Chapter 1. The Calculus of Probabilities.

A century ago, French treatises on the theory of probability were commonly called "Le Calcul des Probabilités"—"The Calculus of Probabilities." The name has fallen out of fashion, perhaps due to the potential confusion with integral and differential calculus, but it seems particularly apt for our present topic. It suggests a system of rules that are generally useful for calculation, where the more modern "probability theory" has a speculative connotation. Theories may be overthrown or superceded; a calculus can be used within many theories. Rules for calculation can be accepted even when, as with probability, there may be different views as to the correct interpretation of the quantities being calculated.

The interpretation of probability has been a matter of dispute for some time, although the terms of the dispute have not remained constant. To say that an event (such as the occurrence of a Head in the toss of a coin) has probability 1/2 will mean to some that if the coin is tossed an extraordinarily large number of times, about half the results will be Heads, while to others it will be seen as a subjective assessment, an expression of belief about the uncertainty of the event that makes no reference to an idealized (and not realizable) infinite sequence of tosses. In this book we will not insist upon any single interpretation of probability; indeed, we will find it convenient to adopt different interpretations at different times, depending upon the scientific context. Probabilities may be interpreted as long run frequencies, in terms of random samples from large populations, or as degrees of belief. While not philosophically pure, this opportunistic approach will have the benefit of permitting us to develop a large body of statistical methodology that can appeal to and be useful to a large number of people in quite varied situations.

We will, then, begin with a discussion of a set of rules for manipulating or calculating with probabilities, rules which show how we can go from one assignment of probabilities to another, without prejudice to the source of the first assignment. Usually we will be interested in reasoning from simple situations to complex situations.

1.1 Probabilities of Events.

The rules will be introduced within the framework of what we will call an *experiment.* We will be purposefully vague as to exactly what we mean by an experiment, only describing it as some process with an observable outcome. The process may be planned or unplanned, a laboratory exercise or a passive historical recording of facts about society. For our purposes, the important point that specifies the experiment is that there is a set, or list, of all possible outcomes of the experiment, called the *sample space* and denoted S. An *event* (say E) is then a set of possible outcomes, a subset of S. We shall see that usually the same actual experiment may be described in terms of different sample spaces, depending upon the purpose of the description.

The notation we use for describing and manipulating events is borrowed from elementary set theory. If E and F are two events, both subsets of the same sample space S, then the *complement* of E (denoted E^c, or sometimes E') is the set of all outcomes not in E, the *intersection* of E and F ($E \cap F$) is the set of all outcomes in both E and F, and the *union* of E and F ($E \cup F$) is the set of all outcomes in E or in F or in both E and F. It is often convenient to represent these definitions, and arguments associated with them, in terms of shaded regions of *Venn diagrams*, where the rectangle S is the sample space and the areas E and F two events.

[Figure 1.1]

If E and F have no common outcomes, they are said to be *mutually exclusive.* Even with only these elementary definitions, fairly complicated relationships can be described. For example, consider the event $(A \cap B) \cup (A \cap B^c)$, where A and B are two events. Then a simple consideration of Venn diagrams shows that in fact this describes the same set of outcomes as A:

$$(A \cap B) \cup (A \cap B^c) = A.$$

[Figure 1.2]

Thus even without any notion of what the symbol P for probability may mean, we would have the identity

$$P((A \cap B) \cup (A \cap B^c)) = P(A).$$

Of course, for such equations to be useful we will need to define probability. As mentioned earlier, we avoid giving a limited interpretation to probability for the present, though we may, whatever the interpretation, think of it as a measure of uncertainty. But for all interpretations, probability will have certain properties, namely those of an additive set function. With respect to our general sample space S these are:

(1.1) *Scaling:* $P(S) = 1$, and $0 \leq P(E)$ for all E in S

(1.2) *Additivity:* If E and F are mutually exclusive, $P(E \cup F) = P(E) + P(F)$.

Property (1.1) is a scaling property; it says, in effect, that we measure uncertainty on a scale from 0 to 1, with 1 representing certainty. If we let ϕ denote the *null* or empty event (so $\phi = S^c$) where no outcome occurs, then since $S = S \cup \phi$, (1.2) tells us $P(\phi) = 0$, so 0 represents impossibility on our scale. If E and F are mutually exclusive, $E \cap F = \phi$, and $P(E \cap F) = 0$.

Property (1.2) may be taken to be a precise way of imposing order on the assignment of probabilities; it requires in particular that if E is smaller than F (that is, E is contained in F, $E \subset F$), then $P(E) \leq P(F)$. This use of a zero to one scale with increasing values representing greater certainty is by no means the only scale that could be used. Another scale that is used in some statistical applications is the *log odds:*

$$\text{log odds}(E) = \log_e \left(\frac{P(E)}{P(E^C)} \right).$$

This measures probability on a scale from $-\infty$ to ∞, with 0 as a middle value (corresponding to $P(E) = 1/2$). But for present purposes, the zero-to-one scale represented by $P(E)$ is convenient.

Together (1.1) and (1.2) imply a number of useful other properties. For example,

(1.3) *Complementarity:* $P(E) + P(E^c) = 1$ for all E in S.

(1.4) *General additivity:* For any E and F in S (not necessarily mutually exclusive),

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

(1.5) *Finite additivity:* For any finite collection of mutually exclusive events $E_1, E_2, \ldots, E_n$,

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i).$$

Properties (1.1) and (1.2) are not sufficiently strong to imply the more general version of (1.5), namely:

(1.6) *Countable additivity:* For any countably infinite collection of mutually exclusive events $E_1, E_2, \ldots$ in S,

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

For our purposes this is not a restrictive additional condition, so we shall add it to (1.1) and (1.2) as an assumption we shall make about the probabilities we deal with. In some advanced applications, for example where the sample space is a set of infinite sequences or a function space, there are useful probability measures that satisfy (1.1) and (1.2) but not (1.6), however.

Probabilities may in some instances be specified by hypothesis for simple outcomes, and the probabilities of more complex events computed from these rules. Indeed, in this chapter we shall only consider such hypothetical probabilities, and turn to empirical questions in a following chapter. A trivial example will illustrate.

Example 1.A: We might describe the experiment of tossing a single six-sided die by the sample space $S = \{1, 2, 3, 4, 5, 6\}$, where the possible outcomes are the numbers on the upper face when the die comes to rest. By hypothesis, we might suppose the die is "fair" and interpret this mathematically as meaning that each of these six outcomes has an equal probability; $P(\{1\}) = 1/6$, $P(\{2\}) = 1/6$, etc. As mentioned earlier, this statement is susceptible to several interpretations: it might represent your subjective willingness to bet on #1 at 5 to 1 odds, or the fact that in an infinite sequence of hypothetical tosses, one-sixth will show #1. But once we accept the hypothesis of equally likely faces under any interpretation, the calculations we make are valid under that interpretation. For example, if

$$E = \text{"an odd number is thrown"}$$

and

$$F = \text{"the number thrown is less than 3,"}$$

then $E = \{1\} \cup \{3\} \cup \{5\}, F = \{1\} \cup \{2\}$, and rule (1.5) implies

$$P(E) = P(\{1\}) + P(\{3\}) + P(\{5\}) = \frac{3}{6}$$

and

$$P(F) = P(\{1\}) + P(\{2\}) = \frac{2}{6}.$$

Furthermore, $E \cap F = \{1\}$, so $P(E \cap F) = 1/6$. Then by rule (1.4),

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6}.$$

In this simple situation, and even in more complicated ones, we have alternative ways of computing the same quantity. Here $E \cup F = \{1, 2, 3, 5\}$ and we can also verify that $P(E \cup F) = 4/6$ from rule (1.5).

1.2 Conditional Probability

In complex experiments it is common to simplify the specification of probabilities by describing them in terms of conditional probabilities. Intuitively, the *conditional probability* of an event E given an event F, written $P(E|F)$, is the probability that E occurs given that F has occurred. Mathematically, we may define this probability in terms of probabilities involving E and F as

(1.7) *Conditional probability:* The probability that E occurs given F has occurred is defined to be

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \quad \text{if } P(F) > 0.$$

If $P(F) = 0$ we leave $P(E|F)$ undefined for now. Conditional probability may be thought of as *relative* probability: $P(E|F)$ is the probability of E relative to the reduced sample space consisting of only those outcomes in the event F. In a sense, all probabilities are conditional since even "unconditional" probabilities are relative to the sample space S, and it is only by custom that we write $P(E)$ instead of the equivalent $P(E|S)$.

The definition (1.7) is useful when the quantities on the right-hand side are known; we shall make frequent use of it in a different form, though, when the conditional probability is given and the composite probability $P(E \cap F)$ is sought:

(1.8) *General multiplication:* For any events E and F in S,

$$P(E \cap F) = P(F)P(E|F).$$

Note we need not specify $P(F) > 0$ here, for if $P(F) = 0$ then $P(E \cap F) = 0$ and both sides are zero regardless of what value might be specified for $P(E|F)$. We can see that (1.7) and (1.8) relate three quantities, any two of which determine the third. The third version of this relationship (namely $P(F) = P(E \cap F)/P(E|F)$) is seldom useful.

Sometimes knowing that F has occurred has no effect upon the specification of the probability of E:

(1.9) *Independent events:* We say events E and F in S are independent if $P(E) = P(E|F)$.

By simple manipulation using the previous rules, this can be expressed in two other equivalent ways.

(1.10) *Independent events:* If E and F are independent then $P(E|F) = P(E|F^c)$.

(1.11) *Multiplication with independent events:* If E and F are independent, then

$$P(E \cap F) = P(E) \cdot P(F).$$

Indeed, this latter condition (which is not to be confused with the almost opposite notion of "mutually exclusive") is often taken as the definition of independence.

Note that independence (unlike, for example, being mutually exclusive) depends crucially upon the values specified for the probabilities. In the previous example of the die, E and F are independent for the given specification of probabilities. Using (1.9),

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \left(\frac{1/6}{2/6}\right) = \frac{1}{2} = P(E).$$

Alternatively, using (1.11),

$$P(E \cap F) = \frac{1}{6} = \left(\frac{3}{6}\right)\left(\frac{2}{6}\right) = P(E)P(F).$$

However, if the die were strongly weighted and $P(\{1\}) = P(\{2\}) = P(\{4\}) = \frac{1}{3}$, then $P(F) = \frac{2}{3}$. $P(E) = \frac{1}{3}$, and $P(E \cap F) = \frac{1}{3}$, so for this specification of probabilities E and F are then *not* independent.

Example 1.B: A Random Star. Early astronomers noticed many patterns in the heavens; one which caught the attention of mathematicians in the eighteenth and nineteenth centuries was the occurrence of six bright stars (the constellation of the Pleiades) within a small section of the celestial sphere 1° square. How likely, they asked, would such a tight grouping be if the stars were distributed at random in the sky? Could the occurrence of such a tight cluster be taken as evidence that a common cause, such as gravitational attraction, tied the six stars together? This turns out to be an extraordinarily difficult question to formulate, much less answer, but a simpler question can be addressed even with the few rules we have introduced, namely: Let A be a given area on the surface of the celestial sphere that is a square, 1° on a side. A single star is placed randomly on the sphere. What is the probability it lands in A? A solution requires that we specify what "placed randomly" means mathematically. Here the sample space S is infinite, namely the points on the celestial sphere. Specifying probabilities on such a set can be challenging. We give two different solutions.

First Solution: The star is placed by specifying a latitude and a longitude. By "placed randomly" we may mean that the latitude and longitude are picked independently, the latitude in the range $-90°$ to $90°$ with a probability of $1/180$ attached to each 1° interval, and the longitude in the range 0° to 360° with a probability of $1/360$ attached to each 1° interval. This is not a full specification of the probabilities of the points on the sphere, but it is sufficient for the present purpose. Suppose A is located at the equator (Figure 1.3). Let

E = "pick latitude in A's range"

F = "pick longitude in A's range"

A = "pick a point within A".

Then $A = E \cap F$, $P(E) = 1/180$, and by independence $P(F|E) = P(F) = 1/360$ and so

$$\begin{aligned} P(A) &= P(E) \cdot P(F) \\ &= \frac{1}{180} \cdot \frac{1}{360} \\ &= \frac{1}{64800}. \end{aligned}$$

Second Solution: We may ask how many 1° squares make up the area of the sphere; if all are equally likely, the probability of A is just the reciprocal of this number. It has been known since Archimedes that if a sphere has radius r, the area of the surface is $4\pi r^2$. (This can be easily remembered as following from the "Orange Theorem": if a spherical orange is sliced into four quarters then for each quarter the area of the two flat juicy sides equals that of the peel. The flat juicy sides are two semicircles of total area πr^2, so the peel of a quarter orange has area πr^2 and the whole peel is $4\pi r^2$.) Now if the units of area are to be square degrees, then since the circumference is $2\pi r$ we will need to choose r so that $2\pi r = 360^\circ$, or $r = \frac{360}{2\pi}$. Then the area of the surface is

$$4\pi r^2 = 4\pi \left(\frac{360}{2\pi}\right)^2 = \frac{360^2}{\pi} = 41253.$$

Each square degree being supposed equally likely, we have

$$P(A) = \frac{1}{41253},$$

which is $\pi/2$ times larger than the first solution.

Both solutions are correct; they are based upon different hypotheses. The hypotheses of the first solution may well be characterized as "placed at random" from one point of view, but they will make it more likely that a square degree near a pole contains the star than that one on the equator does.

The original problem is more difficult than the one we solved because we need to ask, if the approximately 1500 bright stars (of 5th magnitude or brighter) are placed "randomly" and independently on the sphere, and we search out the square degree containing the largest number of stars, what is the chance it contains six or more? And even this already complicated question glosses over the fact that our interest in a section 1° square (rather than 2° square, or 1° triangular, etc.) was determined a posteriori, after looking at the data. We shall discuss some aspects of this problem in later chapters.

1.3 Counting

When the sample space S is finite and the outcomes are specified to be equally likely, the calculation of probabilities becomes an exercise in counting: $P(E)$ is simply the number of outcomes in E divided by the number of outcomes in S. Nevertheless, counting can be difficult. Indeed, an entire branch of mathematics, combinatorics, is devoted to counting. We will require two rules for counting, namely those for determining the numbers of *permutations* and of *combinations* of n distinguishable objects taken r at a time. These are

(1.12) The number of ways of choosing r objects from n distinguishable objects where the order of choice makes a difference is the number of *permutations* of n choose r, given by

$$P_{r,n} = \frac{n!}{(n-r)!}.$$

(1.13) The number of ways of choosing r objects from n distinguishable objects where the order of choice does not make a difference is the number of *combinations* of n choose r, given by

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

In both cases, $n!$ denotes *n factorial,* defined by $n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$ for integer $n > 0$, and we take $0! = 1$ for convenience. Thus we have also

$$P_{r,n} = \frac{1 \cdot 2 \cdot 3 \cdot \cdots \cdot n}{1 \cdot 2 \cdot \cdots \cdot (n-r)} = (n-r+1) \cdots (n-1)n$$

and

$$\binom{n}{r} = \frac{P_{r,n}}{r!} = \frac{(n-r+1) \cdots (n-1)n}{1 \cdot 2 \cdot 3 \cdots (r-1)r}.$$

A variety of identities can be established using these definitions; some easily (eg. $\binom{n}{0} = 1$, $\binom{n}{1} = n$, $\binom{n}{r} = \binom{n}{n-r}$), others with more difficulty (eg. $\sum_{r=0}^{n} \binom{n}{r} = 2^n$, which can, however, be directly established by noting the lefthand side gives the number of ways *any* selection can be made from n objects without regard to order, which is just the number of subsets, or 2^n).

Example 1.C: If $r = 2$ people are to be selected from $n = 5$ to be designated president and vice president respectively, there are $P_{2,5} = 20$ ways the selection can be made. If, however, they are to serve as a committee of two equals (so the committees (A, B) and (B, A) are the same committee), then there are only $\binom{5}{2} = 10$ ways the selection can be made.

Example 1.D: For an example of a more important type, we could ask how many binary numbers of length n (= 15, say) are there with exactly r (= 8, say) l's. That is, how many possible sequences of 15 0's and 1's are there for which the sum of the sequence is 8? The answer is $\binom{n}{r} = \binom{15}{8} = 6435$, as may be easily seen by considering the sequence as a succession of $n = 15$ distinguishable numbered spaces, and the problem as one of selecting $r = 8$ from those 15 spaces as the locations for the 8 1's, the order of the 1's being unimportant and the remaining unfilled slots to be filled in by 0's.

1.4 Stirling's Formula

Evaluating $n!$, $P_{r,n}$, or $\binom{n}{r}$ can be quite difficult if n is at all large. It is also usually unnecessary, due to a very close approximation discovered about 1730 by James Stirling and Abraham De Moivre. *Stirling's formula* states that

$$\log_e(n!) \sim \frac{1}{2}\log_e(2\pi) + \left(n + \frac{1}{2}\right)\log_e(n) - n, \tag{1.14}$$

and thus

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}, \tag{1.15}$$

where "$\sim$" means that the ratio of the two sides tends to 1 as n increases. The approximations can be good for even small n, as Table 1.1 shows.

[Table 1.1]

Stirling's formula can be used to derive approximations to $P_{r,n}$ and $\binom{n}{r}$, namely

$$P_{r,n} \sim \left(1 - \frac{r}{n}\right)^{-\left(n+\frac{1}{2}\right)} (n-r)^r e^{-r}, \tag{1.16}$$

and

$$\binom{n}{r} \sim \frac{1}{\sqrt{2\pi n}} \left(1 - \frac{r}{n}\right)^{-\left(n-r+\frac{1}{2}\right)} \left(\frac{r}{n}\right)^{-\left(r+\frac{1}{2}\right)}. \tag{1.17}$$

These too are reasonably accurate approximations; for example $\binom{10}{5} = 252$, while the approximation gives 258.37. While not needed for most purposes, there are more accurate refinements available. For example, the bounds

$$\sqrt{2\pi}\ n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n+1}} < n! < \sqrt{2\pi}\ n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}} \tag{1.18}$$

give, for $n = 5$, $119.9699 < n! = 120 < 120.0026$. Feller (1957, Chapter II.9) gives a proof of Stirling's formula and a nice discussion.

1.5 Random Variables

Sometimes the outcomes of an experiment are expressed as numbers, and at other times we will be most interested in numerical descriptions that capture only some aspects of the outcomes, even in situations where we find it easiest to specify the probabilities of the outcomes themselves. We will use the term *random variables* for such a description: a function that assigns a numerical value to each outcome in S; a real-valued function defined on S.

Example 1.E: If a coin is tossed three times, the sample space might be described by a list of 8 three-letter words,

$$S = \{TTT, TTH, THT, HTT, HHT, HTH, THH, HHH\},$$

where HHT means that the first two tosses result in Heads, and the third in Tails. One possible random variable is

$$X = \#H\text{'s in the word.}$$

Another is

$$Y = \#T\text{'s in the word.}$$

In both cases, the possible values are 0, 1, 2, and 3.

We will call random variables whose values can be listed sequentially in this manner *discrete* random variables. In such cases, once the probabilities of the values of the random variable have been specified, they can be described rather simply, by listing them. A list of the possible values of a discrete random variable together with the probabilities of these values is called the *probability distribution* of the random variables; we shall denote the probability that the random variable X is equal to the possible value x by $p_X(x)$, or, when there is no likely confusion, by $p(x)$.

For the coin example, the specification of the probability distribution of the random variable $X = \#H$'s can be straightforward. If we assume the coin is "fair" (which we may take to mean that $P(H) = 1/2$ for a single toss), and the tosses are independent, then applying the multiplication rule for independent events (1.11) repeatedly gives us, for example,

$$P(HHT) = P(H) \cdot P(H) \cdot P(T) = \left(\frac{1}{2}\right)^3 = \frac{1}{8},$$

and so the 8 points in S are equally likely. Now the event "$X = 1$" consists of the outcomes $\{HTT, THT, TTH\}$ and by the additivity rule (1.5) it has probability

$$\begin{aligned} p_X(1) &= P(X = 1) \\ &= P(\{HTT, THT, TTH\}) \\ &= P(HTT) + P(THT) + P(TTH) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}. \end{aligned}$$

The full probability distribution is easily found:

x	0	1	2	3
$p_X(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The same probability distribution can be described in many ways. For example, the above table can be compressed into a formula: $p_X(x) = \binom{3}{x}(1/2)^3$, for $x = 0, 1, 2, 3$. We will find an alternative description convenient later, what we will call the *cumulative distribution function,* sometimes abbreviated c.d.f. and denoted $F_X(x)$ (or $F(x)$), defined by

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= \sum_{\substack{\text{all values} \\ a \le x}} p_X(a). \end{aligned}$$

It is often helpful to think of a probability distribution of a random variable as a distribution of a unit mass along the real line, with $p_X(x)$ giving the mass assigned to the point x. Then $F_X(x)$ gives the cumulative mass, starting from the left, up to and including that at the point x.

For the coin example the calculation is simple:

x	0	1	2	3
$p_X(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
$F_X(x)$	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{7}{8}$	$\frac{8}{8} = 1$

Graphically, we can depict $p_X(x)$ as a system of spikes, and then $F_X(x)$ is a jump function with jumps at each possible value of x.

[Figure 1.4]

Note that the probability distribution $p_X(x)$ can be recovered from $F_X(x)$ by evaluating the sizes and locations of the jumps. If $x-1$ and x are two consecutive possible values of X, then

$$p_X(x) = F_X(x) - F_X(x-1).$$

The two alternative specifications are thus completely equivalent, given one the other can be found, and we can choose between them on the grounds of convenience.

1.6 Binomial Experiments

The experiment of tossing a fair coin three times is a special case of a broad class of experiments of immense use in statistics, the class of *binomial experiments*. These are characterized by three conditions:

(1.19) (a) The experiment consists of a series of n independent trials.
(b) The possible outcomes of a single trial are classified as being of one of two types:

A, called "success"

A^c, called "failure".

(c) The probability of success on a single trial, $P(A) = \theta$, is the same for all n trials (and so, of course, is $P(A^c) = 1 - \theta$). This probability θ is called the *parameter* of the experiment.

For the fair coin example, $n = 3$, $A = H$, and $\theta = 1/2$.

The sample space S of the possible outcomes of a binomial experiment consists of a list of "words" of length n, made up entirely of the "letters" A and A^c. These range from all successes to all failures:

n times			
AA	$\cdots$	AA	all successes
AA	$\cdots$	AA^c	
AA	$\cdots$	A^cA	
.		.	
.		.	
AA^c	$\cdots$	AA^c	
.		.	
.		.	
.		.	
.		.	
A^cA^c	$\cdots$	A^cA^c	all failures

There are 2 choices for each letter and thus $2 \times 2 \times \cdots \times 2 = 2^n$ different such words. For the coin example, $n = 3$ and there are $2^3 = 8$ outcomes.

Since the trials are independent by hypothesis, it is easy to compute the probability of a single outcome using the multiplication rule for independent events (1.11). For example:

For $n = 2$:

$$P(AA) = P(A) \cdot P(A) = \theta \cdot \theta = \theta^2$$
$$P(AA^c) = P(A)P(A^c) = \theta(1-\theta).$$

For $n = 3$:

$$P(AA^cA) = P(A)P(A^c)P(A)$$
$$= \theta(1-\theta)\theta = \theta^2(1-\theta).$$

In general, the probability of an outcome will only depend upon the numbers of A's and A^c's in the outcome. If the word

$$AA^c \ldots AA^c$$

consists of

$$\begin{array}{ccc} x & A\text{'s} & \text{and} \\ n-x & A^c\text{'s} & \end{array}$$

then

$$P(AA^c \cdots AA^c) = P(A)P(A^c) \cdots P(A)P(A^c)$$
$$= \theta(1-\theta) \cdots \theta(1-\theta)$$
$$= \theta^x(1-\theta)^{n-x}.$$

Now, for binomial experiments we will frequently only be interested in a numerical summary of the outcome, the random variable

$$X = \#\text{ successes} = \#A\text{'s}.$$

The possible values of X are $0, 1, 2, \ldots, n$, and its probability distribution can be found as follows: The event "$X = x$" consists of exactly those outcomes with x A's and $n - x$ A^c's. We have just found that each such outcome has probability $\theta^x(1-\theta)^{n-x}$. It remains only to determine the number, say C, of outcomes with xA's and $n - x$ A^c's; the desired probability will then be $p_X(x) = P(X = x) = C \cdot \theta^x(1-\theta)^{n-x}$. But C is equal to the number of binary numbers of length n with exactly x 1's and $n - x$ 0's (just think of each A as "1" and each A^c as "0"), and we have found (in Example 1.D) that this number is $\binom{n}{x}$. Therefore $p_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. This probability distribution is called the *Binomial Distribution,* and is sometimes given a special symbol that shows its dependence upon n and θ explicitly:

(1.20) *The Binomial (n, θ) Distribution*

$$b(x; n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n$$
$$= 0 \quad \text{otherwise.}$$

Figure 1.5 illustrates some examples, for $n = 8$. The *parameters* n and θ determine the distribution; for each integer $n \geq 1$ and $0 \leq \theta \leq 1$, we have a different distribution. The Binomial distribution is thus an example of what is called a *parametric family* of probability distributions.

The trials that make up a binomial experiment are often called *Bernoulli trials*, after the Swiss mathematician Jacob Bernoulli (1654–1705) who was instrumental in the early study of this experiment. Bernoulli trials can be conducted in manners other than that we have discussed; the most important of these is where rather than conduct a fixed number n of trials, the trials are conducted until a fixed number of successes r have been observed. Because this is a sort of reversal of the original scheme, it is called the *negative binomial experiment.* For example, if $r = 1$, the trials are conducted until the first success, and the sample space consists of "words" of increasing length with a single A at the end: A, A^cA, A^cA^cA, $A^cA^cA^cA$, etc.

For this experiment the random variable of most interest is

$$Z = \#\text{ ``failures'' before the } r^{th} \text{ ``success''}.$$

For $r = 1$ the probability distribution of Z is easy to compute. For example, we will have $Z = 3$ only for the outcome $A^cA^cA^cA$, and since $P(A^cA^cA^cA) = (1-\theta)^3\theta$ we have $p_Z(3) = (1-\theta)^3\theta$. More generally, $p_Z(z) = (1-\theta)^z\theta$ for $z = 0, 1, 2, \ldots$. Note that Z is a discrete random variable with a countably infinite number of possible values.

To find the probability distribution of Z in general, we can reason analogously to the way we found the binomial distribution. The sample space S will consist of words with r A's, each word ending with an A (since the experiment terminates with the r^{th} success). The outcomes corresponding to $Z = z$ will be those with r A's and z A^c's, and each of these will have probability $\theta^r(1-\theta)^z$. To find the probability distribution of Z we need only find the number C of outcomes in S with $Z = z$; for then $p_Z(z) = C\theta^r(1-\theta)^z$. But C is the number of "words" of length $r+z$ ending in A, with exactly z A^c's. This is the same as the number of "words" of length $r+z-1$ with exactly z A^c's and no restrictions on the last letter, namely $C = \binom{r+z-1}{z} = \binom{r+z-1}{r-1}$. We have therefore found the

(1.21) *Negative Binomial Distribution:* The probability distribution of the number of failures Z before the r^{th} success in a series of Bernoulli trials with probability of success θ is

$$\begin{aligned} nb(z : r, \theta) &= \binom{r+z-1}{r-1}\theta^r(1-\theta)^z \quad \text{for } z = 0, 1, 2, \ldots \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

This distribution is sometimes called the *Pascal distribution,* after an early programming language that will continue to compile until the first bug is encountered. The special case where $r = 1$, namely $p_Z(z) = \theta(1-\theta)^z$, for $z = 0, 1, 2, \ldots$, is called the *geometric distribution.* We shall later see that the negative binomial distribution has a close relationship to another important discrete distribution, the Poisson distribution.

[Figure 1.6]

The Binomial and Negative Binomial distributions are more closely related than the fact that both involve Bernoulli trials. Let

$$B(x; n, \theta) = P(X \le x)$$

$$NB(z; r, \theta) = P(Z \le z)$$

be their respective cumulative distribution functions. Then a little reflection tells us that if we were computing X and Z from the same series of trials, we would have $X \ge r$ if and only if $Z \le n - r$. Since $P(X \ge r) = 1 - P(X \le r-1)$, this means

$$NB(n-r; r, \theta) = 1 - B(r-1; n, \theta), \tag{1.22}$$

and so one set of probabilities can be computed from the other. For example, tables of the binomial distribution can be used to find Negative Binomial probabilities. The Binomial distribution enjoys certain symmetry properties. In particular

$$B(x; n, \theta) = 1 - B(n-x-1; n, 1-\theta). \tag{1.23}$$

This relation allows the computation of Binomial (and hence Negative Binomial) probabilities using a table of the Binomial distribution for $0 < \theta \le \frac{1}{2}$.

1.7 Continuous Distributions

A random variable is called *continuous* if its possible values form an interval, and hence cannot be listed sequentially.

Example 1.F: Consider the spinner, a disc with a pointer rotating freely around the center point, pointing at the edge of the disc which is labelled continuously from a to b.

[Figure 1.7]

If the pointer is spun and allowed to come to rest at an (in some sense random) point X, then the sample space S is the interval $\{x : a \leq x < b\}$ and X is a random variable whose possible values are the numbers in this interval.

Because the values of a continuous random variable cannot be listed, their probabilities cannot be listed, and another device is used to describe the probability distribution. In a direct extension to the interpretation of discrete probability distributions as mass distributions, continuous probability distributions will be described by *probability density functions,* nonnegative functions which give the probabilities of an interval through the area under the function over the interval. Mathematically, since areas are given by integrals, we will define $f_X(x)$ (or $f(x)$ if no confusion arises) to be the *probability density function* of the continuous random variable X if for any numbers c and d, with $c < d$,

$$P(c < X \leq d) = \int_c^d f_X(x)dx.$$

[Figure 1.8]

It will necessarily be true of probability density functions that

(i) $f_X(x) \geq 0$ for all x

and

(ii) $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Indeed, any function satisfying (i) and (ii) may be considered as the probability density of a continuous random variable. Note that the values of $f_X(x)$ do not themselves give probabilities (they may even exceed 1), though we can think heuristically of $f_X(x)dx$ (= height $f_X(x)$ times base dx) as the probability X falls in an infinitesimal interval at x:

$$P(x < X \leq x + dx) = f_X(x)dx$$

[Figure 1.9]

It is frequently helpful to think of the density function $f_X(x)$ as describing the upper boundary of a sheet of unit mass resting upon the line of possible values, the area under that boundary over an interval being equal to the mass over that interval.

One consequence of using probability densities to describe distributions is that individual points are assigned probability zero:

$$P(X = c) = 0 \quad \text{for any } c.$$

The area or mass exactly over each single point must be considered to be zero, or contradictions would ensue, as we shall see. As a consequence, for continuous random variables we have, for any $c < d$,

$$\begin{aligned} P(c < X \leq d) &= P(c < X < d) \\ &= P(c \leq X < d) \\ &= P(c \leq X \leq d), \end{aligned}$$

since, for example,

$$P(c < X \leq d) = P(c < X < d) + P(X = d)$$

by the additivity rule (2).

Example 1.F (Continued): To illustrate probability density functions, and to better understand the apparent paradox that $P(X = c) = 0$ for all c (all values are "impossible") yet $P(-\infty < X < \infty) = 1$ (some value is "certain"), consider the spinner. We may wish to capture the notion that all values in the interval $[a, b]$ are "equally likely" in some sense; we do that by taking the probability of any subinterval as proportional to the length of the subinterval. This is true if we adopt as the probability density of X the *Uniform* (a, b) or *Rectangular distribution:*

$$\begin{aligned} f_X(x) &= \frac{1}{(b-a)} \quad \text{for } a \leq x < b \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{1.25}$$

[Figure 1.10]

Clearly the total area under $f_X(x)$ is 1, and the area or probability over any subinterval (c, d) is $(d-c)/(b-a)$, proportional to the length $d - c$ of the subinterval. The numbers a and b are the parameters of this distribution. If we ask what probability could be assigned to any single number c, we see it must be smaller than that assigned to the interval $c \leq x < c + \epsilon$, for any $\epsilon > 0$, that is, smaller than $P(c \leq X < c + \epsilon) = \epsilon/(b - a)$. But no positive number fits that description, and we are forced by the limitations of our number system to take $P(X = c) = 0$. This will not cause difficulties or contradictions as long as we follow our rules and only insist that probabilities be countably additive: having the probability $P(a \leq X < b) = 1$, yet each $P(X = c) = 0$ does not contradict the additivity rule (1.6) since there are uncountably many c's between a and b.

Similarly to the discrete case, we define the *cumulative distribution function* of a continuous random variable X by

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \int_{-\infty}^{x} f_X(u)du \end{aligned}$$

$F_X(x)$ is thus the area under $f_X(x)$ to the left of x. As before, $F_X(x)$ is a nondecreasing

[Figure 1.11]

function, though it is no longer a jump function. It gives an alternative way to describe continuous distributions. The fundamental theorem of calculus holds that

$$\frac{d}{dx} \int_{-\infty}^{x} f_X(u)du = f_X(x),$$

so

$$\frac{d}{dx} F_X(x) = f_X(x)$$

and we may find the probability density function from the cumulative distribution function as well as vice versa.

Example 1.G: The Exponential Distribution. Consider the experiment of burning a lightbulb until failure. Let X be the time until failure; X may be considered a continuous random variable with possible values $\{x : 0 \leq x < \infty\}$. In order to specify a class of possible probability distributions for X, we would expect to have the probability of survival beyond time t, $P(X > t)$, decreasing as $t \to \infty$. One class of decreasing functions which also have $P(X > 0) = 1$ are the exponentially decreasing functions $P(X > t) = C^t$, where $0 < C < 1$. Equivalently, writing $e^{-\theta t}$ for C, where $\theta > 0$ is a fixed parameter, we have

$$P(X > t) = e^{-\theta t} \quad \text{for } t \geq 0$$

and

$$F_X(t) = P(X \le t) = 1 - e^{-\theta t} \quad \text{for } t \ge 0$$
$$= 0 \quad \text{for } t < 0.$$

The corresponding probability density function is found by differentiation:

$$f_X(t) = \theta e^{-\theta t} \quad \text{for } t \ge 0$$
$$= 0 \quad \text{for } t < 0. \tag{1.27}$$

This is called the *Exponential* (θ) *distribution;* θ is a parameter of the distribution (for each $\theta > 0$ we get a different distribution). When we come to discuss the Poisson process we shall see how the Exponential distribution arises in that context from more natural and less arbitrary assumptions as a common failure time distribution or waiting time distribution.

[Figure 1.12]

1.8 Transformations of Random Variables

In the calculus of probabilities, it is common to specify a probability distribution for a situation where that task is simple, and to then reason to a more complicated one. We may choose to describe the sample space so that the outcomes are equally likely and then deduce the distribution of a random variable whose possible values are not equally likely. Or we may use ideas of conditional probability to break a complicated framework down into a number of simple steps, perhaps even into independent trials. In the case of the binomial experiment, both devices were adopted. Another common route, one which will be particularly useful in statistical applications, is to go from the distribution of one random variable, say X, whose distribution is easily specified or has previously been determined, to that of another random variable which is a transformation or function of X, say $Y = h(X)$.

Example 1.H Suppose X is the time to failure of a lightbulb (or an electronic component), and that we believe X to have an Exponential (θ) distribution with density (1.27),

$$f_X(x) = \theta e^{-\theta x} \quad x \ge 0$$
$$= 0 \quad x < 0.$$

Upon failure, we plan to replace the lightbulb with a second, similar one. The probability that the first survives beyond time t is $P(X > t) = e^{-\theta t}$; the probability the second survives longer than the first is then $Y = e^{-\theta X}$, a random variable that depends upon the time to failure of the first. What is the distribution of Y?

Example 1.I: Suppose a fair coin is tossed three times, and you received \$2 for every Head. The number of Heads, X, has a Binomial distribution, what is the distribution of your winnings, $Y = 2X$? Or what if you receive $Y = X^2$?

To begin to address the general question of finding the distribution of a transformation $Y = h(X)$ of a random variable X, consider first the case where h is a strictly monotone function, at least over the range of possible values of X. This restriction will ensure that each value of Y could have come from only one possible X, and the ideas will be easier to explain in that case. For example, $h(X) = 2X + 3$ is strictly monotone, while $h(X) = X^2$ is not, although it will be allowed in the present discussion if X takes no negative values, since it is strictly monotone for nonnegative x.

Example 1.J: The extremely useful transformation $h(X) = \log_e(X)$ is strictly monotone, though only defined for $X > 0$. We can see what happens to a probability distribution under transformation by looking at this one special case. Figure 1.14 illustrates the effect of this transformation upon the X-scale: it compresses the upper end of the scale by pulling large values down, while spreading out the scale for small values. The

gap between X's of 5 and 6 (namely, 1 X-unit) is narrowed to that between Y's of 1.61 and 1.79 (.18 Y-units), and the gap between X's of .2 and 1.2 (also 1 X-unit) is expanded to that between Y's of -1.61 and .18 (1.79 Y-units). Figure 1.15 illustrates the effect of this transformation upon two probability distributions, one discrete and one continuous. The effect in the discrete case is particularly easy to describe: as the scale is warped by the transformation, the locations of the spikes are changed accordingly, but their heights remain unchanged. In the continuous case, something different occurs. Since the total area that was between 5 and 6 on the X-scale must now fit between 1.61 and 1.79 on the Y-scale, the height of the density over this part of the Y-scale must be increased. Similarly, the height of the density must be decreased over the part of the Y-scale where the scale is being expanded, to preserve areas there. The result is a dramatic change in the appearance of the density. Our object in the remainder of this section is to describe precisely how this can be done.

If $Y = h(X)$ is a strictly monotone transformation of X, then we can solve for X in terms of Y, that is, find the *inverse transformation* $X = g(Y)$. Given $Y = y$, the function g "looks back" to see which possible value x of X produced that value y; it was $x = g(y)$. If $Y = h(X) = 2X + 3$, then $X = g(Y) = (Y - 3)/2$. If $Y = h(X) = \log_e(X)$, for, $X > 0$, then $X = g(Y) = e^Y$. If $Y = h(X) = X^2$, for $X > 0$, then $X = g(Y) = +\sqrt{Y}$.

In terms of this inverse relationship, the solution for the discrete case is immediate. If $p_X(x)$ is the probability distribution function of X, then the probability distribution function of Y is

$$\begin{aligned} p_Y(y) &= P(Y = y) \\ &= P(h(X) = y) \\ &= P(X = g(y)) \\ &= p_X(g(y)). \end{aligned} \tag{1.28}$$

That is, for each value y of Y, simply "look back" to find the x that produced y, namely $x = g(y)$, and assign y the same probability that had been previously assigned to that x, namely $p_X(g(y))$.

Example 1.K: If X has the Binomial distribution of Example 1.E; that is, $p_X(x) = \binom{3}{x}(0.5)^3$ for $x = 0, 1, 2, 3$, and $Y = X^2$, what is the distribution of Y? Here $g(y) = +\sqrt{y}$, and so $p_Y(y) = p_X(\sqrt{y}) = \binom{3}{\sqrt{y}}(0.5)^3$ for $\sqrt{y} = 0, 1, 2, 3$ (or $y = 0, 1, 4, 9$). For all other y's, $p_Y(y) = p_X(\sqrt{y}) = 0$. That is,

$$\begin{aligned} p_Y(y) &= \frac{1}{8} \quad \text{for } y = 0 \\ &= \frac{3}{8} \quad \text{for } y = 1 \\ &= \frac{3}{8} \quad \text{for } y = 4 \\ &= \frac{1}{8} \quad \text{for } y = 9 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

[Figure 1.16]

In the continuous case, an additional step is required, the rescaling of the density to compensate for the compression or expansion of the scale and match corresponding areas. For this reason it is not true that $f_Y(y) = f_X(g(y))$, where g is the inverse transformation, but instead

$$f_Y(y) = f_X(g(y)) \cdot \left| \frac{dg(y)}{dy} \right|. \tag{1.29}$$

The rescaling factor $\left|\frac{dg(y)}{dy}\right| = |g'(y)|$ is called the *Jacobian* of the transformation in advanced calculus, and it is precisely the compensation factor needed to match areas. When $|g'(y)|$ is small, $x = g(y)$ is changing

slowly as y changes (for example, for y near 0 in Figures 1.14 and 1.15), and we scale down. When $g(y)$ changes rapidly with y, $|g'(y)|$ is large (for example, for y near 6 in Figures 1.14 and 1.15), and we scale up. It is easy to verify that this is the correct factor: simply compute $P(Y \leq a)$ in two different ways.

First,

$$P(Y \leq a) = \int_{-\infty}^{a} f_Y(y)dy, \tag{1.30}$$

by the definition of $f_Y(y)$. Second, supposing for a moment that $h(x)$ (and hence $g(y)$ also), is monotone *increasing,* we have

$$\begin{aligned} P(Y \leq a) &= P(h(X) \leq a) \\ &= P(X \leq g(a)) \\ &= \int_{-\infty}^{g(a)} f_X(x)dx. \end{aligned}$$

Now, making the change of variables, $x = g(y)$, and $dx = g'(y)dy$, we have

$$P(Y \leq a) = \int_{-\infty}^{a} f_X(g(y))g'(y)dy. \tag{1.31}$$

Differentiating both (1.30) and (1.31) with respect to a gives $f_Y(y) = f_X(g(y))g'(y)$. If $h(x)$ and $g(y)$ are monotone *decreasing,* the result is the same, but with $-g'(y)$ as the compensation factor; the factor $|g'(y)|$ covers both cases, and gives us (1.29).

Example 1.H (Continued). Let X be the time to failure of the first lightbulb, and Y the probability that the second bulb burns longer than the first. Y depends on X, and is given by $Y = h(X) = e^{-\theta X}$. The random time X has density

$$\begin{aligned} f_X(x) &= \theta e^{-\theta x} \quad x \geq 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Now $\log_e(Y) = -\theta X$, and the inverse transformation is $X = g(Y) = -\log_e(Y)/\theta$. Both $h(x)$ and $g(y)$ are monotone decreasing. The inverse $g(y)$ is only defined for $0 < y$, but the only possible values of Y are $0 < y \leq 1$.

We find $g'(y) = -\frac{1}{\theta} \cdot \frac{1}{y}$, and

$$|g'(y)| = \frac{1}{\theta y}, \quad \text{for } y > 0.$$

Then

$$f_Y(y) = f_X(g(y))|g'(y)|,$$

and, noting that $f_X(g(y)) = 0$ for $y \leq 0$ or $y > 1$, we have

$$\begin{aligned} f_Y(y) &= \theta e^{-\theta(-\log(y)/\theta)} \cdot \frac{1}{\theta y} \quad \text{for } 0 < y \leq 1 \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

or, since $\theta e^{-\theta(-\log(y)/\theta)} = \theta y$,

$$\begin{aligned} f_Y(y) &= 1 \quad \text{for } 0 < y \leq 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

We recognize this as the *Uniform* $(0, 1)$ *distribution,* formula (1.25) of Example F, with $a = 0$, $b = 1$.

Example 1.L. The Probability Integral Transformation. The simple answer we obtained in Example 1.H, namely that if X has an Exponential (θ) distribution, $Y = e^{-\theta X}$ has a Uniform $(0, 1)$ distribution, is an example of a class of results that are useful in theoretical statistics. If X is *any* continuous random variable with cumulative distribution function $F_X(x)$, then both transformations $Y = F_X(X)$ and $Z = 1 - F_X(X)$

$(= 1-Y)$ have Uniform $(0,1)$ distributions. Example 1.H concerned Z for the special case of the Exponential (θ) distribution. Because $F_X(x)$ is the integral of the probability density function $f_X(x)$, the transformation $h(x) = F_X(x)$ has been called the *probability integral transformation.* To find the distribution of $Y = h(X)$, we need to differentiate $g(y) = F_X^{-1}(y)$, defined to be the *inverse cumulative distribution function*, the function that for each y, $0 < y < 1$, gives the value of x for which $F_X(x) = y$. [Figure 1.17] (For continuous random variables X with densities, $F_X(x)$ is continuous and such an x will exist for all $0 < y < 1$. For more general random variables, $F_X^{-1}(y)$ can be defined as $F_X^{-1}(y) = \textit{infimum}\{x : F_X(x) \geq y\}$.) The derivative of $g(y) = F_X^{-1}(y)$ can be found by implicit differentiation:

$$y = F_X(x)$$

so

$$1 = \frac{d}{dy}F_X(x) = f_X(x) \cdot \frac{dx}{dy}$$

by the chain rule, and so

$$\frac{dx}{dy} = \frac{1}{f_X(x)}$$

or, with $x = g(y) = F_X^{-1}(y)$,

$$g'(y) = \frac{d}{dy}F_X^{-1}(y) = \frac{1}{f_X(F_X^{-1}(y))}, \tag{1.32}$$

But then

$$\begin{aligned} f_Y(y) &= f_X(g(y)) \cdot |g'(y)| \\ &= f_X(F_X^{-1}(y)) \cdot \frac{1}{f_X(F_X^{-1}(y))} \quad \text{for } 0 < y < 1 \\ &= 1 \quad \text{for } 0 < y < 1 \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

the Uniform $(0,1)$ distribution. The fact that $Z = 1-Y$ also has this distribution can be shown by repeating this derivation with $h(x) = 1 - F_X(x)$, or more simply by transforming the distribution of Y by $h(y) = 1-y$, whose inverse $g(z) = 1 - z$ has $|g'(z)| = 1$.

Thus far we have considered only strictly monotone transformations $h(x)$. A full treatment for nonmonotone transformations is possible, but it is cumbersome, and not necessary for our anticipated applications. The following example, involving a transformation that can be broken down into two transformations monotone over different ranges, captures all of the important ideas for even more general cases.

Example 1.M. The Standard Normal and Chi-square (1 d.f.) distributions. Suppose X is a continuous random variable with probability density function $f_X(x)$ defined for all x, $-\infty < x < \infty$, by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}. \tag{1.33}$$

This distribution is called the *Standard Normal distribution,* and it is sufficiently important that the symbol $\phi(x)$ is reserved for its density and $\Phi(x)$ for its cumulative distribution function. [Figure 1.18]. The cumulative distribution function cannot be written in closed form in terms of simple functions, but it can be evaluated numerically and is tabled at the back of the book for $x \geq 0$. By the symmetry and continuity of the distribution, this range is sufficient, since

$$P(X \leq -x) = P(X \geq x) = 1 - P(X < x) = 1 - P(X \leq x),$$

or

$$F_X(-x) = 1 - F_X(x),$$

or

$$\Phi(-x) = 1 - \Phi(x) \quad \text{for all } x.$$

Consider the nonmonotone transformation of X, $Y = h(X) = X^2$. Because h is nonmonotone over the range of X, we cannot find a single inverse; rather we exploit the fact that $h(x)$ is separately monotone for negative and for positive values, and we can find two "inverses":

$$x = g_1(y) = -\sqrt{y} \quad \text{for } -\infty < x < 0$$

$$x = g_2(y) = +\sqrt{y} \quad \text{for } 0 < x < \infty.$$

The probability density of Y can then be found by following our earlier logic twice, once for each branch of the inverse, and adding the results to give

$$f_Y(y) = f_X(g_1(y)) \cdot |g_1'(y)| + f_X(g_2(y)) \cdot |g_2'(y))|. \tag{1.34}$$

Why does this work? In essence, for each $y > 0$ it recognizes that y could have come from either of two different x's, so we "look back" to both, namely $x = g_1(y)$ and $x = g_2(y)$. Heuristically, the probability appropriate to a small interval of width dy at y will be the sum of those found from the two separate branches (Figure 1.19).

For our example, the range of y is $y > 0$, and we find

$$f_X(g_1(y)) = \frac{1}{\sqrt{2\pi}} \frac{e^{-(-\sqrt{y})^2}}{2} = \frac{1}{\sqrt{2\pi}}\, e^{\frac{-y}{2}},$$

$$f_X(g_2(y)) = \frac{1}{\sqrt{2\pi}} \frac{e^{-(\sqrt{y})^2}}{2} = \frac{1}{\sqrt{2\pi}}\, e^{\frac{-y}{2}}$$

$$g_1'(y) = \frac{-1}{2\sqrt{y}}$$

and

$$g_2'(y) = \frac{1}{2\sqrt{y}},$$

so

$$|g_1'(y)| = |g_2'(y)| = \frac{1}{2\sqrt{y}},$$

and

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{\frac{-y}{2}} \cdot \frac{1}{2\sqrt{y}} + \frac{1}{\sqrt{2\pi}} e^{\frac{-y}{2}} \cdot \frac{1}{2\sqrt{y}} \quad \text{for } y > 0 \\ &= \frac{1}{\sqrt{2\pi y}} e^{\frac{-y}{2}} \quad \text{for } y > 0 \\ &= 0 \quad \text{for } y \le 0. \end{aligned} \tag{1.35}$$

We shall encounter this density later, it is called the *Chi-square distribution with 1 degree of freedom*, a name that will seem a bit less mysterious later.

Example 1.N. Linear change of scale. A common and mathematically simple example of a transformation of a random variable is a linear change of scale. The random variable X may be measured in inches; what is the distribution of $Y = 2.54X$, the same quantity measured in centimeters? Or if X is measured in degrees Fahrenheit, $Y = (X - 32°)/1.8$ is measured in degrees Celsius. The general situation has

$$Y = aX + b, \tag{1.36}$$

where a and b are constants. For any $a \neq 0$, $h(x) = ax + b$ is a monotone transformation, with inverse $g(y) = (y - b)/a$, $g'(y) = 1/a$, and

$$|g'(y)| = \frac{1}{|a|}.$$

We then have, for any continuous random variable X,

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{|a|}, \tag{1.37}$$

while for any discrete random variable X,

$$p_Y(y) = p_X\left(\frac{y-b}{a}\right). \tag{1.38}$$

Example 1.M (continued) The Normal (μ, σ^2) Distribution. A special case of Example 1.N will be of great use later, namely where X has a standard normal distribution, and Y is related to X by a linear change of scale

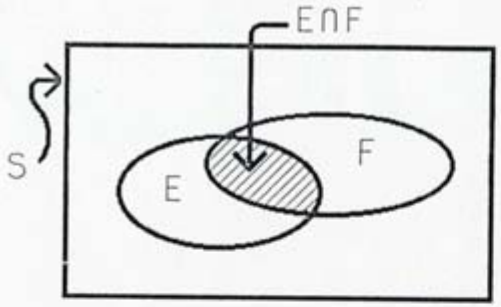
$$Y = \sigma X + \mu, \quad \sigma > 0. \tag{1.39}$$

Then Y has what we will call the *Normal (μ, σ^2) distribution* with density

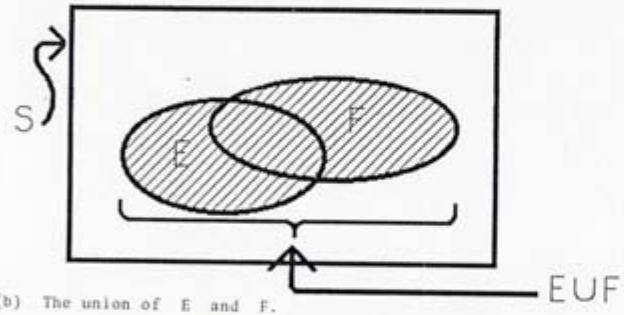
$$\begin{aligned} f_Y(y) &= \phi\left(\frac{y-\mu}{\sigma}\right) \cdot \frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad \text{for } -\infty < y < \infty. \end{aligned} \tag{1.40}$$

[Figure 1.20]

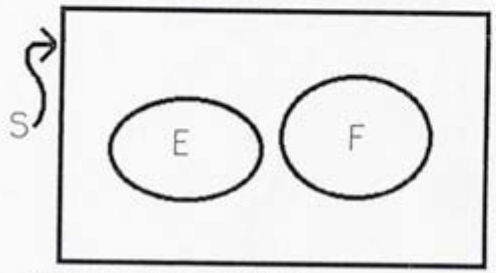
This might be called "general" Normal distribution, as a contrast to the "standard" Normal distribution. Actually, it is of course a parametric family of densities, with parameters μ and σ. When we encounter this family of distributions next we shall justify referring to μ as the *mean* and σ as the *standard deviation* of the distribution.



(a) The intersection of E and F.

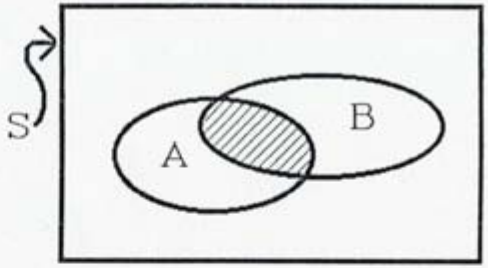


(b) The union of E and F.

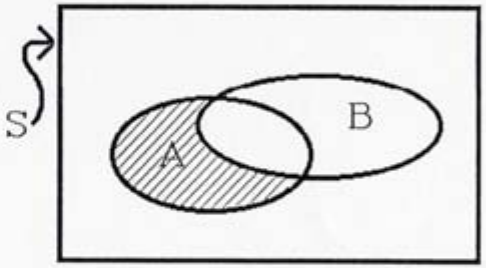


(c) Two mutually exclusive events.

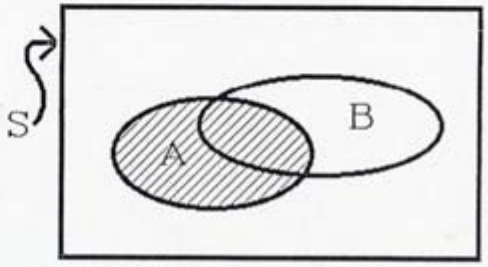
Figure 1.1



(a) $A \cap B$



(b) $A \cap B^c$



(c) $(A \cap B) \cup (A \cap B^c)$

Figure 1.2

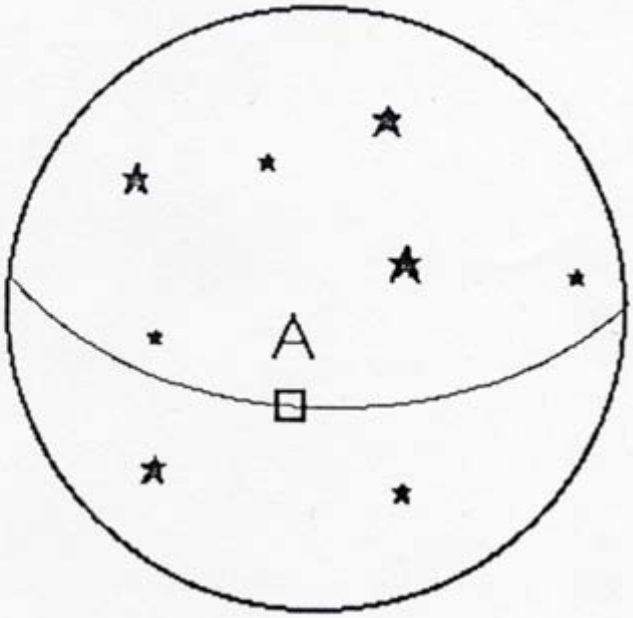
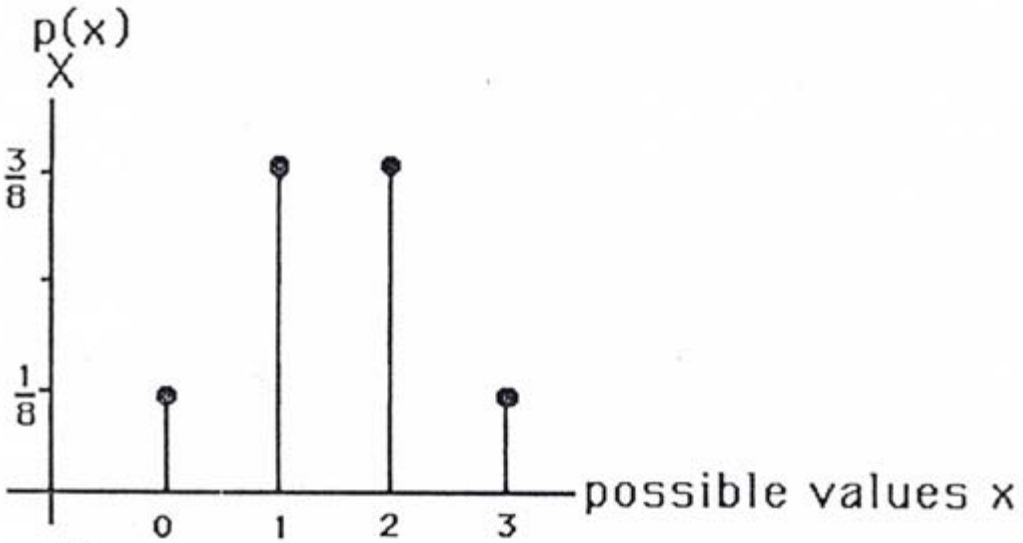


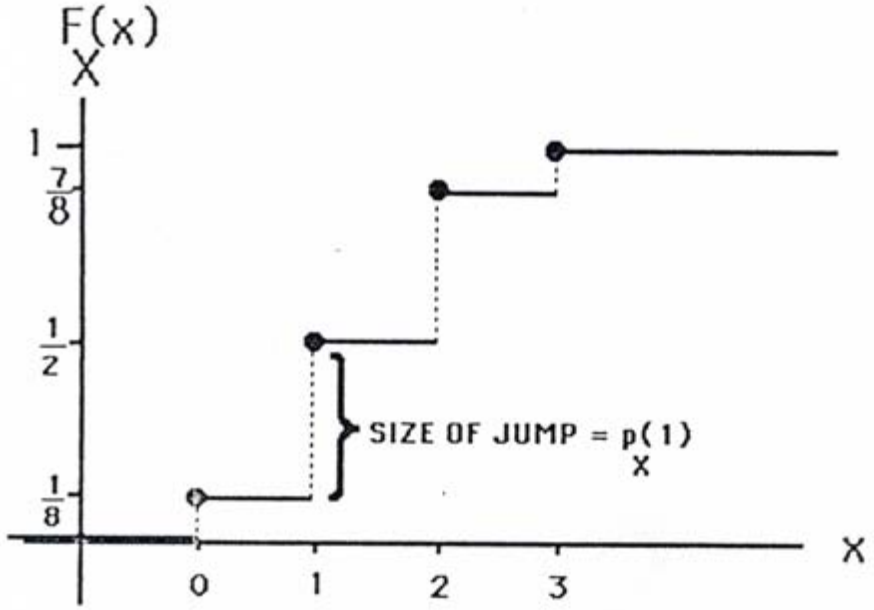
Figure 1.3. The celestial sphere with area A $1^{o} \times 1^{o}$ at the equator.

n	$n!$	Stirling's Approximation $S(n)$	$n!/S(n)$
1	1	.922	1.0844
2	2	1.919	1.0422
3	6	5.836	1.0281
4	24	23.506	1.0210
5	120	118.12	1.0168
6	720	710.08	1.0140
7	5,040	4,980.4	1.0120
8	40,320	39,902.4	1.0105
9	362,880	359,536.9	1.0093
10	3,628,800	3,598,695.6	1.0084

Table 1.1. The accuracy of Stirling's formula for small n.

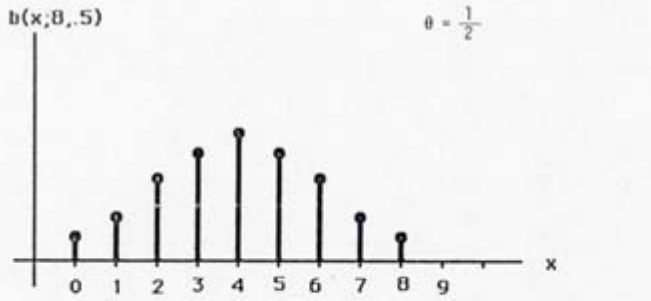


(a) $p_X(x)$ as a mass distribution

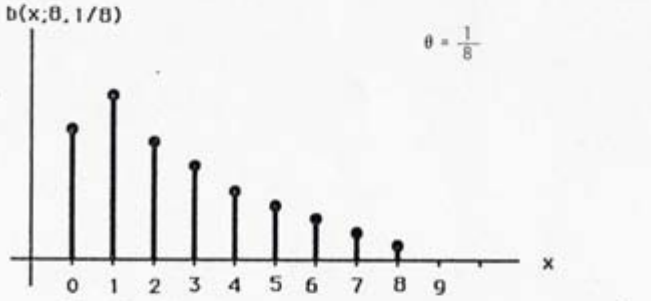


(b) $F_X(x)$ as a jump function

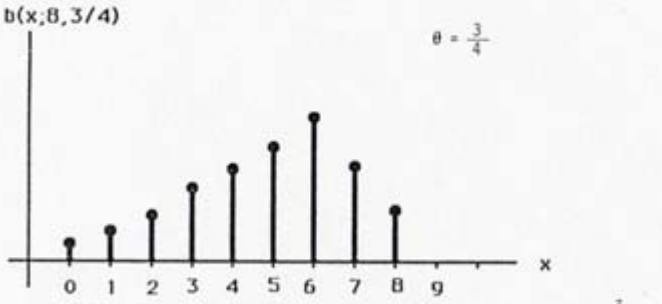
Figure 1.4. Equivalent graphical representations of the distribution of the number of heads in three independent tosses of a fair coin.



(a) The binomial distribution with $n = 8$, $\theta = \frac{1}{2}$.



(b) The binomial distribution with $n = 8$, $\theta = \frac{1}{8}$.



(c) The binomial distribution with $n = 8$, $\theta = \frac{3}{4}$.

Figure 1.5. Three binomial distributions.

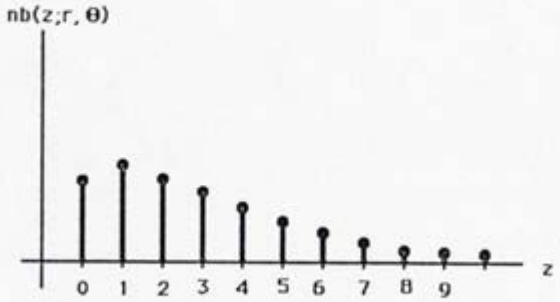


Figure 1.6. The negative binomial distribution for $r = 2$ and $\theta = \frac{1}{2}$.

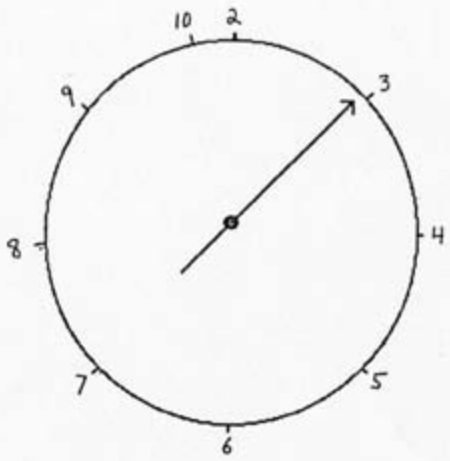


Figure 1.7. A spinner with $a = 2$ and $b = 10.5$.

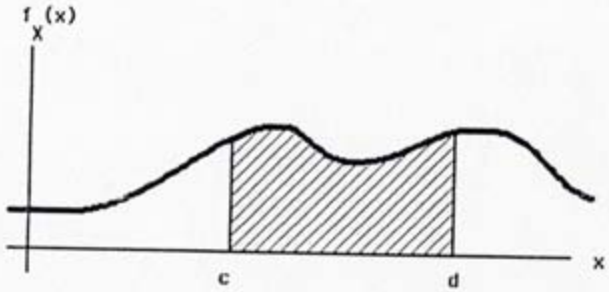


Figure 1.8. A probability density function $f_X(x)$. The shaded area is $\int_c^d f_X(x)\,dx$, the probability that $c < X \leq d$.

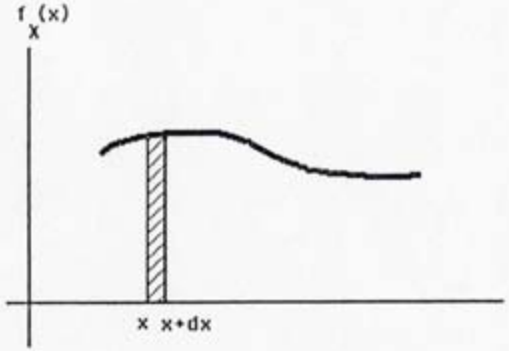


Figure 1.9. $P(x < X \leq x + dx)$ is approximately the area of a rectangle of height $f_X(x)$ and base dx.

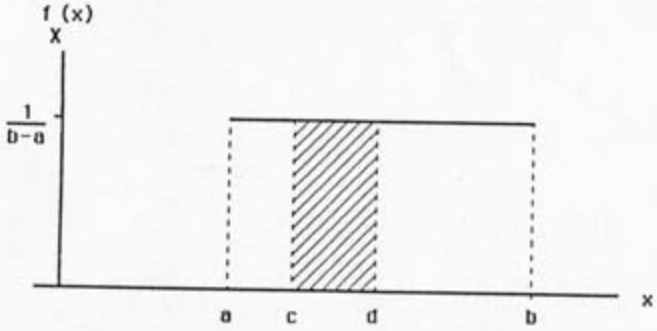
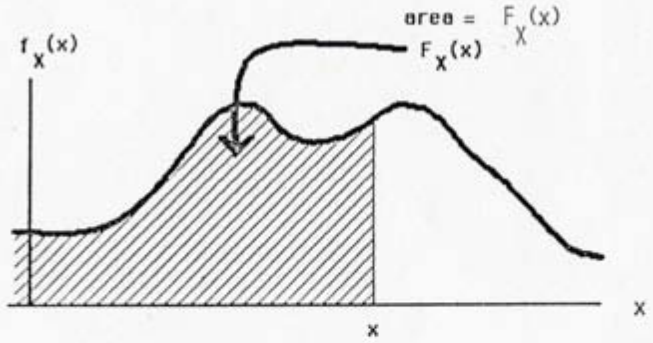


Figure 1.10. The Uniform or Rectangular distribution.



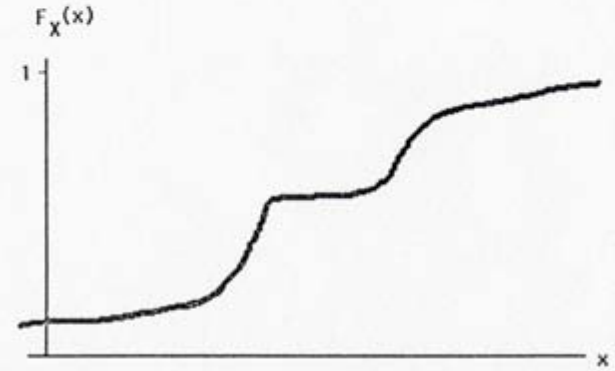
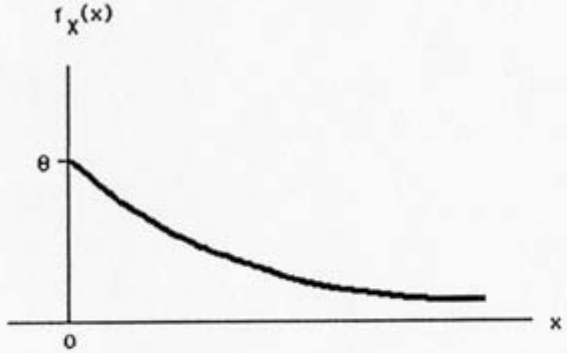
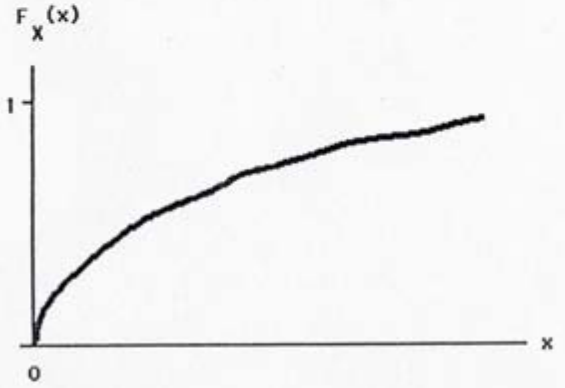


Figure 1.11. The cumulative distribution function of a continuous random variable X.

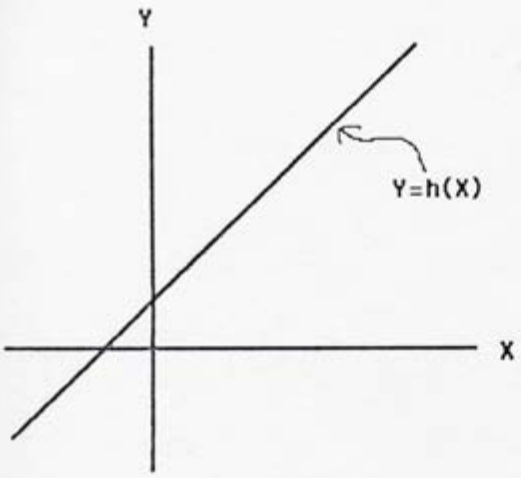


(a) The Exponential (θ) density $f_X(x) = \theta e^{-\theta x}$, for $x \geq 0$.

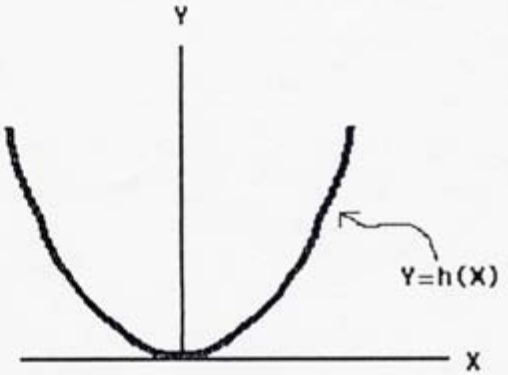


(b) The cumulative distribution function $F(x) = 1 - e^{-\theta x}$, $x \geq 0$, for the Exponential (θ) distribution.

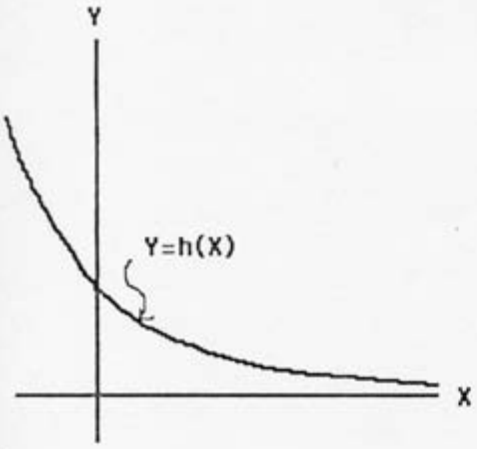
Figure 1.12. The Exponential (θ) distribution.



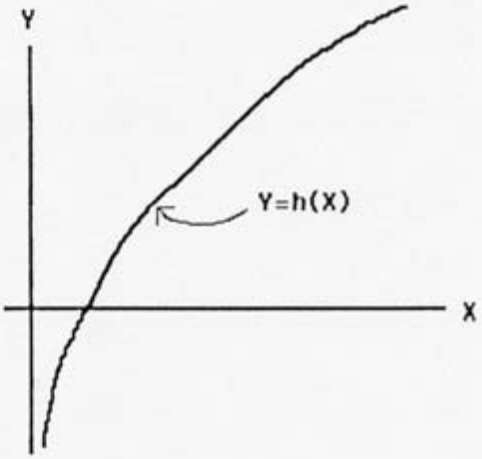
$Y = 2X + 3$, a strictly monotone transformation.



$Y = X^2$, a nonmonotone transformation (but strictly monotone for $x \geq 0$).

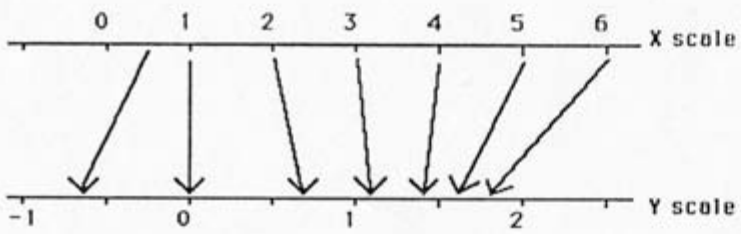


$Y = e^{-X}$, a strictly monotone transformation.



$Y = \log_e(X)$, a strictly monotone transformation for $X > 0$.

Figure 1.13. Four transformations.



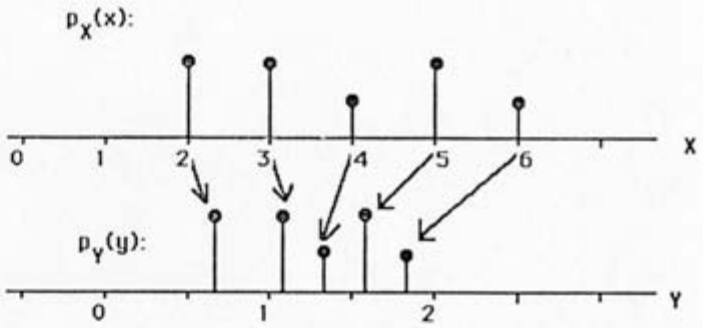
$\log_e(0) = -\infty$ $\log_e(3) = 1.10$

$\log_e(.5) = -.69$ $\log_e(4) = 1.39$

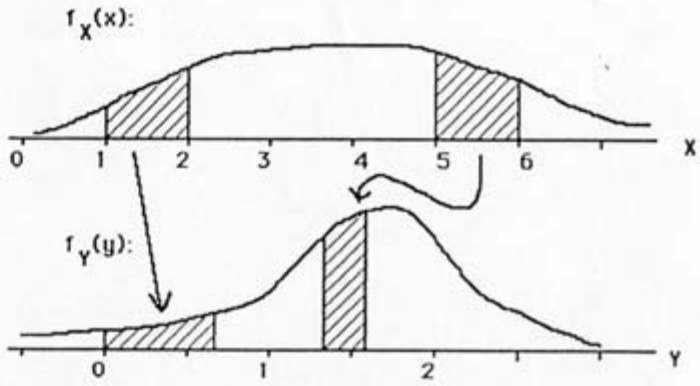
$\log_e(1) = 0$ $\log_e(5) = 1.61$

$\log_e(2) = .69$ $\log_e(6) = 1.79$

Figure 1.14. The effect of the transformation $Y = \log_e(X)$ is to compress large values, stretch out small ones.



(a) Discrete Example, $Y = \log_e(X)$



(b) Continuous Example, $Y = \log_e(X)$

Figure 1.15. The effect of transformation upon probability distribution.

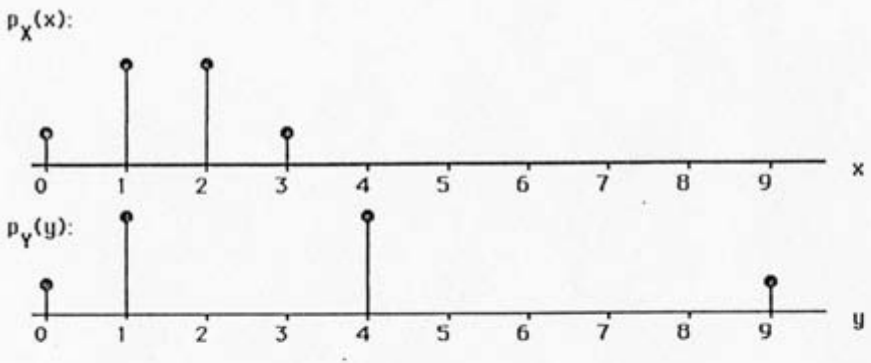
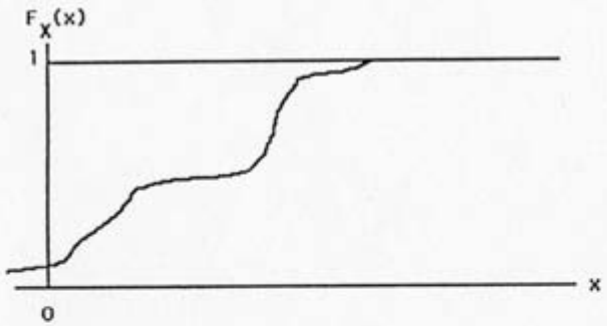
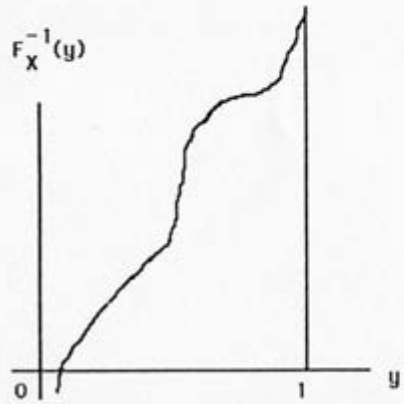


Figure 1.16. The Distributions of X and $Y = X^2$, when X has a Binomial (3, .5) distribution.



(a) The cumulative distribution function $F_X(x)$ of a continuous random variable.



(b) The corresponding inverse cumulative distribution function $F_X^{-1}(y)$, defined only for $0 < y < 1$.

Figure 1.17. The cumulative distribution function and its inverse.

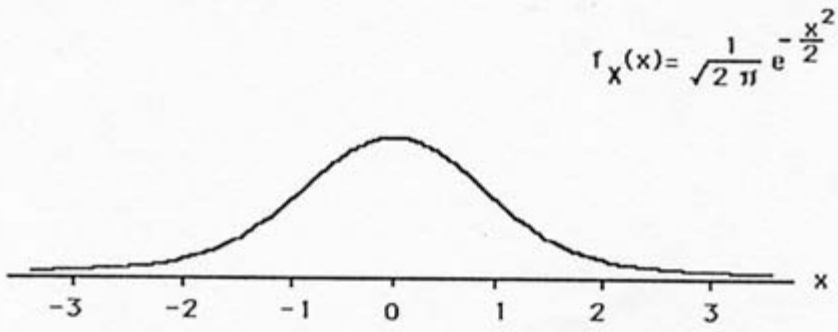


Figure 1.18. The probability density function of the Standard Normal distribution.

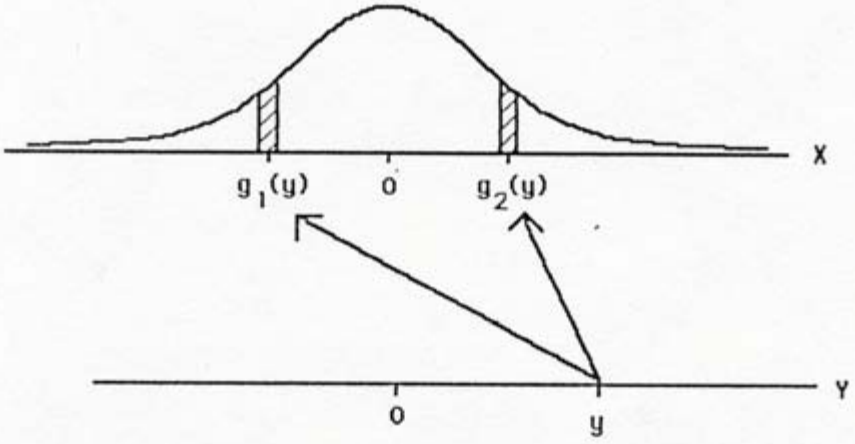


Figure 1.19. The two branches of the inverse of $h(x) = x^2$.

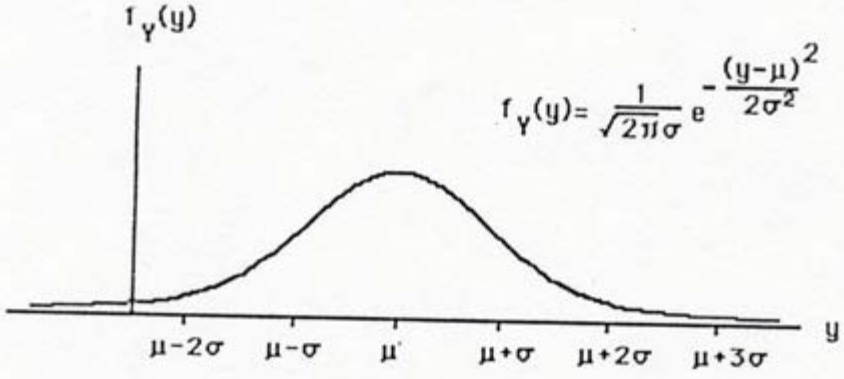


Figure 1.20. The probability density function of the Normal(μ,σ) distribution.