^{-g(x)} dyf(x)dx$$

$$= \int_{x:g(x)>0} g(x)f(x)dx - \int_{x:g(x)<0} -g(x)f(x)dx$$

$$= \int_{x:g(x)>0} g(x)f(x)dx + \int_{x:g(x)<0} g(x)f(x)dx$$

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

5.18. Verify that the gamma density function integrates to 1.

Suppose that $X \sim \text{Gamma}(\alpha, \lambda)$. The density function is

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

We want to prove that

$$\int_R f(x)dx = 1$$

We have that

$$\int_R f(x)dx = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\lambda x}dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1}e^{-\lambda x}dx$$

In order to evaluate the integral, make substitution $s = \lambda x$. That yields $x = \frac{s}{\lambda}$ and $ds = \lambda dx$.

We have that

$$\int_0^\infty x^{\alpha-1}e^{-\lambda x}dx = \int_0^\infty \left(\frac{s}{\lambda}\right)^{\alpha-1} e^{-s}\frac{ds}{\lambda} = \frac{1}{\lambda^\alpha}\int_0^\infty s^{\alpha-1}e^{-s}ds = \frac{\Gamma(\alpha)}{\lambda^\alpha}$$

If we plug that back in $\int_R f(x)dx$, we end up with

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1}e^{-\lambda x}dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha)}{\lambda^\alpha} = 1$$

Hence, we have proved the claimed.

5.19. If X is an exponential random variable with mean $1/\lambda$, show that

$$E[X^k] = \frac{k!}{\lambda^k}, \quad k = 1,2,\cdots$$

Hint: Make use of the gamma density function to evaluate the preceding.

$$E[X^k] = \int_0^\infty x^k \lambda e^{-\lambda x}dx = \lambda^{-k}\int_0^\infty \lambda e^{-\lambda x}(\lambda x)^k dx$$

$$= \lambda^{-k}\Gamma(k+1) = k!/\lambda^k.$$

5.21. Show that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Hint: $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-x}x^{-1/2}dx$. Make the change of variables $y = \sqrt{2x}$ and then relate the resulting expression to the normal distribution.

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-x}x^{-1/2}dx$$

$$= \sqrt{2}\int_0^\infty e^{-\frac{y^2}{2}}dy \quad \left(\text{by } y = \sqrt{2x}, dy = \frac{\sqrt{2}}{2}x^{-1/2}dx\right)$$

$$= \sqrt{2}\frac{\sqrt{2\pi}}{2}$$

$$= \sqrt{\pi}.$$

CHAPTER 6: Jointly Distributed Random Variables

6.1 Joint Distribution Functions

Thus far, 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

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

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

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

Note that in the preceding set of equalities, we have once again made use of the fact that probability is a continuous set (that is, event) function. Similarly, the cumulative distribution function of Y is given by

$$\begin{aligned} F_Y(b) &= P\{Y \le b\} \\ &= \lim_{a\to\infty} F(a,b) \\ &\equiv F(\infty,b) \end{aligned}$$

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. For instance, suppose we wanted to compute the joint probability that X is greater than a and Y is greater than b. This could be done as follows:

$$\begin{aligned} P\{X > a, Y > b\} &= 1 - P(\{X > a, Y > b\}^c) \\ &= 1 - P(\{X > a\}^c \cup \{Y > b\}^c) \\ &= 1 - P(\{X \le a\} \cup \{Y \le b\}) \\ &= 1 - [P\{X \le a\} + P\{Y \le b\} - P\{X \le a, Y \le b\}] \\ &= 1 - F_X(a) - F_Y(b) + F(a,b) \end{aligned}$$

The above equation is a special case of the following equation, whose verification is left as an exercise:

$$\begin{aligned} &P\{a_1 < X \le a_2, b_1 < Y \le b_2\} \\ &= F(a_2,b_2) + F(a_1,b_1) - F(a_1,b_2) - F(a_2,b_1) \end{aligned}$$

whenever $a_1 < a_2$, $b_1 < b_2$. (Hint: the event $\{a_1 < X \le a_2, b_1 < Y \le b_2\}$ can be written as the difference of two event: $\{a_1 < X \le a_2, b_1 < Y \le b_2\} = \{X \le a_2, Y \le b_2\} - (\{X \le a_1, Y \le b_2\} \cup \{X \le a_2, Y \le b_1\})$)

For remembering: $P\{a_1 < X \le a_2, b_1 < Y \le b_2\} = \big(F_X(a_2) - F_X(a_1)\big)\big(F_Y(b_2) - F_Y(b_1)\big)$, but it's not correct.

In the case when X and Y are both **discrete** random variables, it is convenient to define the ***joint*** *probability* ***mass*** *function* of X and Y by

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

The probability mass function of X can be obtained from $p(x,y)$ by

$$p_X(x) = P\{X = x\}$$
$$= \sum_{y:p(x,y)>0} p(x,y)$$

Similarly,

$$p_Y(y) = \sum_{x:p(x,y)>0} p(x,y)$$

Example 1a: Suppose that 3 balls are randomly selected from an urn containing 3 red, 4 white, and 5 blue balls. If we let X and Y denote, respectively, the number of red and white balls chosen, then the joint probability mass function of X and Y, $p(i,j) = P\{X = i, Y = j\}$, is given by

$$p(0,0) = \binom{5}{3} / \binom{12}{3} = \frac{10}{220}$$

$$p(0,1) = \frac{40}{220}$$

$$p(0,2) = \frac{30}{220}$$

$$p(0,3) = \frac{4}{220}$$

$$p(1,0) = \frac{30}{220}$$

$$p(1,1) = \frac{60}{220}$$

$$p(1,2) = \frac{18}{220}$$

$$p(2,0) = \frac{15}{220}$$

$$p(2,1) = \frac{12}{220}$$

$$p(3,0) = \frac{1}{220}$$

These probabilities can most easily be expressed in tabular form, as in Table 6.1.

Table 6.1 $P\{X = i, Y = j\}$.

i \ j	0	1	2	3	Row sum $= P\{X = i\}$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
Column sum $= P\{Y = j\}$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

Because the individual probability mass functions of X and Y thus appear in the margin of

such a table, they are often referred to as the *marginal* probability mass functions of X and Y, respectively.

We say that X and Y are ***jointly continuous*** if there exists a function $f(x,y)$, defined for all real x and y, having the property that for every set C of pairs of real numbers (that is, C is a set in the two-dimensional plane),

$$P\{(X,Y) \in C\} = \iint_{(x,y)\in C} f(x,y)\,dx\,dy$$

The function $f(x,y)$ is called the ***joint*** *probability* ***density*** *function* of X and Y. If A and B are any sets of real numbers, then by defining $C = \{(x,y): x \in A, y \in B\}$, we see from the above equation that

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x,y)\,dx\,dy$$

Because

$$\begin{aligned} F(a,b) &= P\{X \in (-\infty,a], Y \in (-\infty,b]\} \\ &= \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)\,dx\,dy \end{aligned}$$

it follows, upon differentiation, that

$$f(a,b) = \frac{\partial^2}{\partial a \partial b} F(a,b)$$

wherever the partial derivatives are defined. Another interpretation of the joint density function, obtained from $P\{X \in A, Y \in B\} = \int_B \int_A f(x,y)\,dx\,dy$, is

$$\begin{aligned} P\{a < X < a + da, b < Y < b + db\} &= \int_b^{b+db}\int_a^{a+da} f(x,y)\,dx\,dy \\ &\approx f(a,b)dadb \end{aligned}$$

when da and db are small and $f(x,y)$ is continuous at a, b. Hence, $f(a,b)$ is a measure of how likely it is that the random vector (X,Y) will be near (a,b).

If X and Y are jointly continuous, they are individually continuous, and their probability density functions can be obtained as follows:

$$\begin{aligned} P\{X \in A\} &= P\{X \in A, Y \in (-\infty,\infty)\} \\ &= \int_A \int_{-\infty}^{\infty} f(x,y)\,dy\,dx \\ &= \int_A f_X(x)\,dx \end{aligned}$$

where

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

is thus the probability density function of X. Similarly, the probability density function of Y is given by

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

Example 1d: Consider a circle of radius R, and suppose that a point within the circle is randomly chosen in such a manner that all regions within the circle of equal area are equally likely to

contain the point. (Page 225)

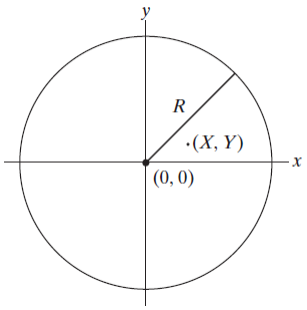


Figure 6. 1 Joint probability distribution

We can also define joint probability distributions for n random variables in the same manner as we did for $n = 2$. For instance, the joint cumulative probability distribution function $F(a_1, a_2, \cdots, a_n)$ of the n random variables $X_1, X_2, \cdots, X_n$ is defined by

$$F(a_1, a_2, \cdots, a_n) = P\{X_1 \le a_1, X_2 \le a_2, \cdots, X_n \le a_n\}$$

Further, the n random variables are said to be *jointly continuous* if there exists a function $f(x_1, x_2, \cdots, x_n)$, called the *joint probability density function*, such that, for any set C in n-space,

$$P\{(X_1, X_2, \cdots, X_n) \in C\} = \iint \cdots \int_{(x_1, \cdots, x_n) \in C} f(x_1, \cdots, x_n) dx_1 dx_2 \cdots dx_n$$

For any n sets of real numbers $A_1, A_2, \cdots, A_n$,

$$P\{X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n\}$$

$$= \int_{A_n} \int_{A_{n-1}} \cdots \int_{A_1} f(x_1, \cdots, x_n) dx_1 dx_2 \cdots dx_n$$

Example 1f: The multinomial distribution (Page 228)

One of the most important joint distributions is the multinomial distribution, which arises when a sequence of n independent and identical experiments is performed. Suppose that each experiment can result in any one of r possible outcomes, with respective probabilities $p_1, p_2, \cdots, p_r, \sum_{i=1}^{r} p_i = 1$. If we let X_i denote the number of the n experiments that result in outcome number i, then

$$P\{X_1 = n_1, X_2 = n_2, \cdots, X_r = n_r\} = \frac{n!}{n_1! n_2! \cdots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$$

whenever $\sum_{i=1}^{r} n_i = n$.

Note that when $r = 2$, the multinomial reduces to the binomial distribution.

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,

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

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.

It can be shown by using the three axioms of probability that the above equation will follow if and only if, for all a, b,

$$P\{X \le a, Y \le b\} = P\{X \le a\} P\{Y \le b\}$$

Hence, in terms of the joint distribution function F of X and Y, X and Y are independent if

$$F(a,b) = F_X(a)F_Y(b) \quad \text{for all } a,b$$

When X and Y are discrete random variables, the condition of independence is equivalent to

$$p(x,y) = p_X(x)p_Y(y) \quad \text{for all } x,y$$

In the jointly continuous case, the condition of independence is equivalent to

$$f(x,y) = f_X(x)f_Y(y) \quad \text{for all } x,y$$

Thus, loosely speaking, X and Y are independent if knowing the value of one does not change the distribution of the other. Random variables that are not independent are said to be *dependent*.

Example 2b: Suppose that the number of people who enter a post office on a given day is a Poisson random variable with parameter λ. (Page 229)

Example 2d: Buffon's needle problem. (Page 231)

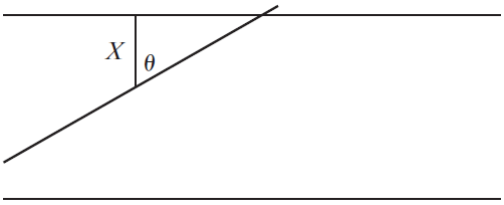


Figure 6. 2

Example 2e: Characterization of the normal distribution. (Page 232)

Proposition 2.1: The continuous (discrete) random variables X and Y are independent if and only if their joint probability density (mass) function can be expressed as

$$f_{X,Y}(x,y) = h(x)g(y) \quad -\infty < x < \infty, -\infty < y < \infty$$

The concept of independence may, of course, be defined for more than two random variables. In general, the n random variables $X_1, X_2, \cdots, X_n$ are said to be independent if, for all sets of real numbers $A_1, A_2, \cdots, A_n$,

$$P\{X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n\} = \prod_{i=1}^{n} P\{X_i \in A_i\}$$

As before, it can be shown that this condition is equivalent to

$$P\{X_1 \le a_1, X_2 \le a_2, \cdots, X_n \le a_n\} = \prod_{i=1}^{n} P\{X_i \le a_i\} \quad \text{for all } a_1, a_2, \cdots, a_n$$

Example 2g: How can a computer choose a random subset? (Page 234)

Remark The foregoing method for generating a random subset has a very low memory requirement. A faster algorithm that requires somewhat more memory is presented in Section 10.1. (The latter algorithm uses the last k elements of a random permutation of $1,2,\cdots,n$.)

Example 2i: Probabilistic interpretation of half-life. (Page 236)

Remark *Independence is a symmetric relation*. As a result, in considering whether X is independent of Y in situations where it is not at all intuitive that knowing the value of Y will not change the probabilities concerning X, it can be beneficial to interchange the roles of X and Y and ask instead whether Y is independent of X.

6.3 Sums of Independent Random Variables

It is often important to be able to calculate the distribution of $X + Y$ from the distributions of X and Y when X and Y are independent. Suppose that X and Y are independent, **continuous** random variables having probability density functions f_X and f_Y. The cumulative distribution function of $X + Y$ is obtained as follows:

$$F_{X+Y}(a) = P\{X + Y \le a\}$$

$$= \iint_{x+y\le a} f_X(x)f_Y(y)\,dx\,dy$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)\,dx\,f_Y(y)dy$$
$$= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy$$

The cumulative distribution function F_{X+Y} is called the **convolution** of the distributions F_X and F_Y (the cumulative distribution functions of X and Y, respectively).

By differentiating the above equation, we find that the probability density function f_{X+Y} of $X+Y$ is given by

$$f_{X+Y}(a) = \frac{d}{da}\int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy$$
$$= \int_{-\infty}^{\infty} \frac{d}{da}F_X(a-y)f_Y(y)dy$$
$$= \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$$

● Identically Distributed Uniform Random Variables

It is not difficult to determine the density function of the sum of two independent uniform $(0,1)$ random variables.

Example 3a: **Sum of two independent uniform random variables**

If X and Y are independent random variables, both uniformly distributed on $(0,1)$, calculate the probability density of $X+Y$.

Solution From the equation $f_{X+Y}(a)$ above, since

$$f_X(a) = f_Y(a) = \begin{cases} 1 & 0 < a < 1 \\ 0 & \text{otherwise} \end{cases}$$

we obtain

$$f_{X+Y}(a) = \int_0^1 f_X(a-y)dy$$

(or using a graph) For $0 < a-y < 1$, we get $y < a < 1+y$. So when $0 < y < a$ and $0 \le a \le 1$, this yields

$$f_{X+Y}(a) = \int_0^a dy = a$$

When $a-1 < y < 1$ and $0 \le a-1 < 1 \to 1 \le a < 2$, we get

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2-a$$

Hence,

$$f_{X+Y}(a) = \begin{cases} a & 0 \le a \le 1 \\ 2-a & 1 < a < 2 \\ 0 & \text{otherwise} \end{cases}$$

Because of the shape of its density function (see Figure 6.3), the random variable $X+Y$ is said to have a *triangular* distribution.

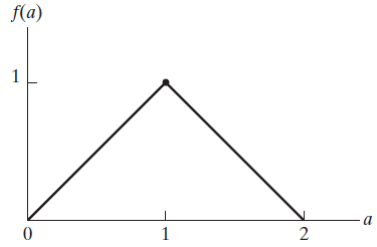


Figure 6. 3 Triangular density function.

Now, suppose that $X_1, X_2, \cdots, X_n$ are independent uniform $(0,1)$ random variables, and let

$$F_n(x) = P\{X_1 + \cdots + X_n \le x\}$$

Whereas a general formula for $F_n(x)$ is messy, it has a particularly nice form when $x \le 1$.

Indeed,

$$F_n(x) = x^n/n!, \quad 0 \le x \le 1$$

For an interesting application of the preceding formula, let us use it to determine the expected number of independent uniform $(0,1)$ random variables that need to be summed to exceed 1. (Page 241)

● Gamma Random Variables

Recall that a gamma random variable has a density of the form

$$f(y) = \frac{\lambda e^{-\lambda y}(\lambda y)^{t-1}}{\Gamma(t)} \quad 0 < y < \infty$$

An important property of this family of distributions is that for a fixed value of λ, it is closed under convolutions.

Proposition 3.1: If X and Y are independent gamma random variables with respective parameters (s,λ) and (t,λ), then $X+Y$ is a gamma random variable with parameters $(s+t,\lambda)$.

$$f_{X+Y}(a) = \frac{\lambda e^{-\lambda a}(\lambda a)^{s+t-1}}{\Gamma(s+t)}$$

It is now a simple matter to establish, by using Proposition 3.1 and induction, that if $X_i, i = 1,\cdots,n$ are independent gamma random variables with respective parameters (t_i,λ), $i = 1,\cdots,n$, then $\sum_{i=1}^n X_i$ is gamma with parameters $(\sum_{i=1}^n t_i, \lambda)$.

Example 3b: Let $X_1, X_2, \cdots, X_n$ be n independent exponential random variables, each having parameter λ. Then, since an exponential random variable with parameter λ is the same as a gamma random variable with parameters $(1,\lambda)$, it follows from Proposition 3.1 that $X_1 + X_2 + \cdots + X_n$ is a gamma random variable with parameters (n,λ).

If $Z_1, Z_2, \cdots, Z_n$ are independent standard normal random variables, then $Y \equiv \sum_{i=1}^n Z_i^2$ is said to have the ***chi-squared*** (sometimes seen as χ^2) distribution with n degrees of freedom. Let us compute the density function of Y. When $n = 1$, $Y = Z_1^2$, and from Example 7b of Chapter 5, we see that its probability density function is given by

$$\begin{aligned} f_{Z^2}(y) &= \frac{1}{2\sqrt{y}}\left[f_Z(\sqrt{y}) + f_Z(-\sqrt{y})\right] \\ &= \frac{1}{2\sqrt{y}}\frac{2}{\sqrt{2\pi}}e^{-y/2} \\ &= \frac{\frac{1}{2}e^{-y/2}(y/2)^{1/2-1}}{\sqrt{\pi}} \end{aligned}$$

But we recognize the preceding as the gamma distribution with parameters $\left(\frac{1}{2},\frac{1}{2}\right)$. [A by-product of this analysis is that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.] But since each Z_i^2 is gamma $\left(\frac{1}{2},\frac{1}{2}\right)$, it follows from Proposition 3.1 that the chi-squared distribution with n degrees of freedom is just the gamma distribution with parameters $\left(n/2,\frac{1}{2}\right)$ and hence has a probability density function given by

$$\begin{aligned} f_Y(y) &= \frac{\frac{1}{2}e^{-y/2}\left(\frac{y}{2}\right)^{n/2-1}}{\Gamma\left(\frac{n}{2}\right)} \quad y > 0 \\ &= \frac{e^{-y/2}y^{n/2-1}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \quad y > 0 \end{aligned}$$

When n is an even integer, $\Gamma(n/2) = [(n/2)-1]!$, whereas when n is odd, $\Gamma(n/2)$ can be obtained from iterating the relationship $\Gamma(t) = (t-1)\Gamma(t-1)$ and then using the previously obtained result that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. [For instance, $\Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3}{2}\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{3}{4}\sqrt{\pi}$.]

In practice, the chi-squared distribution often arises as the distribution of the square of the error involved when one attempts to hit a target in n-dimensional space when the coordinate errors are taken to be independent standard normal random variables. It is also important in statistical analysis.

● Normal Random Variables

We can also use $f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy$ to prove the following important result about normal random variables.

Proposition 3.2: If $X_i, i = 1, \cdots, n$, are independent random variables that are normally distributed with respective parameters $\mu_i, \sigma_i^2, i = 1, \cdots, n$, then $\sum_{i=1}^{n} X_i$ is normally distributed with parameters $\sum_{i=1}^{n} \mu_i$ and $\sum_{i=1}^{n} \sigma_i^2$.

The random variable Y is said to be a *lognormal* random variable with parameters μ and σ if $\log(Y)$ is a normal random variable with mean μ and variance σ^2. That is, Y is lognormal if it can be expressed as

$$Y = e^X$$

where X is a normal random variable.

● Poisson and Binomial Random Variables

Rather than attempt to derive a general expression for the distribution of $X + Y$ in the discrete case, we shall consider some examples.

Example 3e: **Sums of independent Poisson random variables**

If X and Y are independent Poisson random variables with respective parameters λ_1 and λ_2, compute the distribution of $X + Y$.

Solution Because the event $\{X + Y = n\}$ may be written as the union of the disjoint events $\{X = k, Y = n - k\}$, $0 \le k \le n$, we have

$$\begin{aligned} P\{X+Y=n\} &= \sum_{k=0}^{n} P\{X=k, Y=n-k\} \\ &= \sum_{k=0}^{n} P\{X=k\} P\{Y=n-k\} \\ &= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!\,(n-k)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!\,(n-k)!} \lambda_1^k \lambda_2^{n-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1+\lambda_2)^n \end{aligned}$$

Thus, $X + Y$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Example 3f: **Sums of independent binomial random variables**

Let X and Y be independent binomial random variables with respective parameters (n, p) and (m, p). Calculate the distribution of $X + Y$.

Solution Recalling the interpretation of a binomial random variable, and without any computation at all, we can immediately conclude that $X + Y$ is binomial with parameters $(n + m, p)$. This follows because X represents the number of successes in n independent trials, each of which results in a success with probability p; similarly, Y represents the number of successes in m independent trials, each of which results in a success with probability p. Hence, given that X and Y are assumed independent, it follows that $X + Y$ represents the number of successes in $n + m$ independent trials when each trial has a probability p of resulting in a success. Therefore, $X + Y$ is a binomial random variable with parameters $(n + m, p)$. To check this conclusion analytically, note that

$$\begin{aligned} P\{X+Y=k\} &= \sum_{i=0}^{n} P\{X=i, Y=k-i\} \\ &= \sum_{i=0}^{n} P\{X=i\} P\{Y=k-i\} \end{aligned}$$

$$= \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \binom{m}{k-i} p^{k-i} q^{m-k+i}$$

$$= p^k q^{n+m-k} \sum_{i=0}^{n} \binom{n}{i} \binom{m}{k-i}$$

where $q = 1 - p$, following the application of the combinatorial identity

$$\binom{n+m}{k} = \sum_{i=0}^{n} \binom{n}{i} \binom{m}{k-i}$$

$$P\{X + Y = k\} = p^k q^{n+m-k} \binom{n+m}{k}$$

6.4 Conditional Distributions: Discrete Case

If X and Y are discrete random variables, it is natural to define the conditional probability mass function of X given that $Y = y$, by

$$p_{X|Y}(x|y) = P\{X = x|Y = y\}$$

$$= \frac{P\{X = x, Y = y\}}{P\{Y = y\}}$$

$$= \frac{p(x, y)}{p_Y(y)}$$

for all values of y such that $p_Y(y) > 0$. Similarly, the conditional probability distribution function of X given that $Y = y$ is defined, for all y such that $p_Y(y) > 0$, by

$$F_{X|Y}(x|y) = P\{X \leq x|Y = y\}$$

$$= \sum_{a \leq x} p_{X|Y}(a|y)$$

In other words, the definitions are the same as in the unconditional case, except that everything is now conditional on the event that $Y = y$. If X is independent of Y, then the conditional mass function and the distribution function are the same as the respective unconditional ones. This follows because if X is independent of Y, then

$$p_{X|Y}(x|y) = P\{X = x\}$$

Example 4b: If X and Y are independent Poisson random variables with respective parameters λ_1 and λ_2, calculate the conditional distribution of X given that $X + Y = n$.

$$P\{X = k|X + Y = n\} = \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k}$$

In other words, the conditional distribution of X given that $X + Y = n$ is the binomial distribution with parameters n and $\lambda_1/(\lambda_1 + \lambda_2)$. (Page 249)

6.5 Conditional Distributions: Continuous Case

If X and Y have a joint probability density function $f(x, y)$, then the conditional probability density function of X given that $Y = y$ is defined, for all values of y such that $f_Y(y) > 0$, by

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

To motivate this definition, multiply the left-hand side by dx and the right-hand side by $(dx\, dy)/\, dy$ to obtain

$$f_{X|Y}(x|y)\, dx = \frac{f(x, y)\, dx\, dy}{f_Y(y)\, dy}$$

$$\approx \frac{P\{x \leq X \leq x + dx, y \leq Y \leq y + dy\}}{P\{y \leq Y \leq y + dy\}}$$

$$= P\{x \leq X \leq x + dx\, |y \leq Y \leq y + dy\}$$

In other words, for small values of dx and dy, $f_{X|Y}(x|y)\, dx$ represents the conditional probability that X is between x and $x + dx$ given that Y is between y and $y + dy$.

The use of conditional densities allows us to define conditional probabilities of events associated with one random variable when we are given the value of a second random variable. That is, if X and Y are jointly continuous, then, for any set A,

$$P\{X \in A|Y = y\} = \int_A f_{X|Y}(x|y)\,dx$$

In particular, by letting $A = (-\infty, a)$ we can define the conditional cumulative distribution function of X given that $Y = y$ by

$$F_{X|Y}(a|y) \equiv P\{X \le a|Y = y\} = \int_{-\infty}^{a} f_{X|Y}(x|y)\,dx$$

If X and Y are independent continuous random variables, the conditional density of X given that $Y = y$ is just the unconditional density of X,

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

Example 5c: The **t-distribution**

If Z and Y are independent, with Z having a standard normal distribution and Y having a chi-squared distribution with n degrees of freedom, then the random variable T defined by

$$T = \frac{Z}{\sqrt{Y/n}} = \sqrt{n}\frac{Z}{\sqrt{Y}}$$

is said to have a *t-distribution* with n degrees of freedom. As will be seen in Section 7.8, the t-distribution has important applications in statistical inference. At present, we will content ourselves with computing its density function. This will be accomplished by using the conditional density of T given Y to obtain the joint density function of T and Y, from which we will then obtain the marginal density of T. To begin, note that because of the independence of Z and Y, it follows that the conditional distribution of T given that $Y = y$ is the distribution of $\sqrt{n/y}Z$, which is normal with mean 0 and variance n/y. Hence, the conditional density of T given that $Y = y$ is

$$f_{T|Y}(t|y) = \frac{1}{\sqrt{2\pi n/y}} e^{-t^2 y/2n}, \quad -\infty < t < \infty$$

Using the preceding, along with the following formula for the chi-squared density given in Example 3b of this chapter,

$$f_Y(y) = \frac{e^{-y/2}y^{n/2-1}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \quad y > 0$$

we obtain that the joint density of T,Y is

$$f_{T,Y}(t,y) = f_{T|Y}(t|y)f_Y(y) = \frac{1}{\sqrt{2\pi n}2^{n/2}\Gamma(n/2)} e^{-t^2 y/2n} e^{-y/2} y^{(n-1)/2}$$

$$= \frac{1}{\sqrt{\pi n}2^{(n+1)/2}\Gamma(n/2)} e^{-\frac{t^2+n}{2n}y} y^{(n-1)/2}, \quad y > 0, -\infty < t < \infty$$

Letting $c = \frac{t^2+n}{2n}$, and integrating the preceding over all y, gives

$$f_T(t) = \int_0^\infty f_{T,Y}(t,y)\,dy$$

$$= \frac{1}{\sqrt{\pi n}2^{(n+1)/2}\Gamma(n/2)} \int_0^\infty e^{-cy} y^{(n-1)/2}\,dy$$

$$= \frac{c^{-(n+1)/2}}{\sqrt{\pi n}2^{(n+1)/2}\Gamma(n/2)} \int_0^\infty e^{-x} x^{(n-1)/2}\,dx \quad \text{(by letting } x = cy\text{)}$$

$$= \frac{n^{(n+1)/2}\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}(t^2+n)^{(n+1)/2}\Gamma\left(\frac{n}{2}\right)} \quad \left(\text{because } \frac{1}{c} = \frac{2n}{t^2+n}\right)$$

$$= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < \infty.$$

Example 5d: **The bivariate normal distribution**

One of the most important joint distributions is the bivariate normal distribution. We say that the random variables X,Y have a bivariate normal distribution if, for constants μ_x, μ_y, $\sigma_x > 0$, $\sigma_y > 0$, $-1 < \rho < 1$, their joint density function is given, for all $-\infty < x, y < \infty$, by

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}$$

We now determine the conditional density of X given that $Y=y$. In doing so, we will continually collect all factors that do not depend on x and represent them by the constants C_i. The final constant will then be found by using that $\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,dx=1$. We have

$$\begin{aligned}f_{X|Y}(x|y)&=\frac{f(x,y)}{f_Y(y)}\\&=C_1 f(x,y)\\&=C_2\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\frac{x(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}\\&=C_3\exp\left\{-\frac{1}{2{\sigma_x}^2(1-\rho^2)}\left[x^2-2x\left(\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right)\right]\right\}\\&=C_4\exp\left\{-\frac{1}{2{\sigma_x}^2(1-\rho^2)}\left[x-\left(\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right)\right]^2\right\}\end{aligned}$$

Recognizing the preceding equation as a normal density, we can conclude that given $Y=y$, the random variable X is normally distributed with mean $\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)$ and variance ${\sigma_x}^2(1-\rho^2)$. Also, because the joint density of Y,X is exactly the same as that of X,Y, except that μ_x, σ_x are interchanged with μ_y, σ_y, it similarly follows that the conditional distribution of Y given $X=x$ is the normal distribution with mean $\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$ and variance ${\sigma_y}^2(1-\rho^2)$. It follows from these results that the necessary and sufficient condition for the bivariate normal random variables X and Y to be independent is that $\rho=0$ (a result that also follows directly from their joint density, because it is only when $\rho=0$ that the joint density factors into two terms, one depending only on x and the other only on y).

With $C=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$, the marginal density of X can be obtained from

$$\begin{aligned}f_X(x)&=\int_{-\infty}^{\infty}f(x,y)\,dy\\&=C\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}dy\end{aligned}$$

Making the change of variables $w=\frac{y-\mu_y}{\sigma_y}$ gives

$$\begin{aligned}f_X(x)&=C\sigma_y\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right\}\times\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[w^2-2\rho\frac{(x-\mu_x)}{\sigma_x}w\right]\right\}dw\\&=C\sigma_y\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_x}{\sigma_x}\right)^2(1-\rho^2)\right\}\times\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[w-\rho\frac{x-\mu_x}{\sigma_x}\right]^2\right\}dw\end{aligned}$$

Because

$$\frac{1}{\sqrt{2\pi(1-\rho^2)}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[w-\rho\frac{x-\mu_x}{\sigma_x}\right]^2\right\}dw=1$$

we see that

$$\begin{aligned}f_X(x)&=C\sigma_y\sqrt{2\pi(1-\rho^2)}e^{-(x-\mu_x)^2/2{\sigma_x}^2}\\&=\frac{1}{\sqrt{2\pi}\sigma_x}e^{-(x-\mu_x)^2/2{\sigma_x}^2}\end{aligned}$$

That X is normal with mean μ_x and variance σ_x^2. Similarly, Y is normal with mean μ_y and variance σ_y^2.

We can also talk about conditional distributions when the random variables are neither jointly continuous nor jointly discrete. (Page 255) (Page 251) For example, suppose that X is a continuous random variable having probability density function f and N is a discrete random variable, and consider the conditional distribution of X given that $N=n$. Then

$$f_{X|N}(x|n)=\frac{P\{N=n|X=x\}}{P\{N=n\}}f(x)$$

Example 5e: It states that if the original or *prior* (to the collection of data) distribution of a trial success probability is uniformly distributed over $(0,1)$ [or, equivalently, is beta with parameters $(1,1)$], then the posterior (or conditional) distribution given a total of n successes in $n+m$ trials is beta with parameters $(1+n, 1+m)$. (Page 255)

6.6 Order Statistics

Let $X_1, X_2, \cdots, X_n$ be n **independent** and **identically** distributed (iid) continuous random variables having a common density f and distribution function F. Define

$$X_{(1)} = \text{smallest of } X_1, X_2, \cdots, X_n$$
$$X_{(2)} = \text{second smallest of } X_1, X_2, \cdots, X_n$$
$$\vdots$$
$$X_{(j)} = j\text{th smallest of } X_1, X_2, \cdots, X_n$$
$$\vdots$$
$$X_{(n)} = \text{largest of } X_1, X_2, \cdots, X_n$$

The ordered values $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are known as the ***order statistics*** corresponding to the random variables $X_1, X_2, \cdots, X_n$. In other words, $X_{(1)}, \cdots, X_{(n)}$ are the ordered values of $X_1, \cdots, X_n$.

The **joint** density function of the order statistics is **obtained** by noting that the order statistics $X_{(1)}, \cdots, X_{(n)}$ will take on the values $x_1 \le x_2 \le \cdots \le x_n$ if and only if, for some permutation $(i_1, i_2, \cdots, i_n)$ of $(1,2,\cdots,n)$,

$$X_1 = x_{i_1}, X_2 = x_{i_2}, \cdots, X_n = x_{i_n}$$

Since, for any permutation $(i_1, \cdots, i_n)$ of $(1,2,\cdots,n)$,

$$P\left\{x_{i_1} - \frac{\varepsilon}{2} < X_1 < x_{i_1} + \frac{\varepsilon}{2}, \cdots, x_{i_n} - \frac{\varepsilon}{2} < X_n < x_{i_n} + \frac{\varepsilon}{2}\right\}$$
$$\approx \varepsilon^n f(x_1) \cdots f(x_n)$$

it follows that, for $x_1 < x_2 < \cdots < x_n$,

$$P\left\{x_1 - \frac{\varepsilon}{2} < X_{(1)} < x_1 + \frac{\varepsilon}{2}, \cdots, x_n - \frac{\varepsilon}{2} < X_{(n)} < x_n + \frac{\varepsilon}{2}\right\}$$
$$\approx n!\, \varepsilon^n f(x_1) \cdots f(x_n)$$

Dividing by ε^n and letting $\varepsilon \to 0$ yields

$$f_{X_{(1)},\cdots,X_{(n)}}(x_1, x_2, \cdots, x_n) = n!\, f(x_1) \cdots f(x_n) \quad x_1 < x_2 < \cdots < x_n$$

Example 6a: Along a road 1 mile long are 3 people "distributed at random." Find the probability that no 2 people are less than a distance of d miles apart when $d \le \frac{1}{2}$.

Solution Let us assume that "distributed at random" means that the positions of the 3 people are independent and uniformly distributed over the road. If X_i denotes the position of the ith person, then the desired probability is $P\{X_{(i)} > X_{(i-1)} + d, i = 2,3\}$. Because

$$f_{X_{(1)},X_{(2)},X_{(3)}}(x_1, x_2, x_3) = 3! \quad 0 < x_1 < x_2 < x_3 < 1$$

it follows that

$$P\{X_{(i)} > X_{(i-1)} + d, i = 2,3\} = \iiint_{x_i > x_{i-1}+d} f_{X_{(1)},X_{(2)},X_{(3)}}(x_1, x_2, x_3)\, dx_1 dx_2 dx_3$$
$$= 3! \int_0^{1-2d} \int_{x_1+d}^{1-d} \int_{x_2+d}^{1} dx_3 dx_2 dx_1$$
$$= (1-2d)^3$$

Hint: $x_3 < 1$ and $x_3 - x_2 > d \rightarrow x_2 + d < x_3 < 1$;
$x_2 - x_1 > d$ and $x_3 - x_2 > d \rightarrow x_1 + d < x_2 < 1 - d$;
$0 < x_1$ and $x_2 - x_1 > d \rightarrow 0 < x_1 < 1 - 2d$.

The **density** function of the jth-**order** statistic $X_{(j)}$ can be obtained either by integrating the joint density function $f_{X_{(1)},\cdots,X_{(n)}}(x_1, x_2, \cdots, x_n)$ or by direct reasoning as follows: In order for $X_{(j)}$ to equal x, it is necessary for $j-1$ of the n values $X_1, \cdots, X_n$ to be less than x, $n-j$ of them to be greater than x, and 1 of them to equal x. Now, the probability density that any given set of $j-1$ of the X_i's are less than x, another given set of $n-j$ are all greater than x, and the remaining value is equal to x equals

$$[F(x)]^{j-1}[1 - F(x)]^{n-j} f(x)$$

Hence, since there are

$$\binom{n}{j-1, n-j, 1} = \frac{n!}{(n-j)!\,(j-1)!}$$

different partitions of the n random variables $X_1, \cdots, X_n$ into the preceding three groups, it follows that the density function of $X_{(j)}$ is given by

$$f_{X_{(j)}}(x) = \frac{n!}{(n-j)!\,(j-1)!}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x)$$

Example 6b: When a sample of $2n+1$ random variables (that is, when $2n+1$ independent and identically distributed random variables) is observed, the $(n+1)$ smallest is called the ***sample median***. (Page 258)

The cumulative distribution function of $X_{(j)}$ can be found by integrating $f_{X_{(j)}}(x) = \frac{n!}{(n-j)!(j-1)!}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x)$. That is,

$$F_{X_{(j)}}(y) = \frac{n!}{(n-j)!\,(j-1)!}\int_{-\infty}^{y}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x)\,dx$$

However, $F_{X_{(j)}}(y)$ could also have been derived directly by noting that the jth order statistic is less than or equal to y if and only if there are j or more of the X_i's that are less than or equal to y. Thus, because the number of X_i's that are less than or equal to y is a binomial random variable with parameters $n, p = F(y)$, it follows that

$$F_{X_{(j)}}(y) = P\{X_{(j)} \le y\} = P\{j \text{ or more of the } X_i\text{'s are } \le y\}$$
$$= \sum_{k=j}^{n}\binom{n}{k}[F(y)]^k[1-F(y)]^{n-k}$$

If, in the above two equations, we take F to be the uniform $(0,1)$ distribution [that is, $f(x) = 1, 0 < x < 1$], then we obtain the interesting analytical identity

$$\sum_{k=j}^{n}\binom{n}{k}y^k(1-y)^{n-k} = \frac{n!}{(n-j)!\,(j-1)!}\int_{-\infty}^{y}x^{j-1}(1-x)^{n-j}\,dx \quad 0 \le y \le 1$$

By employing the same type of argument that we used in establishing equation $f_{X_{(j)}}(x) = \frac{n!}{(n-j)!(j-1)!}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x)$, we can show that the joint density function of the order statistics $X_{(i)}$ and $X_{(j)}$ when $i < j$ is

$$f_{X_{(i)},X_{(j)}}(x_i, x_j) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}[F(x_i)]^{i-1}$$
$$\times [F(x_j)-F(x_i)]^{j-i-1}[1-F(x_j)]^{n-j}f(x_i)f(x_j)$$

for all $x_i < x_j$.

Example 6c: Distribution of the range of a random sample (Page 259)

6.7 Joint Probability Distribution of Functions of Random Variables

Let X_1 and X_2 be jointly continuous random variables with joint probability density function f_{X_1,X_2}. It is sometimes necessary to obtain the joint distribution of the random variables Y_1 and Y_2, which arise as functions of X_1 and X_2. Specifically, suppose that $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ for some functions g_1 and g_2.

Assume that the functions g_1 and g_2 satisfy the following conditions:

1. The equations $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$ can be uniquely solved for x_1 and x_2 in terms of y_1 and y_2, with solutions given by, say, $x_1 = h_1(y_1, y_2), x_2 = h_2(y_1, y_2)$.
2. The functions g_1 and g_2 have continuous partial derivatives at all points (x_1, x_2) and are such that the 2×2 determinant

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix} \equiv \frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1} \neq 0$$

at all points (x_1, x_2).

Under these two conditions, it can be shown that the random variables Y_1 and Y_2 are jointly

continuous with joint density function given by

$$f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(x_1,x_2)|J(x_1,x_2)|^{-1}$$

where $x_1 = h_1(y_1,y_2), x_2 = h_2(y_1,y_2)$.

A proof of the above equation would proceed in page 260 along the following lines:

$$P\{Y_1 \le y_1, Y_2 \le y_2\} = \iint\limits_{\substack{(x_1,x_2):\\ g_1(x_1,x_2)\le y_1\\ g_2(x_1,x_2)\le y_2}} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$$

Example 7a: Let X_1 and X_2 be jointly continuous random variables with probability density function f_{X_1,X_2}. Let $Y_1 = X_1 + X_2$, $Y_2 = X_1 - X_2$. Find the joint density function of Y_1 and Y_2 in terms of f_{X_1,X_2}.

Solution Let $g_1(x_1,x_2) = x_1 + x_2$ and $g_2(x_1,x_2) = x_1 - x_2$. Then

$$J(x_1,x_2) = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2$$

Also, since the equations $y_1 = x_1 + x_2$ and $y_2 = x_1 - x_2$ have $x_1 = (y_1 + y_2)/2$, $x_2 = (y_1 - y_2)/2$ as their solution, it follows from $f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(x_1,x_2)|J(x_1,x_2)|^{-1}$ that the desired density is

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2} f_{X_1,X_2}\left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right)$$

For instance, if X_1 and X_2 are independent uniform $(0,1)$ random variables, then

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \frac{1}{2} & 0 \le y_1 + y_2 \le 2, 0 \le y_1 - y_2 \le 2 \\ 0 & \text{otherwise} \end{cases}$$

or if X_1 and X_2 are independent exponential random variables with respective parameters λ_1 and λ_2, then

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \frac{\lambda_1\lambda_2}{2} \exp\left\{-\lambda_1\left(\frac{y_1+y_2}{2}\right) - \lambda_2\left(\frac{y_1-y_2}{2}\right)\right\} & y_1 + y_2 \ge 0, y_1 - y_2 \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

Finally, if X_1 and X_2 are independent standard normal random variables, then

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \frac{1}{4\pi} e^{-[(y_1+y_2)^2/8 + (y_1-y_2)^2/8]} \\ &= \frac{1}{4\pi} e^{-(y_1^2+y_2^2)/4} \\ &= \frac{1}{\sqrt{4\pi}} e^{-y_1^2/4} \frac{1}{\sqrt{4\pi}} e^{-y_2^2/4} \end{aligned}$$

Thus, not only do we obtain (in agreement with Proposition 3.2) that both $X_1 + X_2$ and $X_1 - X_2$ are normal with mean 0 and variance 2, but we also conclude that these two random variables are independent. (In fact, it can be shown that if X_1 and X_2 are independent random variables having a common distribution function F, then $X_1 + X_2$ will be independent of $X_1 - X_2$ if and only if F is a normal distribution function.

Example 7b: Let (X,Y) denote a random point in the plane, and assume that the rectangular coordinates X and Y are independent standard normal random variables. We are interested in the joint distribution of R, Θ, the polar coordinate representation of (x,y).

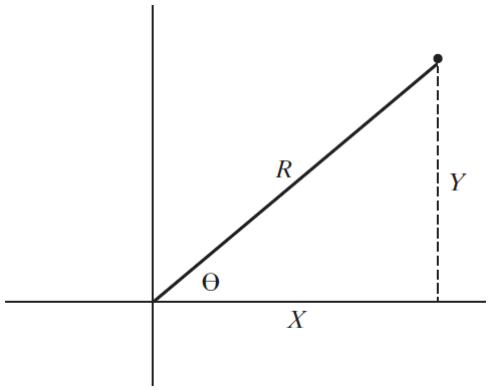


Figure 6. 4 •=Random point. $(X,Y)=(R,\Theta)$.

This result is quite interesting, for it certainly is not evident a priori that a random vector whose coordinates are independent standard normal random variables will have an angle of orientation that not only is uniformly distributed, but also is independent of the vector's distance from the origin. (Page 261)

The preceding result can be used to simulate (or generate) normal random variables by making a suitable transformation on uniform random variables. (Page 263)

Example 7c: This entire result is quite interesting. (Page 263)

When the joint density function of the n random variables $X_1, X_2, \cdots, X_n$ is given and we want to compute the joint density function of $Y_1, Y_2, \cdots, Y_n$, where

$$Y_1 = g_1(X_1, \cdots, X_n) \quad Y_2 = g_2(X_1, \cdots, X_n), \cdots \quad Y_n = g_n(X_1, \cdots, X_n)$$

the approach is the same—namely, we assume that the functions g_i have continuous partial derivatives and that the Jacobian determinant

$$J(x_1, \cdots, x_n) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_n} \end{vmatrix} \neq 0$$

at all points $(x_1, \cdots, x_n)$. Furthermore, we suppose that the equations $y_1 = g_1(x_1, \cdots, x_n), y_2 = g_2(x_1, \cdots, x_n), \cdots, y_n = g_n(x_1, \cdots, x_n)$ have a unique solution, say, $x_1 = h_1(y_1, \cdots, y_n), \cdots, x_n = h_n(y_1, \cdots, y_n)$. Under these assumptions, the joint density function of the random variables Y_i is given by

$$f_{Y_1,\cdots,Y_n}(y_1, \cdots, y_n) = f_{X_1,\cdots,X_n}(x_1, \cdots, x_n)|J(x_1, \cdots, x_n)|^{-1}$$

where $x_i = h_i(y_1, \cdots, y_n), i = 1,2, \cdots, n$.

6.8 Exchangeable Random Variables

The random variables $X_1, X_2, \cdots, X_n$ are said to be *exchangeable* if, for every permutation $i_1, \cdots, i_n$ of the integers $1, \cdots, n$,

$$P\{X_{i_1} \leq x_1, X_{i_2} \leq x_2, \cdots, X_{i_n} \leq x_n\} = P\{X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n\}$$

for all $x_1, \cdots, x_n$. That is, the n random variables are exchangeable if their joint distribution is the same no matter in which order the variables are observed.

Discrete random variables will be exchangeable if

$$P\{X_{i_1} = x_1, X_{i_2} = x_2, \cdots, X_{i_n} = x_n\} = P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$$

for all permutations $i_1, \cdots, i_n$, and all values $x_1, \cdots, x_n$. This is equivalent to stating that $p(x_1, x_2, \cdots, x_n) = P\{X_1 = x_1, \cdots, X_n = x_n\}$ is a symmetric function of the vector $(x_1, \cdots, x_n)$, which means that its value does not change when the values of the vector are permuted.

It is easily seen that if $X_1, X_2, \cdots, X_n$ are exchangeable, then each X_i has the same probability distribution.

CHAPTER 7: Properties of Expectation

7.1 Introduction

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

$$E[X] = \sum_x xp(x)$$

where X is a discrete random variable with probability mass function $p(x)$, and 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

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

then

$$a \le E[X] \le b$$

7.2 Expectation of Sums of Random Variables

For a two-dimensional analog of Propositions 4.1 of Chapter 4 and 2.1 of Chapter 5, which give the computational formulas for the expected value of a function of a random variable, suppose that X and Y are random variables and g is a function of two variables. Then we have the following result.

Proposition 2.1: 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

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

Example 2a: An accident occurs at a point X that is uniformly distributed on a road of length L. At the time of the accident, an ambulance is at a location Y that is also uniformly distributed on the road. Assuming that X and Y are independent, find the expected distance between the ambulance and the point of the accident.

Solution We need to compute $E[|X-Y|]$. Since the joint density function of X and Y is

$$f(x,y) = \frac{1}{L^2}, \quad 0 < x < L, 0 < y < L$$

it follows from Proposition 2.1 that

$$E[|X-Y|] = \frac{1}{L^2}\int_0^L \int_0^L |x-y|\,dy\,dx$$

Then,

$$E[|X-Y|] = \frac{1}{L^2}\left(\int_0^L \int_0^x (x-y)\,dy\,dx + \int_0^L \int_0^y (y-x)\,dx\,dy\right)$$

or

$$E[|X-Y|] = \frac{1}{L^2}\int_0^L \left(\int_0^x (x-y)\,dy + \int_x^L (y-x)\,dy\right)dx$$

Finally,

$$E[|X-Y|] = \frac{L}{3}$$

For an important application of Proposition 2.1, suppose that $E[X]$ and $E[Y]$ are both finite and let $g(X,Y) = X+Y$. Then, in the continuous case,

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

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,dy\,dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty} x f_X(x)\, dx + \int_{-\infty}^{\infty} y f_Y(y)\, dy$$
$$= E[X] + E[Y]$$

The same result holds in general; thus, whenever $E[X]$ and $E[Y]$ are finite,

$$E[X + Y] = E[X] + E[Y]$$

If $E[X_i]$ is finite for all $i = 1, \cdots, n$, then

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$$

The above equation is an extremely useful formula whose utility will now be illustrated by a series of examples.

Example 2c: **The sample mean**

Let $X_1, \cdots, X_n$ 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

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

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 2d: **Boole's inequality**

Let $A_1, \cdots, A_n$ denote events, and define the indicator variables X_i, $i = 1, \cdots, n$. 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,

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

Example 2e: Expectation of a binomial random variable. (Page 284)

Example 2f: Mean of a negative binomial random variable. (Page 284)

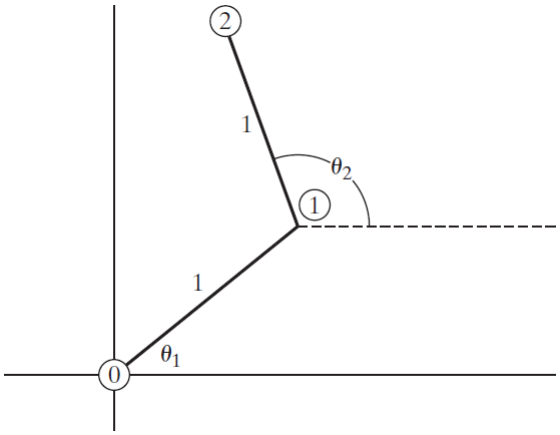
Example 2g: Mean of a hypergeometric random variable. (Page 285)

Example 2h: Expected number of matches. (Page 285)

Example 2i: Coupon-collecting problems. (Page 286) (Need check)

Example 2k: Expected number of runs. (Page 287)

Example 2l: A random walk in the plane. (Page 288)



Example 2m: Analyzing the quick-sort algorithm

Suppose that we are presented with a set of n distinct values $x_1, x_2, \cdots, x_n$ and that we desire to put them in increasing order, or as it is commonly stated, to *sort* them. (Page 289)

Example 2n: The probability of a union of events

Let $A_1, \cdots, A_n$ denote events, and define the indicator variables $X_i, i = 1, \cdots, n$, by

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

Now, note that

$$1 - \prod_{i=1}^{n}(1 - X_i) = \begin{cases} 1 & \text{if } \cup A_i \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Hence, (Page 291)

$$E\left[1 - \prod_{i=1}^{n}(1 - X_i)\right] = P\left(\bigcup_{i=1}^{n} A_i\right)$$

$$= E\left[\sum_{i=1}^{n} X_i - \sum\sum_{i<j} X_i X_j + \sum\sum_{i<j<k}\sum X_i X_j X_k - \cdots + (-1)^{n+1} X_1 \cdots X_n\right]$$

Example 2o: Consider any nonnegative, integer-valued random variable X. If, for each $i \geq 1$, we define

$$X_i = \begin{cases} 1 & \text{if } X \geq i \\ 0 & \text{if } X < i \end{cases}$$

Then

$$\sum_{i=1}^{\infty} X_i = \sum_{i=1}^{X} X_i + \sum_{i=X+1}^{\infty} X_i$$

$$= \sum_{i=1}^{X} 1 + \sum_{i=X+1}^{\infty} 0$$

$$= X$$

Hence, since the X_i are all nonnegative, we obtain

$$E[X] = \sum_{i=1}^{\infty} \mathrm{E}(X_i)$$

$$= \sum_{i=1}^{\infty} P\{X \geq i\}$$

a useful identity.

Example 2p: Showing that ordering the elements in decreasing order of the probability that they are requested minimizes the expected position of the element requested. (Page 293)

● Obtaining Bounds from Expectations via the Probabilistic Method

The probabilistic method is a technique for analyzing the properties of the elements of a set by introducing probabilities on the set and then studying an element chosen according to those

probabilities. The technique was previously seen in Example 4l of Chapter 3, where it was used to show that a set contained an element that satisfied a certain property. In this subsection, we show how it can sometimes be used to bound complicated functions.

Example 2q: The maximum number of Hamiltonian paths in a tournament. (Page 293)

● The Maximum–Minimums Identity

We start with an identity relating the maximum of a set of numbers to the minimums of the subsets of these numbers.

Proposition 2.2: For arbitrary numbers $x_i, i = 1, \cdots, n$,

$$\max_i x_i = \sum_i x_i - \sum_{i<j} \min(x_i, x_j) + \sum_{i<j<k} \min(x_i, x_j, x_k) + \cdots + (-1)^{n+1} \min(x_1, \cdots, x_n)$$

It follows from the above proposition that for any random variables $X_1, \cdots, X_n$,

$$\max_i X_i = \sum_i X_i - \sum_{i<j} \min(X_i, X_j) + \cdots + (-1)^{n+1} \min(X_1, \cdots, X_n)$$

Taking expectations of both sides of this equality yields the following relationship between the expected value of the maximum and those of the partial minimums:

$$E\left[\max_i X_i\right] = \sum_i E[X_i] - \sum_{i<j} E[\min(X_i, X_j)] + \cdots + (-1)^{n+1} E[\min(X_1, \cdots, X_n)]$$

Example 2s: Coupon collecting with unequal probabilities. (Page 297)

7.3 Moments of the Number of Events that Occur

Many of the examples solved in the previous section were of the following form: For given events $A_1, \cdots, A_n$, find $E[X]$, where X is the number of these events that occur. The solution then involved defining an indicator variable I_i for event A_i such that

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

Because

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

we obtained the result

$$E[X] = E\left[\sum_{i=1}^{n} I_i\right] = \sum_{i=1}^{n} E[I_i] = \sum_{i=1}^{n} P(A_i)$$

Now suppose we are interested in the number of *pairs* of events that occur. Because $I_i I_j$ will equal 1 if both A_i and A_j occur and will equal 0 otherwise, it follows that the number of pairs is equal to $\sum_{i<j} I_i I_j$. But because X is the number of events that occur, it also follows that the number of pairs of events that occur is $\binom{X}{2}$. Consequently,

$$\binom{X}{2} = \sum_{i<j} I_i I_j$$

where there are $\binom{n}{2}$ terms in the summation. Taking expectations yields

$$E\left[\binom{X}{2}\right] = \sum_{i<j} E[I_i I_j] = \sum_{i<j} P(A_i A_j)$$

or

$$E\left[\frac{X(X-1)}{2}\right] = \sum_{i<j} P(A_i A_j)$$

giving that

$$E[X^2] - E[X] = 2 \sum_{i<j} P(A_i A_j)$$

which yields $E[X^2]$, and thus $\mathrm{Var}(X) = E[X^2] - (E[X])^2$.

Moreover, by considering the number of distinct subsets of k events that all occur, we see that

$$\binom{X}{k} = \sum_{i_1<i_2<\cdots<i_k} I_{i_1} I_{i_2} \cdots I_{i_k}$$

Taking expectations gives the identity

$$E\left[\binom{X}{k}\right] = \sum_{i_1<i_2<\cdots<i_k} E\left[I_{i_1} I_{i_2} \cdots I_{i_k}\right] = \sum_{i_1<i_2<\cdots<i_k} P\left(A_{i_1} A_{i_2} \cdots A_{i_k}\right)$$

Example 3a: Moments of binomial random variables. (Page 299)

Example 3b: Moments of hypergeometric random variables. (Page 299)

Example 3c: Moments in the match problem. (Page 300)

Example 3d: Another coupon-collecting problem. (Page 301)

Example 3e: The negative hypergeometric random variables.

Suppose an urn contains $n+m$ balls, of which n are special and m are ordinary. These items are removed one at a time, with each new removal being equally likely to be any of the balls that remain in the urn. The random variable Y, equal to the number of balls that need be withdrawn until a total of r special balls have been removed, is said to have a *negative hypergeometric distribution*. (Page 302)

Example 3f: Singletons in the coupon collector's problem. (Page 303) (unfinished)

7.4 Covariance, Variance of Sums, and Correlations

The following proposition shows that the expectation of a product of independent random variables is equal to the product of their expectations.

Proposition 4.1: If X and Y are independent, then, for any functions h and g,

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

Just as the expected value and the variance of a single random variable give us information about that random variable, so does the covariance between two random variables give us information about the relationship between the random variables.

Definition

The **covariance** between X and Y, denoted by $\mathrm{Cov}(X,Y)$, is defined by

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

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

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

Note that if X and Y are independent, then, by Proposition 4.1, $\mathrm{Cov}(X,Y) = 0$. However, the converse is **not** true.

The following proposition lists some of the properties of covariance.

Proposition 4.2:

i. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$

ii. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$

iii. $\mathrm{Cov}(aX,X) = a\,\mathrm{Cov}(X,X)$

iv. $\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$

It follows from parts (ii) and (iv) of the above proposition, upon taking $Y_j = X_j, j = 1,\cdots,n$, that

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

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

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

If $X_1, \cdots, X_n$ are pairwise independent, in that X_i and X_j are independent for $i \neq j$, then the above equation reduces to

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

Example 4a: Let $X_1, \cdots, X_n$ be independent and identically distributed random variables having expected value μ and variance σ^2, and as in Example 2c, let $\bar{X} = \sum_{i=1}^{n} X_i/n$ be the sample mean. The quantities $X_i - \bar{X}$, $i = 1, \cdots, n$, 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)

$$\text{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \text{Var}(X_i) \quad \text{by independence}$$

$$= \frac{\sigma^2}{n}$$

(b) We start with the following algebraic identity:

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

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

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

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

Taking expectations of the preceding yields

$$(n-1)\,E[S^2] = \sum_{i=1}^{n} E[(X_i - \mu)^2] - n\,E[(\bar{X} - \mu)^2]$$

$$= n\sigma^2 - n\,\text{Var}(\bar{X})$$

$$= (n-1)\sigma^2$$

Dividing through by $n-1$ shows that the expected value of the sample variance is the distribution variance σ^2. ($E[S^2] = \sigma^2$)

Example 4b: Variance of a binomial random variable. (Page 308)

Example 4c: Sampling from a finite population. (Page 309)

The **correlation** of two random variables X and Y, denoted by $\rho(X, Y)$, is defined, as long as $\text{Var}(X)\,\text{Var}(Y)$ is positive, by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

It can be shown that

$$-1 \leq \rho(X, Y) \leq 1$$

To prove the above inequality, suppose that X and Y have variances given by σ_x^2 and σ_y^2,

respectively. Then, on the one hand,

$$0 \le \mathrm{Var}\left(\frac{X}{\sigma_x}+\frac{Y}{\sigma_y}\right)$$
$$=\frac{\mathrm{Var}(X)}{\sigma_x^2}+\frac{\mathrm{Var}(Y)}{\sigma_y^2}+\frac{2\,\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}$$
$$=2[1+\rho(X,Y)]$$

implying that

$$-1 \le \rho(X,Y)$$

On the other hand,

$$0 \le \mathrm{Var}\left(\frac{X}{\sigma_x}-\frac{Y}{\sigma_y}\right)$$
$$=\frac{\mathrm{Var}(X)}{\sigma_x^2}+\frac{\mathrm{Var}(Y)}{(-\sigma_y)^2}-\frac{2\,\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}$$
$$=2[1-\rho(X,Y)]$$

implying that

$$\rho(X,Y) \le 1$$

In fact, since $\mathrm{Var}(Z)=0$ implies that Z is constant with probability 1 (this intuitive relationship will be rigorously proven in Chapter 8), it follows from the proof of the equation $-1 \le \rho(X,Y) \le 1$ that $\rho(X,Y)=1$ implies that $Y=a+bX$, where $b=\sigma_y/\sigma_x>0$ and $\rho(X,Y)=-1$ implies that $Y=a+bX$, where $b=-\sigma_y/\sigma_x<0$. And the reverse is also true: that if $Y=a+bX$, then $\rho(X,Y)$ is either $+1$ or -1, depending on the sign of b.

The correlation coefficient is a measure of the degree of linearity between X and Y. A value of $\rho(X,Y)$ near $+1$ or -1 indicates a high degree of linearity between X and Y, whereas a value near 0 indicates that such linearity is absent. A positive value of $\rho(X,Y)$ indicates that Y tends to increase when X does, whereas a negative value indicates that Y tends to decrease when X increases. If $\rho(X,Y)=0$, then X and Y are said to be *uncorrelated*.

Example 4d: We obtain the quite intuitive result that the indicator variables for A and B are either positively correlated, uncorrelated, or negatively correlated, depending on whether $P(A|B)$ is, respectively, greater than, equal to, or less than $P(A)$. (Page 311)

Example 4e: The example shows that the sample mean and a deviation from the sample mean are uncorrelated. (Page 311)

Although $\bar{X}$ and the deviation $X_i-\bar{X}$ are uncorrelated, they are not, in general, independent. However, in the special case where the X_i are normal random variables, it turns out that not only is $\bar{X}$ independent of a single deviation, but it is independent of the entire sequence of deviations $X_j-\bar{X}, j=1,\cdots,n$. (Page 312)

Example 4f: Consider m independent trials, each of which results in any of r possible outcomes with probabilities $p_1,\cdots,p_r$, $\sum_{i=1}^{r}p_i=1$. If we let $N_i, i=1,\cdots,r$, denote the number of the m trials that result in outcome i, then $N_1,N_2,\cdots,N_r$ have the multinomial distribution (Page 312)

$$P\{N_1=n_1,\cdots,N_r=n_r\}=\frac{m!}{n_1!\cdots n_r!}p_1^{n_1}\cdots p_r^{n_r},\quad \sum_{i=1}^{r}n_i=m$$

7.5 Conditional Expectation

● Definitions

Recall that if X and Y are jointly discrete random variables, then the conditional probability mass function of X, given that $Y=y$, is defined for all y such that $P\{Y=y\}>0$, by

$$p_{X|Y}(x|y)=P\{X=x|Y=y\}=\frac{p(x,y)}{p_Y(y)}$$

It is therefore natural to define, in this case, the conditional expectation of X given that $Y=y$, for all values of y such that $p_Y(y)>0$, by

$$E[X|Y=y]=\sum_x xP\{X=x|Y=y\}$$

$$= \sum_x x p_{X|Y}(x|y)$$

Similarly, let us recall that if X and Y are jointly continuous with a joint probability density function $f(x,y)$, then the conditional probability density of X, given that $Y = y$, is defined for all values of y such that $f_Y(y) > 0$ by

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

It is natural, in this case, to define the conditional expectation of X, given that $Y = y$, by

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx$$

provided that $f_Y(y) > 0$.

Remark Just as conditional probabilities satisfy all the properties of ordinary probabilities, so do conditional expectations satisfy the properties of ordinary expectations. For instance, such formulas as

$$E[g(X)|Y = y] = \begin{cases} \sum_x g(x) p_{X|Y}(x|y) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\, dx & \text{in the continuous case} \end{cases}$$

and

$$E\left[\sum_{i=1}^{n} X_i | Y = y\right] = \sum_{i=1}^{n} E[X_i|Y = y]$$

remain valid. As a matter of fact, conditional expectation given that $Y = y$ can be thought of as being an ordinary expectation on a reduced sample space consisting only of outcomes for which $Y = y$.

● Computing Expectations by Conditioning

Let us denote by $E[X|Y]$ that function of the random variable Y whose value at $Y = y$ is $E[X|Y = y]$. Note that $E[X|Y]$ is itself a random variable. An extremely important property of conditional expectations is given by the following proposition.

Proposition 5.1:

$$E[X] = E\big[E[X|Y]\big]$$

If Y is a discrete random variable, then the above equation states that

$$E[X] = \sum_y E[X|Y = y] P\{Y = y\}$$

whereas if Y is continuous with density $f_Y(y)$, then the equation states

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y] f_Y(y)\, dy$$

One way to understand the equation $E[X] = \sum_y E[X|Y = y] P\{Y = y\}$ is to interpret it as follows: To calculate $E[X]$, we may take a weighted average of the conditional expected value of X given that $Y = y$, each of the terms $E[X|Y = y]$ being weighted by the probability of the event on which it is conditioned. (Of what does this remind you?) This is an extremely useful result that often enables us to compute expectations easily by first conditioning on some appropriate random variable.

Example 5c: A miner is trapped in a mine containing 3 doors. (Page 316)

Example 5d: Expectation of a sum of a random number of random variables. (Page 317)

Example 5f: Define the correlation. (Page 319)

Example 5h: Variance of the geometric distribution. (Page 321)

Example 5i: Consider a gambling situation in which there are r players, with player i initially having n_i units, $n_i > 0$, $i = 1, \cdots, r$. At each stage, two of the players are chosen to play a game, with the winner of the game receiving 1 unit from the loser.

It is interesting to note that while our argument shows that the mean number of stages does not depend on the manner in which the teams are selected at each stage, the same is not true

for the distribution of the number of stages. (Page 322)

Example 5j: Let $U_1, U_2, \cdots$ be a sequence of independent uniform $(0,1)$ random variables. Find $E[N]$ when

$$N = \min\left\{n: \sum_{i=1}^{n} U_i > 1\right\}$$

The expected number of uniform $(0,1)$ random variables that need to be added until their sum exceeds 1, is equal to e. (Page 325)

● Computing Probabilities by Conditioning

Not only can we obtain expectations by first conditioning on an appropriate random variable, but we can also use this approach to compute probabilities. To see this, let E denote an arbitrary event, and define the indicator random variable X by

$$X = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{if } E \text{ does not occur} \end{cases}$$

It follows from the definition of X that

$$E[X] = P(E)$$

$$E[X|Y = y] = P(E|Y = y) \quad \text{for any random variable } Y$$

Therefore, from the extremely useful equation in the chapter of Computing Expectations by Conditioning, we obtain

$$P(E) = \sum_{y} P(E|Y = y)P(Y = y) \quad \text{if } Y \text{ is discrete}$$

$$= \int_{-\infty}^{\infty} P(E|Y = y)f_Y(y)\,dy \quad \text{if } Y \text{ is continuous}$$

Note that if Y is a discrete random variable taking on one of the values $y_1, \cdots, y_n$, then by defining the events F_i, $i = 1, \cdots, n$, by $F_i = \{Y = y_i\}$, the above equation reduces to the familiar equation

$$P(E) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

where $F_1, \cdots, F_n$ are mutually exclusive events whose union is the sample space.

Example 5k: The best-prize problem. (Page 326)

● Conditional Variance

Just as we have defined the conditional expectation of X given the value of Y, we can also define the conditional variance of X given that $Y = y$:

$$\mathrm{Var}(X|Y) \equiv E[(X - E[X|Y])^2|Y]$$

That is, $\mathrm{Var}(X|Y)$ is equal to the (conditional) expected square of the difference between X and its (conditional) mean when the value of Y is given. In other words, $\mathrm{Var}(X|Y)$ is exactly analogous to the usual definition of variance, but now all expectations are conditional on the fact that Y is known.

There is a very useful relationship between $\mathrm{Var}(X)$, the unconditional variance of X, and $\mathrm{Var}(X|Y)$, the conditional variance of X given Y, that can often be applied to compute $\mathrm{Var}(X)$. (Page 329)

Proposition 5.2: The conditional variance formula

$$\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] + \mathrm{Var}(E[X|Y])$$

Example 5p: Variance of a sum of a random number of random variables. (Page 330)

7.6 Conditional Expectation and Prediction

Sometimes a situation arises in which the value of a random variable X is observed and then, on the basis of the observed value, an attempt is made to predict the value of a second random variable Y. Let $g(X)$ denote the predictor; that is, if X is observed to equal x, then $g(X)$ is our prediction for the value of Y. Clearly, we would like to choose g so that $g(X)$ tends to be close to Y. One possible criterion for closeness is to choose g so as to minimize $E\left[\left(Y - g(X)\right)^2\right]$. We now show that, under this criterion, the best possible predictor of Y is $g(X) = E[Y|X]$.

Proposition 6.1: (Page 330)

$$E\left[\left(Y-g(X)\right)^2\right] \geq E[(Y-E[Y|X])^2]$$

Remark A second, more intuitive, although less rigorous, argument verifying Proposition 6.1 is as follows: It is straightforward to verify that $E[(Y-c)^2]$ is minimized at $c=E[Y]$. Thus, if we want to predict the value of Y when there are no data available to use, the best possible prediction, in the sense of minimizing the mean square error, is to predict that Y will equal its mean. However, if the value of the random variable X is observed to be x, then the prediction problem remains exactly as in the previous (no-data) case, with the exception that all probabilities and expectations are now conditional on the event that $X=x$. Hence, the best prediction in this situation is to predict that Y will equal its conditional expected value given that $X=x$, thus establishing Proposition 6.1.

Example 6b: Suppose that if a signal value s is sent from location A, then the signal value received at location B is normally distributed with parameters $(s,1)$. (Page 331)

Example 6c: In digital signal processing, raw continuous analog data X must be quantized, or discretized, in order to obtain a digital representation. (Page 332) (Page 333: the best linear predictor of Y with respect to X)

Example 6d: when X and Y have a bivariate normal distribution, the best linear predictor of Y with respect to X is the best overall predictor, in which the conditional expectation of Y given X is linear in X. (Page 334)

7.7 Moment Generating Functions

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

$$M(t)=E[e^{tX}]$$

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

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,

$$\begin{aligned}M'(t)&=\frac{d}{dt}E[e^{tX}]\\&=E\left[\frac{d}{dt}(e^{tX})\right]\\&=E[Xe^{tX}]\end{aligned}$$

where we have assumed that the interchange of the differentiation and expectation operators is legitimate. That is, we have assumed that

$$\frac{d}{dt}\left[\sum_x e^{tx}p(x)\right]=\sum_x \frac{d}{dt}[e^{tx}p(x)]$$

in the discrete case and

$$\frac{d}{dt}\left[\int e^{tx}f(x)\,dx\right]=\int \frac{d}{dt}[e^{tx}f(x)]\,dx$$

in the continuous case. This assumption can almost always be justified and, indeed, is valid for all of the distributions considered in this book.

At $t=0$, we obtain

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

Similarly,

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

Thus,

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

In general, the nth derivative of $M(t)$ is given by

$$M^n(t) = E[X^n e^{tX}] \quad n \geq 1$$

implying that

$$M^n(0) = E[X^n] \quad n \geq 1$$

Example 7a: Binomial distribution with parameters n and p

If X is a binomial random variable with parameters n and p, then (Page 336)

$$M(t) = (pe^t + 1 - p)^n$$

Example 7b: Poisson distribution with mean λ

If X is a Poisson random variable with parameter λ, then (Page 336)

$$M(t) = \exp\{\lambda(e^t - 1)\}$$

Example 7c: Exponential distribution with parameter λ (Page 337)

$$M(t) = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda$$

Example 7d: 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

$$M_Z(t) = E[e^{tZ}] = e^{t^2/2}$$

we recall that $X = \mu + \sigma Z$ will have a normal distribution with parameters μ and σ^2 whenever Z is a unit normal random variable. (Page 337)

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

Figure 7.1 and Figure 7.2 give the moment generating functions for some common discrete and continuous distributions.

	Probability mass function, $p(x)$	Moment generating function, $M(t)$	Mean	Variance
Binomial with parameters n, p; $0 \leq p \leq 1$	$\binom{n}{x} p^x(1-p)^{n-x}$ $x = 0, 1, \ldots, n$	$(pe^t + 1 - p)^n$	np	$np(1-p)$
Poisson with parameter $\lambda > 0$	$e^{-\lambda}\frac{\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$	$\exp\{\lambda(e^t - 1)\}$	λ	λ
Geometric with parameter $0 \leq p \leq 1$	$p(1-p)^{x-1}$ $x = 1, 2, \ldots$	$\frac{pe^t}{1-(1-p)e^t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative binomial with parameters r, p; $0 \leq p \leq 1$	$\binom{n-1}{r-1} p^r(1-p)^{n-r}$ $n = r, r+1, \ldots$	$\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

Figure 7. 1 Discrete Probability Distribution

	Probability density function, $f(x)$	Moment generating function, $M(t)$	Mean	Variance
Uniform over (a, b)	$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential with parameter $\lambda > 0$	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$\frac{\lambda}{\lambda - t}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma with parameters $(s, \lambda), \lambda > 0$	$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{s-1}}{\Gamma(s)} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$\left(\frac{\lambda}{\lambda - t}\right)^s$	$\frac{s}{\lambda}$	$\frac{s}{\lambda^2}$
Normal with parameters (μ, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$	μ	σ^2

Figure 7. 2 Continuous Probability Distribution

An important property of moment generating functions is that the moment generating function of the sum of independent random variables equals the product of the individual moment generating functions,

$$M_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$$

where the next-to-last equality follows from Proposition 4.1, since X and Y are independent. Another important result is that the moment generating function uniquely determines the distribution. That is, if $M_X(t)$ exists and is finite in some region about $t = 0$, then the distribution of X is uniquely determined.

Example 7f: Sums of independent binomial random variables. (Page 339)

Example 7g: Sums of independent Poisson random variables. (Page 341)

Example 7h: Sums of independent normal random variables. (Page 341)

Example 7i: Compute the moment generating function of a chi-squared random variable with n degrees of freedom (Page 341)

$$M(t) = (1 - 2t)^{-n/2}$$

Example 7j: Moment generating function of the sum of a random number of random variables. (Page 342)

● Joint Moment Generating Functions

It is also possible to define the joint moment generating function of two or more random variables. This is done as follows: For any n random variables $X_1, \cdots, X_n$, the joint moment generating function, $M(t_1, \cdots, t_n)$, is defined, for all real values of $t_1, \cdots, t_n$, by

$$M(t_1, \cdots, t_n) = E[e^{t_1X_1 + \cdots + t_nX_n}]$$

The individual moment generating functions can be obtained from $M(t_1, \cdots, t_n)$ by letting all but one of the t_i's be 0. That is,

$$M_{X_i}(t) = E[e^{tX_i}] = M(0, \cdots, 0, t, 0, \cdots, 0)$$

where the t is in the ith place.

The joint moment generating function $M(t_1, \cdots, t_n)$ uniquely determines the joint distribution of $X_1, \cdots, X_n$. This result can then be used to prove that the n random variables $X_1, \cdots, X_n$ are independent if and only if

$$M(t_1, \cdots, t_n) = M_{X_1}(t_1) \cdots M_{X_n}(t_n)$$

7.8 Additional Properties of Normal Random Variables

● The Multivariate Normal Distribution

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

$$\begin{aligned} X_1 &= a_{11}Z_1 + \cdots + a_{1n}Z_n + \mu_1 \\ X_2 &= a_{21}Z_1 + \cdots + a_{2n}Z_n + \mu_2 \\ &\vdots \\ X_i &= a_{i1}Z_1 + \cdots + a_{in}Z_n + \mu_i \\ &\vdots \\ X_m &= a_{m1}Z_1 + \cdots + a_{mn}Z_n + \mu_m \end{aligned}$$

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

$$E[X_i] = \mu_i$$

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

Let us now consider

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

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

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

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

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

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

● The Joint Distribution of the Sample Mean and Sample Variance

Proposition 8.1: If $X_1, \cdots, X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2, then the sample mean $\bar{X}$ and the sample variance S^2 are independent. $\bar{X}$ is a normal random variable with mean μ and variance σ^2/n; $(n-1)S^2/\sigma^2$ is a chi-squared random variable with $n-1$ degrees of freedom.

7.9 General Definition of Expectation

Up to this point, we have defined expectations only for discrete and continuous random variables. However, there also exist random variables that are neither discrete nor continuous, and they, too, may possess an expectation.

For any distribution function F, we define the Stieltjes integral of the nonnegative function g over the interval $[a, b]$ by

$$\int_a^b g(x)dF(x) = \lim \sum_{i=1}^{n} g(x_i)[F(x_i) - F(x_{i-1})]$$

Further, we define the Stieltjes integral over the whole

$$\int_{-\infty}^{\infty} g(x)dF(x) = \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_a^b g(x)dF(x)$$

Finally, if g is not a nonnegative function, we define g^+ and g^- by

$$g^+(x) = \begin{cases} g(x) & \text{if } g(x) \geq 0 \\ 0 & \text{if } g(x) < 0 \end{cases}$$

$$g^-(x) = \begin{cases} 0 & \text{if } g(x) \geq 0 \\ -g(x) & \text{if } g(x) < 0 \end{cases}$$

Because $g(x) = g^+(x) - g^-(x)$ and g^+ and g^- are both nonnegative functions, it is natural to define

$$\int_{-\infty}^{\infty} g(x)dF(x) = \int_{-\infty}^{\infty} g^+(x)dF(x) - \int_{-\infty}^{\infty} g^-(x)dF(x)$$

and we say that $\int_{-\infty}^{\infty} g(x)dF(x)$ exists as long as $\int_{-\infty}^{\infty} g^+(x)dF(x)$ and $\int_{-\infty}^{\infty} g^-(x)dF(x)$ are not both equal to $+\infty$.

If X is an arbitrary random variable having cumulative distribution F, we define the expected value of X by

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

CHAPTER 8: Limit Theorems

8.1 Introduction

The most important theoretical results in probability theory are limit theorems. Of these, the most important are those classified either under the heading ***laws of large numbers*** or under the heading ***central limit theorems***. Usually, theorems are considered to be laws of large numbers if they are concerned with stating conditions under which the average of a sequence of random variables converges (in some sense) to the expected average. By contrast, central limit theorems are concerned with determining conditions under which the sum of a large number of random variables has a probability distribution that is approximately normal.

8.2 Chebyshev's Inequality and the Weak Law of Large Numbers

Proposition 2.1: Markov's inequality

If X is a random variable that takes only nonnegative values, then for any value $a > 0$,

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

Proof:

For $a > 0$, let

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

and note that, since $X \geq 0$,

$$I \leq \frac{X}{a}$$

Taking expectations of the preceding inequality yields

$$E[I] \leq \frac{E[X]}{a}$$

which, because $E[I] = P\{X \geq a\}$, proves the result.

Proposition 2.2: Chebyshev's inequality

If X is a random variable with finite mean μ and variance σ^2, then for any value $k > 0$,

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

Proof

Since $(X - \mu)^2$ is a nonnegative random variable, we can apply Markov's inequality (with $a = k^2$) to obtain

$$P\{(X - \mu)^2 \geq k^2\} \leq \frac{E[(X - \mu)^2]}{k^2}$$

But since $(X - \mu)^2 \geq k^2$ if and only if $|X - \mu| \geq k$, the above is equivalent to

$$P\{|X - \mu| \geq k\} \leq \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

and the proof is complete.

The importance of Markov's and Chebyshev's inequalities is that they enable us to derive bounds on probabilities when only the mean or both the mean and the variance of the probability distribution are known. Of course, if the actual distribution were known, then the desired probabilities could be computed exactly and we would not need to resort to bounds.

As Chebyshev's inequality is valid for all distributions of the random variable X, we cannot expect the bound on the probability to be very close to the actual probability in most cases.

Chebyshev's inequality is often used as a theoretical tool in proving results. This use is illustrated first by the following proposition and then, most importantly, by the weak law of large numbers.

Proposition 2.3: If $\mathrm{Var}(X) = 0$, then

$$P\{X = E[X]\} = 1$$

In other words, the only random variables having variances equal to 0 are those that are constant with probability 1.

Proof

By Chebyshev's inequality, we have, for any $n \geq 1$,

$$P\left\{|X-\mu| > \frac{1}{n}\right\} = 0 \cdot n^2 = 0$$

Letting $n \to \infty$ and using the continuity property of probability yields

$$0 = \lim_{n\to\infty} P\left\{|X-\mu| > \frac{1}{n}\right\} = P\left\{\lim_{n\to\infty}\left\{|X-\mu| > \frac{1}{n}\right\}\right\}$$
$$= P\{X \neq \mu\}$$

and the result is established.

Theorem 2.1: **The weak law of large numbers**

Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables, each having finite mean $E[X_i] = \mu$. Then, for any $\varepsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right\} \to 0 \quad \text{as} \quad n \to \infty$$

Proof

We shall prove the theorem only under the additional assumption that the random variables have a finite variance σ^2. Now, since

$$\mathrm{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu \quad \text{and} \quad \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

it follows from Chebyshev's inequality that

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{n\varepsilon^2}$$

and the result is proven.

8.3 The Central Limit Theorem

The central limit theorem is one of the most remarkable results in probability theory. Loosely put, it states that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped (that is, normal) curves.

Theorem 3.1: **The central limit theorem**

Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}\,dx \quad \text{as} \quad n \to \infty$$

The key to the proof of the central limit theorem is the following lemma, which we state without proof.

Let $Z_1, Z_2, \cdots$ be a sequence of random variables having distribution functions F_{Z_n} and moment generating functions M_{Z_n}, $n \geq 1$, and let Z be a random variable having distribution function F_Z and moment generating function M_Z. If $M_{Z_n}(t) \to M_Z(t)$ for all t, then $F_{Z_n}(t) \to F_Z(t)$ for all t at which $F_Z(t)$ is continuous.

Remark Although the central limit theorem states only that, for each a,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \Phi(a)$$

it can, in fact, be shown that the convergence is uniform in a. [We say that $f_n(a) \to f(a)$ uniformly in a if, for each $\varepsilon > 0$, there exists an N such that $|f_n(a) \to f(a)| < \varepsilon$ for all a whenever $n \geq N$.]

Central limit theorems also exist when the X_i are independent, but not necessarily identically distributed random variables. One version, by no means the most general, is as follows.

Theorem 3.2: Central limit theorem for independent random variables

Let $X_1, X_2, \cdots$ be a sequence of independent random variables having respective means and

variances $\mu_i = E[X_i]$, $\sigma_i^2 = \text{Var}(X_i)$. If (a) the X_i are uniformly bounded—that is, if for some M, $P\{|X_i| < M\} = 1$ for all i, and (b) $\sum_{i=1}^{\infty} \sigma_i^2 = \infty$-then

$$P\left\{\frac{\sum_{i=1}^{n}(X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \le a\right\} \to \Phi(a) \quad \text{as} \quad n \to \infty$$

8.4 The Strong Law of Large Numbers

The *strong law of large numbers* is probably the best-known result in probability theory. It states that the average of a sequence of independent random variables having a common distribution will, with probability 1, converge to the mean of that distribution.

Theorem 4.1: **The strong law of large numbers**

Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables, each having a finite mean $\mu = E[X_i]$. Then, with probability 1,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \quad \text{as} \quad n \to \infty$$

That is, the strong law of large numbers states that

$$P\left\{\lim_{n\to\infty}(X_1 + X_2 + \cdots + X_n)/n = \mu\right\} = 1$$

As an application of the strong law of large numbers, suppose that a sequence of independent trials of some experiment is performed. Let E be a fixed event of the experiment, and denote by $P(E)$ the probability that E occurs on any particular trial. Letting

$$X_i = \begin{cases} 1 & \text{if } E \text{ occurs on the } i\text{th trial} \\ 0 & \text{if } E \text{ does not occur on the } i\text{th trial} \end{cases}$$

we have, by the strong law of large numbers, that with probability 1,

$$\frac{X_1 + \cdots + X_n}{n} = X \to E[X] = P(E)$$

Since $X_1 + \cdots + X_n$ represents the number of times that the event E occurs in the first n trials, we may interpret the above equation as stating that with probability 1, the limiting proportion of time that the event E occurs is just $P(E)$.

Many students are initially confused about the difference between the weak and the strong laws of large numbers. The weak law of large numbers states that for any specified large value n^*, $(X_1 + \cdots + X_{n^*})/n^*$ is likely to be near μ. However, it does not say that $(X_1 + \cdots + X_n)/n$ is bound to stay near μ for all values of n larger than n^*. Thus, it leaves open the possibility that large values of $|(X_1 + \cdots + X_n)/n - \mu|$ can occur infinitely often (though at infrequent intervals). The strong law shows that this cannot occur. In particular, it implies that, with probability 1, for any positive value ε,

$$\left|\sum_{1}^{n}\frac{X_i}{n} - \mu\right|$$

will be greater than ε only a finite number of times.

8.5 Other Inequalities

Proposition 5.1: One-sided Chebyshev inequality

If X is a random variable with mean 0 and finite variance σ^2, then, for any $a > 0$,

$$P\{X \ge a\} \le \frac{\sigma^2}{\sigma^2 + a^2}$$

Suppose now that X has mean μ and variance σ^2. Since both $X - \mu$ and $\mu - X$ have mean 0 and variance σ^2, it follows from the one-sided Chebyshev inequality that, for $a > 0$,

$$P\{X - \mu \ge \text{a}\} \le \frac{\sigma^2}{\sigma^2 + a^2}$$

and

$$P\{\mu - X \ge \text{a}\} \le \frac{\sigma^2}{\sigma^2 + a^2}$$

Thus, we have the following corollary.

Corollary 5.1: If $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$, then, for $a > 0$,

$$P\{X \ge \mu + \text{a}\} \le \frac{\sigma^2}{\sigma^2 + a^2}$$

$$P\{X \leq \mu - a\} \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

When the moment generating function of the random variable X is known, we can obtain even more effective bounds on $P\{X \geq a\}$. Let

$$M(t) = E[e^{tX}]$$

be the moment generating function of the random variable X. Then, for $t > 0$,

$$P\{X \geq a\} = P\{e^{tX} \geq e^{ta}\}$$
$$\leq E[e^{tX}]e^{-ta} \quad by\ Markov's\ inequality$$

Similarly, for $t < 0$,

$$P\{X \leq a\} = P\{e^{tX} \geq e^{ta}\}$$
$$\leq E[e^{tX}]e^{-ta}$$

Thus, we have the following inequalities, known as *Chernoff bounds*.

Proposition 5.2: Chernoff bounds

$$P\{X \geq a\} \leq e^{-ta}M(t) \quad \text{for all} \quad t > 0$$
$$P\{X \leq a\} \leq e^{-ta}M(t) \quad \text{for all} \quad t < 0$$

Since the Chernoff bounds hold for all t in either the positive or negative quadrant, we obtain the best bound on $P\{X \geq a\}$ by using the t that minimizes $e^{-ta}M(t)$.

Example 5c: Chernoff bounds for the standard normal random variable (Page 385)

Example 5d: Chernoff bounds for the Poisson random variable (Page 385)

The next inequality is one having to do with expectations rather than probabilities. Before stating it, we need the following definition.

> **Definition**
>
> A twice-differentiable real-valued function $f(x)$ is said to be *convex* if $f''(x) \geq 0$ for all x; similarly, it is said to be *concave* if $f''(x) \leq 0$.

Proposition 5.3: Jensen's inequality

If $f(x)$ is a convex function, then

$$E[f(X)] \geq f(E[X])$$

provided that the expectations exist and are finite.

Proof

Expanding $f(x)$ in a Taylor's series expansion about $\mu = E[X]$ yields

$$f(x) = f(\mu) + f'(\mu)(x - \mu) + \frac{f''(\xi)(x - \mu)^2}{2}$$

where ξ is some value between x and μ. Since $f''(\xi) \geq 0$, we obtain

$$f(x) \geq f(\mu) + f'(\mu)(x - \mu)$$

Hence,

$$f(X) \geq f(\mu) + f'(\mu)(X - \mu)$$

Taking expectations yields

$$E[f(X)] \geq f(\mu) + f'(\mu)E[X - \mu] = f(\mu)$$

and the inequality is established.

8.6 Bounding the Error Probability When Approximating a Sum of Independent Bernoulli Random Variables by a Poisson Random Variable

In this section, we establish bounds on how closely a sum of independent Bernoulli random variables is approximated by a Poisson random variable with the same mean.

$$|P\{X \in A\} - P\{Y \in A\}| \leq \sum_{i=1}^{n} p_i^2$$

where for any set of real numbers A, X is the sum of independent Bernoulli random variables and Y is a Poisson random variable.

Remark When all the p_i are equal to p, X is a binomial random variable. Hence, the preceding inequality shows that, for any set of nonnegative integers A,

$$\left| \sum_{i \in A} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i \in A} \frac{e^{-np}(np)^i}{i!} \right| \leq np^2$$

CHAPTER 9: Additional Topics in Probability

9.1 The Poisson Process

Before we define a Poisson process, let us recall that a function f is said to be $o(h)$ if

$$\lim_{h\to 0}\frac{f(h)}{h}=0.$$

That is, f is $o(h)$ if, for small values of h, $f(h)$ is small even in relation to h. Suppose now that "events" are occurring at random points at time, and let $N(t)$ denote the number of events that occur in the time interval $[0,t]$. The collection of random variables $\{N(t), t\geq 0\}$ is said to be a *Poisson process having rate* λ, $\lambda>0$, if

i. $N(0)=0$
ii. The numbers of events that occur in disjoint time intervals are independent.
iii. The distribution of the number of events that occur in a given interval depends only on the length of that interval and not on its location.
iv. $P\{N(h)=1\}=\lambda h+o(h)$.
v. $P\{N(h)\geq 2\}=o(h)$.

Thus, condition (i) states that the process begins at time 0. Condition (ii), the *independent increment* assumption, states, for instance, that the number of events that occur by time t [that is, $N(t)$] is independent of the number of events that occur between t and $t+s$ [that is, $N(t+s)-N(t)$]. Condition (iii), the *stationary increment* assumption, states that the probability distribution of $N(t+s)-N(t)$ is the same for all values of t.

Lemma 1.1: For a Poisson process with rate λ,

$$P\{N(t)=0\}=e^{-\lambda t}$$

For a Poisson process, let T_1 denote the time the first event occurs. Further, for $n>1$, let T_n denote the time elapsed between the $(n-1)$ and the nth event. The sequence $\{T_n, n=1,2\cdots\}$ is called the *sequence of interarrival times*. For instance, if $T_1=5$ and $T_2=10$, then the first event of the Poisson process would have occurred at time 5 and the second at time 15.

Proposition 1.1: $T_1, T_2, \cdots$ are independent exponential random variables, each with mean $1/\lambda$.

Another quantity of interest is S_n, the arrival time of the nth event, also called the *waiting time* until the nth event. It is easily seen that

$$S_n=\sum_{i=1}^{n}T_i \quad n\geq 1$$

hence, S_n has a gamma distribution with parameters n and λ. That is, the probability density of S_n is given by

$$f_{S_n}(x)=\lambda e^{-\lambda x}\frac{(\lambda x)^{n-1}}{(n-1)!} \quad x\geq 0$$

Theorem 1.1: For a Poisson process with rate λ,

$$P\{N(t)=n\}=\frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

9.2 Markov Chains

Consider a sequence of random variables $X_0, X_1, \cdots$, and suppose that the set of possible values of these random variables is $\{0,1,\cdots,M\}$. It will be helpful to interpret X_n as being the state of some system at time n, and, in accordance with this interpretation, we say that the system is in state i at time n if $X_n=i$. The sequence of random variables is said to form a ***Markov chain*** if, each time the system is in state i, there is some fixed probability—call it P_{ij}-that the system will next be in state j. That is, for all $i_0,\cdots,i_{n-1},i,j$,

$$P\{X_{n+1}=j|X_n=i,X_{n-1}=i_{n-1},\cdots,X_1=i_1,X_0=i_0\}=P_{ij}$$

The values $P_{ij}, 0\leq i\leq M, 0\leq j\leq M$, are called the *transition probabilities* of the Markov chain, and they satisfy

$$P_{ij}\geq 0, \quad \sum_{j=0}^{M}P_{ij}=1, \quad i=0,1,\cdots,M$$

(Why?) It is convenient to arrange the transition probabilities P_{ij} in a square array as follows:

$$\begin{Vmatrix} P_{00} & P_{01} & \cdots & P_{0M} \\ P_{10} & P_{11} & \cdots & P_{1M} \\ \vdots & & & \\ P_{M0} & P_{M1} & \cdots & P_{MM} \end{Vmatrix}$$

Such an array is called a *matrix*.

Knowledge of the **transition probability matrix** and of the distribution of X_0 enables us, in theory, to compute all probabilities of interest. For instance, the joint probability mass function of $X_0, \cdots, X_n$ is given by

$$\begin{aligned} &P\{X_n = i_n, X_{n-1} = i_{n-1}, \cdots, X_1 = i_1, X_0 = i_0\} \\ &= P\{X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\} P\{X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\} \\ &= P_{i_{n-1}, i_n} P\{X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\} \end{aligned}$$

and continual repetition of this argument demonstrates that the preceding is equal to

$$P_{i_{n-1}, i_n} P_{i_{n-2}, i_{n-1}} \cdots P_{i_1, i_2} P_{i_0, i_1} P\{X_0 = i_0\}$$

Example 2a: Suppose that whether it rains tomorrow depends on previous weather conditions only through whether it is raining today. Suppose further that if it is raining today, then it will rain tomorrow with probability α, and if it is not raining today, then it will rain tomorrow with probability β.

If we say that the system is in state 0 when it rains and state 1 when it does not, then the preceding system is a two-state Markov chain having transition probability matrix

$$\begin{Vmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{Vmatrix}$$

That is, $P_{00} = \alpha = 1 - P_{01}$, $P_{10} = \beta = 1 - P_{11}$.

For a Markov chain, P_{ij} represents the probability that a system in state i will enter state j at the next transition. We can also define the two-stage transition probability $P_{ij}^{(2)}$ that a system presently in state i will be in state j after two additional transitions. That is,

$$P_{ij}^{(2)} = P\{X_{m+2} = j | X_m = i\} = \sum_{k=0}^{M} P_{kj} P_{ik}$$

Proposition 2.1: The Chapman–Kolmogorov equations

$$P_{ij}^{(n)} = \sum_{k=0}^{M} P_{ik}^{(r)} P_{kj}^{(n-r)} \quad \text{for all } 0 < r < n$$

Example 2d: A random walk (Page 399)

An example of a Markov chain having a countably infinite state space is the *random walk*, which tracks a particle as it moves along a one-dimensional axis.

For a large number of Markov chains, it turns out that $P_{ij}^{(n)}$ converges, as $n \to \infty$, to a value π_j that depends only on j. That is, for large values of n, the probability of being in state j after n transitions is approximately equal to π_j, no matter what the initial state was. It can be shown that a sufficient condition for a Markov chain to possess this property is that for some $n > 0$,

$$P_{ij}^{(n)} > 0 \quad \text{for all } i, j = 0, 1, \cdots, M$$

Markov chains that satisfy the above equation are said to be *ergodic*. Since the Chapman–Kolmogorov equations yields

$$P_{ij}^{(n+1)} = \sum_{k=0}^{M} P_{ik}^{(n)} P_{kj}$$

it follows, by letting $n \to \infty$, that for ergodic chains,

$$\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$$

Furthermore, since $1 = \sum_{j=0}^{M} P_{ij}^{(n)}$, we also obtain, by letting $n \to \infty$,

$$\sum_{j=0}^{M} \pi_j = 1$$

In fact, it can be shown that the π_j, $0 \le j \le M$, are the unique nonnegative solutions of equations $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ and $\sum_{j=0}^{M} \pi_j = 1$.

Theorem 2.1: For an ergodic Markov chain,

$$\pi_j = \lim_{n\to\infty} P_{ij}^{(n)}$$

exists, and the π_j, $0 \le j \le M$, are the unique nonnegative solutions of

$$\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$$

$$\sum_{j=0}^{M} \pi_j = 1$$

The quantity π_j is also equal to the long-run proportion of time that the Markov chain is in state j, $j = 0, \cdots, M$.

9.3 Surprise, Uncertainty, and Entropy

Consider an event E that can occur when an experiment is performed. How surprised would we be to hear that E does, in fact, occur? It seems reasonable to suppose that the amount of surprise engendered by the information that E has occurred should depend on the probability of E. For instance, if the experiment consists of rolling a pair of dice, then we would not be too surprised to hear that E has occurred when E represents the event that the sum of the dice is even (and thus has probability $\frac{1}{2}$), whereas we would certainly be more surprised to hear that E has occurred when E is the event that the sum of the dice is 12 (and thus has probability $\frac{1}{36}$).

To begin, let us agree to suppose that the surprise one feels upon learning that an event E has occurred depends only on the probability of E, and let us denote by $S(p)$ the surprise evoked by the occurrence of an event having probability p. We assume throughout that $S(p)$ is defined for all $0 < p \le 1$ but is not defined for events having $p = 0$.

Our first condition is just a statement of the intuitive fact that there is no surprise in hearing that an event that is sure to occur has indeed occurred.

Axiom 1

$$S(1) = 0$$

Our second condition states that the more unlikely an event is to occur, the greater is the surprise evoked by its occurrence.

Axiom 2

$S(p)$ is a strictly decreasing function of p; that is, if $p < q$, then $S(p) > S(q)$.

The third condition is a mathematical statement of the fact that we would intuitively expect a small change in p to correspond to a small change in $S(p)$.

Axiom 3

$S(p)$ is a continuous function of p.

To motivate the final condition, consider two independent events E and F having respective probabilities $P(E) = p$ and $P(F) = q$. Since $P(EF) = pq$, the surprise evoked by the information that both E and F have occurred is $S(pq)$. Since $S(p)$ is the surprise evoked by the occurrence of E, it follows that $S(pq) - S(p)$ represents the additional surprise evoked when we are informed that F has also occurred. However, because F is independent of E, the knowledge that E occurred does not change the probability of F; hence, the additional surprise should just be $S(q)$. This reasoning suggests the final condition.

Axiom 4

$$S(pq) = S(p) + S(q) \quad 0 < p \le 1, \quad 0 < q \le 1$$

Theorem 3.1: If $S(\cdot)$ satisfies Axioms 1 through 4, then

$$S(p) = -C \log_2 p$$

where C is an arbitrary positive integer. But in proof, $C = S\left(\frac{1}{2}\right) > S(1) = 0$ by Axioms 2 and 1. (Page 403)

It is usual to let C equal 1, in which case the surprise is said to be expressed in units of ***bits***

(short for *binary digits*).

Next, consider a random variable X that must take on one of the values $x_1, \cdots, x_n$ with respective probabilities $p_1, \cdots, p_n$. Since $-\log_2 p_i$ represents the surprise evoked if X takes on the value x_i, it follows that the **expected** amount of surprise we shall receive upon learning the value of X is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

The quantity $H(X)$ is known in information theory as the ***entropy*** of the random variable X. (In case one of the $p_i = 0$, we take $0 \log_2 0$ to equal 0.) It can be shown (and we leave it as an exercise) that $H(X)$ is maximized when all of the p_i are equal.

Since $H(X)$ represents the average amount of surprise one receives upon learning the value of X, it can also be interpreted as representing the amount of *uncertainty* that exists as to the value of X. In fact, in information theory, $H(X)$ is interpreted as the average amount of *information* received when the value of X is observed. Thus, the average surprise evoked by X, the uncertainty of X, or the average amount of information yielded by X all represent the same concept viewed from three slightly different points of view.

Now consider two random variables X and Y that take on the respective values $x_1, \cdots, x_n$ and $y_1, \cdots, y_m$ with joint mass function

$$p(x_i, y_j) = P\{X = x_i, Y = y_j\}$$

It follows that the uncertainty as to the value of the random vector (X, Y), denoted by $H(X, Y)$, is given by

$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log_2 p(x_i, y_j)$$

Suppose now that Y is observed to equal y_j. In this situation, the amount of uncertainty remaining in X is given by

$$H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log_2 p(x_i|y_j)$$

where

$$p(x_i|y_j) = P\{X = x_i|Y = y_j\}$$

Hence, the average amount of uncertainty that will remain in X after Y is observed is given by

$$H_Y(X) = \sum_j H_{Y=y_j}(X) p_Y(y_j)$$

where

$$p_Y(y_j) = P\{Y = y_j\}$$

Proposition 3.1 relates $H(X, Y)$ to $H(Y)$ and $H_Y(X)$. It states that the uncertainty as to the value of X and Y is equal to the uncertainty of Y plus the average uncertainty remaining in X when Y is to be observed.

Proposition 3.1: $H(X, Y) = H(Y) + H_Y(X)$

It is a fundamental result in information theory that the amount of uncertainty in a random variable X will, on the average, decrease when a second random variable Y is observed.

Lemma 3.1:

$$\ln x \leq x - 1 \quad x > 0$$

with equality only at $x = 1$.

Theorem 3.2:

$$H_Y(X) \leq H(X)$$

with equality if and only if X and Y are independent.

9.4 Coding Theory and Entropy

Suppose that the value of a discrete random vector X is to be observed at location A and then transmitted to location B via a communication network that consists of two signals, 0 and 1. In order to do this, it is first necessary to encode each possible value of X in terms of a sequence of 0's and 1's. To avoid any ambiguity, it is usually required that no encoded

sequence can be obtained from a shorter encoded sequence by adding more terms to the shorter.

One of the objectives in devising a code is to minimize the expected number of bits (that is, binary digits) that need to be sent from location A to location B.

For a given random vector X, what is the maximum efficiency achievable by an encoding scheme? The answer is that for any coding, the average number of bits that will be sent is at least as large as the entropy of X. To prove this result, known in information theory as the *noiseless coding theorem*, we shall need the lemma below.

Lemma 4.1: Let X take on the possible values $x_1, \cdots, x_N$. Then, in order to be able to encode the values of X in binary sequences (none of which is an extension of another) of respective lengths $n_1, \cdots, n_N$, it is necessary and sufficient that

$$\sum_{i=1}^{N} \left(\frac{1}{2}\right)^{n_i} \leq 1$$

Theorem 4.1: The noiseless coding theorem

Let X take on the values $x_1, \cdots, x_N$ with respective probabilities $p(x_1), \cdots, p(x_N)$. Then, for any coding of X that assign n_i bits to x_i,

$$\sum_{i=1}^{N} n_i p(x_i) \geq H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i)$$

For most random vectors, there does not exist a coding for which the average number of bits sent attains the lower bound $H(X)$. However, it is always possible to devise a code such that the average number of bits is within 1 of $H(X)$:

We can associate sequences of bits having lengths n_i with the $x_i, i = 1, \cdots, N$. The average length of such a sequence,

$$L = \sum_{i=1}^{N} n_i p(x_i)$$

satisfies

$$H(X) \leq L \leq H(X) + 1$$

Example 4b: Suppose that 10 independent tosses of a coin having probability p of coming up heads are made at location A and the result is to be transmitted to location B. (Page 409)

Up to this point, we have assumed that the message sent at location A is received without error at location B. However, there are always certain errors that can occur because of random disturbances along the communications channel.

Theorem 4.2: The noisy coding theorem

There is a number C such that for any value R that is less than C, and for any $\varepsilon > 0$, there exists a coding–decoding scheme that transmits at an average rate of R bits sent per signal and with an error (per bit) probability of less than ε. The largest such value of C-call it C^*- is called the channel capacity, and for the binary symmetric channel,

$$C^* = 1 + p \log_2 p + (1-p) \log_2 (1-p)$$