Notes on Probability

James S. Howland

Copyright December 1, 2009.
Revised january 2013.

Contents.

Chapter 1
Combinatorics.

1.1 The Basic Principle of Counting .

Combinatorics is concerned with counting the number of objects in various sets. If the set is small enough that the objects can be listed, this is not difficult. But if the set is large, this method is no longer available.

For example, a Poker hand is a set of five cards, selected from a standard set of 52. There are over two and a half million such sets. Exactly how many are there? Listing them all is out of the question, but their number is easily found, as we shall see below.

In Probability theory, we will often need to compute the number of objects in various large sets. In preparation for this, we shall prove a few basic combinatorial results that will be useful in the following chapters.

Elementary Combinatorics is founded on one Basic Principle.

Basic Counting Principle. *If a job can be done in two stages, with n choices at the first stage and - after the first stage is complete - m choices at the second stage, then there are nm possible ways to do the job.*

Example 1. How many symbols are there of the type $A\ 1$, where the first symbol is a letter is and the second a digit from 0 to 9 ?

Solution. Construct the symbol in two steps by first choosing the letter and then choosing the digit. There are 26 choices at the first step and 10 at the second. By the Basic Principle, the total number is

$$26 \cdot 10 = 260.\blacksquare$$

Example 2. How many three letter words are there if repeated letters are allowed?

Solution. Here 'word' refers to any sequence of letters, regardless of whether it has a meaning in any language.

Construct the word in three steps by choosing the first letter, then the second and then the third. At each step we may choose any of the 26 letters. The total number of words is therefore

$$26 \cdot 26 \cdot 26 = 26^3 = 17,576.\blacksquare$$

Example 3. How many three letter words are there if repeated letters are *not* allowed?

Solution. Construct the word in the same way as in Example 2. This time, however, there are 26 choices at the first step, but only 25 at the second, since one letter has already been used, and similarly, only 24 at the third. The total number of words is therefore

$$26 \cdot 25 \cdot 24 = 15,600.\blacksquare$$

Theorem 1. *The total number of subsets, including the empty set, of a set of n elements is 2^n.*

Proof. List the elements in order: $e_1, e_2, ..., e_n$. Go down the list and for each element, decide whether or not it is to be in the set. For each element there are 2 choices - in or out. The total number is therefore

$$2 \cdot 2 \cdot 2 \cdots 2 = 2^n.$$

This count includes the empty set when the answer is "*No*" for each element, and the whole set when it is always "*Yes*".□

1.2 Permutations and Combinations.

Examples 2 and 3 of the preceding section both deal with *ordered lists of objects chosen from a certain set*, specifically, in that case, the alphabet. (*footnote*: I do not know what an unordered list would be.)

The same argument gives the general case.

Theorem 2. *The number of ordered lists **with repetitions** of length k from a set of n elements is n^k.*

Proof. For each entry on the list from 1 to k, there are n choices, since repetitions are allowed. Thus, the total number of lists is

$$n \cdot n \cdot n \cdots n = n^k.\square$$

Example 1. In how many ways can 2 cards be drawn from a standard deck, if the first card is replaced after it is drawn?

Solution. By Theorem 2, there are $52 \cdot 52 = 2704$ ways.■

The problem of lists *without repetitions* is solved in the same way.

Theorem 3. *The number of ordered lists **without repetitions** of length k from a set of n elements is*

$$P(n,k) = n(n-1)\,(n-2)\cdots(n-k+1)\,.$$

Proof. There are n choices for the first item on the list, and after it has been chosen, there are $n-1$ remaining items to choose from for the second, etc. on down to the k^{th} choice, for which $(n-k+1)$ items are available, so the total number is the product of k factors

$$n\,(n-1)\,(n-2)\cdots(n-k+1)\,.\square$$

Example 2. In how many ways can 2 cards be drawn from a standard deck, if the first card is *not* replaced after it is drawn?

Solution. By Theorem 3, there are $52 \cdot 51 = 2652$ ways.■

The formula for $P(n,k)$ can be written in a more compact form. Define

$$n! = n(n-1)\,(n-2)\cdots 2\cdot 1$$

to be the product of the first n integers, and make the convention that $0! = 1$.

Corollary 1.

$$P(n,k) = \frac{n!}{(n-k)!}.$$

Proof.

$$\begin{aligned} P(n,k) &= n(n-1)\,(n-2)\cdots(n-k+1) \\ &= n(n-1)\,(n-2)\cdots(n-k+1)\,\frac{(n-k)\,(n-k-1)\cdots 2\cdot 1}{(n-k)\,(n-k-1)\cdots 2\cdot 1} \\ &= \frac{n!}{(n-k)!}.\square \end{aligned}$$

Permutations.

A *permutation* of a set is an arrangement of the elements of the set into an ordered list. For example, the six permutations of the letters a, b, c are

$$\begin{array}{lll} abc & bca & cab \\ acb & bac & cba \end{array}$$

A permutation is therefore just an *ordered list without repetitions* of n elements from a set on n elements. We therefore have, by Theorem 2,

Corollary 2. *The number of permutations of a set of* n *elements is*

$$n! = P(n,n) = n(n-1)\,(n-2)\cdots 2\cdot 1.$$

Example 3. How many rearrangements are there of the letters of the word *ROMA*?

Solution. By Corollary 2, there are $4! = 24$ such rearrangements. ■

An *ordered list* of k elements *without repetitions* from a set of n elements is also called a *permutation of the n elements taken k at a a time*. This accounts for the notation $P(n, k)$.

Combinations.

We now derive a very important result.

Theorem 4. *The number of subsets of size k of a set of n elements is*

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}.$$

Proof. Let $\binom{n}{k}$ denotes the number of k-element subsets of a set of size n. The trick is to count the number $P(n, k)$ of permutations of n objects, taken k at a time in another way. First choose a set of k elements; there are $\binom{n}{k}$ possible choices. Then order these elements into a list; there are $k!$ ways to do this. Thus

$$P(n, k) = \binom{n}{k} k!$$

and hence, dividing by $k!$,

$$\binom{n}{k} = \frac{P(n, k)}{k!} = \frac{n!}{k!\,(n-k)!}.\square$$

Example 4. How many five card poker hands are there?

Solution. A poker hand is a set of five cards from a standard deck of 52 cards. Thus, their number is

$$\binom{52}{5} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 2,598,960.\blacksquare$$

The number $\binom{n}{k}$ is sometimes called the number of "*combinations of n things taken k at a time*", especially in the older literature. It is frequently denoted by C_k^n, especially on scientific calculators which have this capability.

The word "*combination*", however, does not seem to be particularly suggestive of a subset. A common usage is to refer to $\binom{n}{k}$ as "n *choose* k", and think of reaching into a set of n objects and pulling out a handful of k objects from the set, without any regard to order.

The number $\binom{n}{k}$ is also referred to as a "*Binomial coefficient*", for reasons that will be made clear in section 1.4.

Symmetry.

Binomial coefficients have the symmetry:

$$\binom{n}{k} = \binom{n}{n-k}.$$

This is clear by inspection of the formula of Theorem 4, but it can also be seen combinatorially, since selecting the k objects that are *in* a set is the same as selecting the $n-k$ objects that are *not in* the set.

Stirling's Approximation.

Since the number $n!$ appears quite frequently in Combinatorics, we shall offer a few comments on its calculation. Standard calculators will compute $n!$, and many will also compute $\binom{n}{r} = C_r^n$ and $P(n,r)$ as well. However, the number $n!$ *grows very rapidly*. Indeed, $10!$ already exceeds one million. Most calculators will compute $n!$ only up to $69!$, but fail after that because $70!$ exceeds 10^{100}, causing an overflow.

For large n, one can use the asymptotic formula of *Stirling*

$$n! \sim \sqrt{2\pi n} n^n e^{-n} = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

The precise statement is

Theorem 5. (Stirling's Approximation.)

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n} n^n e^{-n}} = 1.$$

The approximation is good to less than 1% for $n = 10$, and improves as n increases.

For large values of n, the logarithmic form is more convenient

$$\log(n!) \sim \left(n + \frac{1}{2}\right)\log n - n + \frac{1}{2}\log(2\pi).$$

Example 5. Estimate $100!$.

Solution. This number causes an overflow on most calculators, since it exceeds 10^{100}. By Stirling's approximation

$$\log(100!) \sim (100.5)\log(100) - 100 + \frac{1}{2}\log(2\pi) \simeq 363.7385.$$

The logarithm is the natural logarithm. For computation, the base 10 logarithm is more appropriate:

$$\log_{10}(100!) = \log(100!)/\log(10) \simeq 157.970$$

Taking the exponent gives

$$100! \simeq 10^{0.970} \times 10^{157} = 9.325 \times 10^{157}.\blacksquare$$

1.3 Distinct Permutations and Assignments.

Distinct Permutations.

If some of the letters of a word of n letters are repeated, there are clearly less than $n!$ *distinct permutations*. For example, the word *AAA* has only one distinct rearrangement, rather than $3! = 6$. Similarly, the word *AAH* has only 3 distinct permutations:

AAH, *AHA* and *HAA*..

Example 1. How many distinct permutations are there of the word *MISSISSIPPI* ?

Solution. There are 11 letters in the word, but some are repeated. There are 4 *S's*, 4 *I's*, 2 *P's* and 1 *M*. If all the letters are regarded as distinct objects, there are $11!$ permutations. However, if the 4 *S's* are permuted among one another in their $4!$ ways, we obtain the same word. Similarly for the other letters. Thus there are a total of

$$4! \cdot 4! \cdot 2! \cdot 1!$$

permutations of the letters which give the same word. The number of distinct permutations of *MISSISSIPPI* is therefore

$$\frac{11!}{4!\ 4!\ 2!\ 1!} = 34,650.\blacksquare$$

The same argument gives the general case.

Theorem 6. *The number of* ***distinct permutations*** *of n objects, with n_1 objects of the 1^{st} type, n_2 objects of the 2^{nd} type, ..., and n_r objects of the r^{th} type is*

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

The quantity

$$\binom{n}{n_1\ n_2\ \cdots\ n_r}$$

is called a *multinomial coefficient*, for reasons that will are explained below, in section 1.4.

Example 2. In an election A receives a votes, B receives b votes and C receives c votes. The ballots are collected in a box, and then drawn out one at a time and counted. In how many ways can the counting go?

Solution. The counting sequence can be represented by a sequence of a letters A, b letters B, and c letters C. The number of distinct permutations of this sequence is

$$\binom{a+b+c}{a\ \ b\ \ c} = \frac{(a+b+c)!}{a!b!c!}.\blacksquare$$

Assignments.

Theorem 7. *The number ways to assign n distinct objects to r boxes, with n_1 objects in the 1^{st} box, n_2 objects in the 2^{nd} box, ..., and n_r objects in the r^{th} box is*

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

where $n_1 + n_2 + \cdots + n_r = n$.

Proof. # 1. First choose n_1 objects for the first box. There are $\binom{n}{n_1}$ choices.

Next choose n_2 objects for the second box; there are $n - n_1$ remaining objects to choose from, so there are $\binom{n-n_1}{n_2}$ choices. Continuing in this manner we find the total number of choices to be

$$\begin{aligned}
& \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-n_2-\cdots-n_{r-1}}{n_r} \\
= & \binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n_r}{n_r} \\
= & \frac{n!}{n_1!\,(n-n_1)!}\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!}\frac{(n-n_1-n_2)!}{n_3!\,(n-n_1-n_2-n_3)!}\cdots\frac{n_r!}{0!n_r!} \\
= & \frac{n!}{n_1!n_2!\cdots n_r!} = \binom{n}{n_1\ n_2\ \cdots\ n_r}.\square
\end{aligned}$$

Remark. This is a good way to compute multinomial coefficients on your calculator if you have a C_r^n key; e.g.

$$\binom{12}{5\ 4\ 3} = \binom{12}{5}\binom{7}{4}\binom{3}{3} = 792 \cdot 35 \cdot 1 = 27,720.$$

Note that the result of Theorem 5 is the same as the number of distinct permutations. The second proof shows why.

Proof #2. Align the n objects in a list. Take n_1 letters B_1, n_2 letters B_2,...and, n_r letters B_r, where $n = n_1 + n_2 + \cdots + n_r$. Beside each object of the list, place one of the letters B_k, denoting that that object goes into the k^{th} Box. The sequence of $B_k's$ determines the assignment of objects. But this sequence is just a word made up of n_1 letters B_1, n_2 letters B_2, etc., so the number of assignments is just the number of distinct permutations of this word.■

Looked at another way, a distinct permutation of *MISSISSIPPI* can be obtained by taking 11 balls numbered 1 through 11, and assigning them to the *M, I, S* and *P* boxes with occupation numbers $1, 4, 4$ and 2, with the number on the ball corresponding to the position at which the letter of the box is found. Thus *MISSISSIPPI* corresponds to

M	I	S	P
1	2,5,8,11	3,4,6,7	9,10

Example 2. A total of 100 students are to be assigned to 3 classrooms with 20 in the first, 30 in the second and 50 in the third. How many ways are there to do this?

Solution. By Theorem 5, the number is

$$\binom{100}{20\ 30\ 50} = \frac{100!}{20!\ 30!\ 50!} \simeq 4.8 \times 10^{42}.\blacksquare$$

Indistinguishable Balls.

Another occupation problem is that of assigning n identical indistinguishable balls to r boxes. In this problem, *only the number of balls in each box matters*. For example, with 7 balls and 5 boxes, a possible assignment is a shown in Figure 3.1.

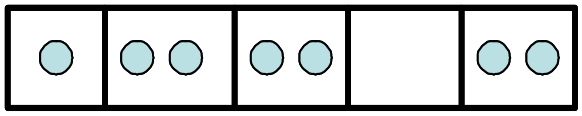

Figure 3.1.

All that matters is the number of balls in each box; that is, the configuration is determined by the *occupation numbers*

$$1 - 2 - 2 - 0 - 2.$$

Like many Combinatorial problems, this problem is easy *if looked at in the right way*. If we remove the superfluous outer boundary of Figure 3.1, we see that the configuration is determined simply by putting in dividers between the balls. Thus, the configuration of Figure 3.1 is determined by Figure 3.2.

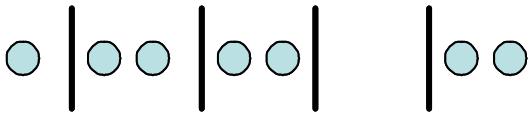

Figure 3.2.

The number of assignments is therefore the number of distinct permutations of 7 balls and 4 dividers, which is

$$\binom{11}{7\ \ 4} = \binom{11}{4} = 330.$$

The general case is the following.

Theorem 8. *The number of ways to put n identical balls into r boxes is*

$$\binom{n+r-1}{n} = \binom{n+r-1}{r-1}.$$

Proof. The configurations are determined by the distinct permutations of n balls and $r-1$ dividers. Their number is therefore

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

Example 3. (a.) How many non negative integer solutions are there to the equation

$$x_1 + x_2 + x_3 + x_4 = 20\ ?$$

(b.) How many *positive* integer solutions?

Solution. Consider 4 boxes with 20 balls. A solution of this equation is determined by letting x_k be the number of balls placed in the k^{th} box.

The solution to (a.) is therefore

$$\binom{23}{3} = 1771.$$

For (b.), *no box can be empty*, so before starting, we must put one ball in each box. That leaves 16 to distribute to 4 boxes, so the solution is

$$\binom{19}{3} = 969.\blacksquare$$

1.4 The Binomial Theorem.

The numbers $\binom{n}{k}$ appear in elementary algebra as the coefficients in the Binomial expansion. The problem is to expand the binomial $(x+y)^n$ into a sum of powers of x and y. This is done as follows. Expanding by the Distributive Law, and *preserving the order of factors*, we have

$$(x+y)^2 = (x+y)(x+y) = xx + xy + yx + yy = x^2 + xy + yx + y^2$$

Collecting terms gives

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

Similarly,

$$\begin{aligned}(x+y)^3 &= (x+y)(x+y)(x+y)\\ &= xxx+xxy+xyx+xyy+yxx+yxy+yyx+yyy\\ &= x^3+3x^2y+3xy^2+y^3.\end{aligned}$$

The term $3x^2y$ is the sum $xxy+xyx+yxx$ of the *three distinct permutations* of xxy. Thus the coefficient of x^2y in the expansion of $(x+y)^3$ is

$$3=\frac{3!}{2!1!}=\binom{3}{2}.$$

In general, if $(x+y)^n$ is expanded, the result consists of a sum of terms x^ky^{n-k}. The coefficient x^ky^{n-k} is the number of distinct permutations of k letters x and $n-k$ letters y; that is

$$\frac{n!}{k!\,(n-k)!}=\binom{n}{k}.$$

Thus, we have

Theorem 9. (**Binomial Theorem.**)

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

The Binomial theorem is useful for evaluating sums involving the Binomial coefficients.

Example 1. For example, if we set $x=y=1$, we get the sum of the Binomial coefficients

$$\sum_{k=0}^{n}\binom{n}{k}=2^n.$$

This is clear, since $\binom{n}{k}$ is the number of k-element subsets of a set of n elements, and 2^n is the total number of subsets.■

Example 2. A less obvious result is the alternating sum, obtained with $n>0$ and $x=-1$, $y=1$:

$$0=(1-1)^n=\sum_{k=0}^{n}(-1)^k\binom{n}{k}=1-\binom{n}{1}+\binom{n}{2}-\cdots+(-1)^{n-1}\binom{n}{n-1}+(-1)^n.$$■

For *even* n , this is clear from the symmetry $\binom{n}{k}=\binom{n}{n-k}$. We will use this identity below in section 1.5.

Example 3. For a final example, differentiate the formula with respect to x. This brings down a factor of k into the sum:

$$\frac{\partial}{\partial x}(x+y)^n = n(x+y)^{n-1} = \frac{\partial}{\partial x}\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k} = \sum_{k=0}^{n}\binom{n}{k}kx^{k-1}y^{n-k}.$$

Setting $x = y = 1$ then gives

$$\sum_{k=0}^{n} k\binom{n}{k} = n2^n.\blacksquare$$

Pascal's Triangle.

A pictorial representation which is useful for generating the Binomial coefficients for small values of n goes by the name of "*Pascal's Triangle*." It was studied by Pascal in the 17th century, but was known much earlier in China and India.

To obtain the triangle, first write out the first few instances of the Binomial theorem.

$$(x+y)^0 = 1$$
$$(x+y)^1 = x+y$$
$$(x+y)^2 = x^2+2xy+y^2$$
$$(x+y)^3 = x+3xy+3xy+y$$
$$(x+y)^4 = x^4+4x^3y+6x^2y^2+4xy^3+y^4$$
$$(x+y)^5 = x^5+5x^4y+10x^3y^2+10x^2y^3+5xy^4+y^5$$
$$(x+y)^5 = x^6+6x^5y+15x^4y^2+20x^3y^3+15x^2y^4+6xy^5+y^5$$

If we write only the coefficients in triangular form, we find

					1							
				1		**1**						
			1		**2**		**1**					
		1		**3**		**3**		**1**				
	1		**4**		**6**		**4**		**1**			
1		**5**		**10**		**10**		**5**		**1**		
1		**6**		**15**		**20**		**15**		**6**		**1**

Now, observe that each entry can be computed from the preceding row by adding the two entries just above it in the table. For example, the entry 6 in the fourth row is the sum $3+3$ of the two numbers above it, while the entry 10 in the fifth row is the sum $4+6$, etc. Thus, the next row of the triangle will be

1 7 21 35 35 21 7 1

This arrangement of the Binomial Coefficients is known as *Pascal's Triangle*. What we

are saying is that the coefficient $\binom{n+1}{k}$ in the $(n+1)^{st}$ row is the sum of the coefficients $\binom{n}{k-1}$ and $\binom{n}{k}$ just above it. This is known as the *Law of Pascal's Triangle*.

Theorem 10. (Law of Pascal's Triangle.)

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}.$$

We shall give three proofs. The first is essentially *by induction*.

Proof # 1. Write

$$\begin{aligned}(x+y)^n &= x^n + \cdots + \binom{n}{k-1}x^{k-1}y^{n-k+1} + \binom{n}{k}x^k y^{n-k} + \cdots + y^n \\ &= \cdots + \binom{n}{k-1}x^k y^{n-k+1} + \binom{n}{k}x^k y^{n-k+1} + \cdots .\end{aligned}$$

and multiply by $(x+y)$ to get

$$\begin{aligned}&(x+y)\left(x^n + \cdots + \binom{n}{k-1}x^{k-1}y^{n-k+1} + \binom{n}{k}x^k y^{n-k} + \cdots + y^n\right) \\ = \;& \cdots + \binom{n}{k-1}x^k y^{n-k+1} + \binom{n}{k}x^k y^{n-k+1} + \cdots \\ = \;& (x+y)^{n+1} = \cdots + \binom{n+1}{k}x^k y^{n+1-k} + \cdots\end{aligned}$$

Equating the coefficients of $x^k y^{n+1-k}$ gives

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}.\square$$

The second proof is just *algebra with factorials*.

Proof # 2. We simply compute that

$$\begin{aligned}\binom{n}{k-1} + \binom{n}{k} &= \frac{n!}{(k-1)!\,(n-k+1)!} + \frac{n!}{k!\,(n-k)!} \\ &= \frac{n!}{(k-1)!\,(n-k)!}\left(\frac{1}{(n-k+1)} + \frac{1}{k}\right) \\ &= \frac{n!}{(k-1)!\,(n-k)!}\left(\frac{n+1}{(n-k+1)\,k}\right) \\ &= \frac{(n+1)!}{k!\,(n+1-k)!} = \binom{n+1}{k}.\square\end{aligned}$$

From our point of view, the third proof is the most interesting. It is *combinatorial*.

Proof # 3. Select a set of k balls from an urn containing n White balls and one Black ball. The number of sets with all White balls is $\binom{n}{k}$. The number having the one Black ball is $\binom{n}{k-1}$, since we must select $k-1$ Whites to complete the set. Thus the total number of sets is

$$\binom{n}{k-1} + \binom{n}{k}.$$

But the total number of sets is also just $\binom{n+1}{k}$.□

1.5 *Inclusion-Exclusion.

Next, we shall discuss a counting method known as *Inclusion-Exclusion*. In order to introduce the method, we consider an example.

Example 1. How many numbers from 1 to $420 = 3 \cdot 7 \cdot 5 \cdot 2^2$ are divisible

(a.) by 5 ?

(b.) by either 3 or 5 ?

(c.) by either 3 or 5 or 7 ?

Solution. For (a.), every fifth number is divisible by 5. So there are $420/5 = 84$ such numbers.

For (b.), there are 84 numbers divisible by 5, and $420/3 = 140$ numbers divisible by 3. If we add these to get 224, we have counted twice the $420/15 = 28$ numbers that are *divisible by both* 3 and 5. We must *exclude* these, so the answer is

$$140 + 84 - 28 = 196.$$

For (c.), there are 84 numbers divisible by 5, 140 numbers divisible by 3, and $420/7 = 60$ divisible by 7. If we add these to get 284, we have counted twice:

the $420/15 = 28$ numbers that are *divisible by both* 3 and 5,.

the $420/21 = 20$ numbers that are *divisible by both* 3 and 7.

the $420/35 = 12$ numbers that are *divisible by both* 5 and 7.

We must *exclude* these, so we subtract them out. However, there are $420/105 = 4$ numbers *divisible by all three* numbers 3, 5 and 7. These have been included three times, but then subtracted out three times. So we must add them back in. Thus, the final answer is

$$(140 + 84 + 60) - (28 + 20 + 12) + 4 = 228.\blacksquare$$

The general rule is as follows. If S is a set, let $N(S)$ be the number of elements of S.

Theorem 11. (Inclusion-Exclusion.) *Let*

$$S_r = \sum_{i_1<\cdots<i_r} N(E_{i_1} E_{i_2} \cdots E_{i_r})$$

Then

$$N(E_1 \cup E_2 \cup \cdots \cup E_n) = S_1 - S_2 + S_3 - \cdots + (-1)^n S_n. \tag{1.1}$$

Proof. Consider a point p of the union that is in exactly m of the sets E_k. For definiteness, say that p is in the first m sets $E_1, E_2, , ..., E_m$. It will then contribute to $\binom{m}{r}$ terms of the sum S_r, and will therefore be counted $\binom{m}{r}$ times in the sum S_r times. Taking signs into account, it is counted in the sum on the right side of (1.1) exactly

$$\binom{m}{1} - \binom{m}{2} + \binom{m}{3} - \cdots + (-1)^m \binom{m}{r}$$

times. But by Example 2 of the preceding section this is equal to 1.□

We will give two more proofs of this identity, in sections 2.7 and 5.11.

Results of this type are sometimes known as *Sieve formulas* in Combinatorial Theory.

Example 2. Find the number of 5 card hands which contain at least one card of each suit.

Solution. Let S be the set of hands containing *no spades*, etc. We compute the number of the hands which are *void in one suit*, that is the number of elements in the set

$$S \cup H \cup D \cup C.$$

We have

$$\begin{aligned}
& N(S \cup H \cup D \cup C) \\
= \; & N(S) + N(H) + N(D) + N(C) \\
& -N(SH) - N(SD) - N(SC) - N(HD) - N(HC) - N(CD) \\
& +N(SHD) + N(SHC) + NSDC) + N(HDC) - N(SHDC) \\
= \; & 4\,N(S) - 6\,N(SH) + 4\,N(SHD) - N(SHDC) \\
= \; & 4\binom{39}{5} - 6\binom{26}{5} + 4\binom{13}{5} - 0 = 1,913,496.
\end{aligned}$$

Since there are $\binom{52}{5} = 2,598,960$ five card hands, the number of hands with at least one card of each suit is therefore $685,464$.

Second count. A good way to check your answer in counting problems is to *count things in two different ways to see if you get the same answer*. In the present case, one can first pick which suit will have two cards, and then choose two cards from that suit and one from each of the others. This gives

$$\binom{4}{1}\binom{13}{2}\binom{13}{1}^3 = 685,464.\blacksquare$$

1.6 The Multinomial Theorem.

We can generalize the Binomial Theorem to powers of multinomials.

Theorem 12. (Multinomial Theorem.)

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

Proof. As in the proof of the Binomial theorem, the coefficient of $x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$ is the number of distinct permutations of n_1 x_1's, n_2 x_2's,...and n_r x_r's. But this number is just

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

by Theorem 5. □

Setting $x_1 = x_2 = \cdots = x_r = 1$ gives

Corollary 3.

$$\sum_{n_1+n_2+\cdots+n_r=n} \binom{n}{n_1\ n_2\ \cdots\ n_r} = r^n.$$

Example 1. Expand $(x + y + z)^4$.

Solution. We have

$$\begin{aligned}
& (x+y+z)^4 \\
= & \sum_{n_1+n_2+n_3=4} \binom{n}{n_1\ n_2\ n_3} x^{n_1} y^{n_2} z^{n_3} \\
= & \binom{4}{4\,0\,0}\left(x^4+y^4+z^4\right) + \binom{4}{2\,2\,0}\left(x^2y^2+x^2z^2+y^2z^2\right) \\
& + \binom{4}{2\,1\,1}\left(x^2yz+xy^2z+xyz^2\right) \\
& + \binom{4}{3\,1\,0}\left(x^3y+x^3z+xy^3+xz^3+y^3z+yz^3\right) \\
= & \left(x^4+y^4+z^4\right) + 6\left(x^2y^2+x^2z^2+y^2z^2\right) + 12\left(x^2yz+xy^2z+xyz^2\right) \\
& + 4\left(x^3y+x^3z+xy^3+xz^3+y^3z+yz^3\right).
\end{aligned}$$

Corollary 3 provides a check on this expansion. Setting $x = y = z = 1$ gives

$$3 + 6 \cdot 3 + 12 \cdot 3 + +4 \cdot 6 = 81 = 3^4$$

which is correct. ■

Theorem 13. *The number of terms in a multinomial expansion is*

$$\binom{n+r-1}{r-1}.$$

Proof. We are counting the number of terms of the form $x_1^{n_1}x_2^{n_2}\cdots x_r^{n_r}$ where $n_1+n_2+\cdots+n_r=n$. Such a term may be determined by setting up a box for each of the r variables, and distributing n identical balls to the boxes. Thus, for example, the term x^2yz in the expansion of $(x+y+z)^4$ corresponds to the configuration

x	y	z
OO	O	O

By Theorem 8, the number of such configurations is $\binom{n+r-1}{r-1}$.□

Example 2. Theorem 13 provides another check on the expansion of Example 1. We compute that there should be

$$\binom{4+3-1}{3-1}=\binom{6}{2}=15$$

terms, and so there are. So we have them all. ■

1.7 Problems.

(1.) A multiple choice quiz consists of 10 questions, with possible answers (a.), (b.), (c.), and (d.).

(a.) In how many ways can the quiz be answered?

(b.) In how many ways if questions can be left blank?

(2.) A list of ten rivers is given. How many ways are there to answer the following questions?

(a.) List the rivers in order of length.

(b.) List the four longest in order of length.

(c.) What are the four longest ?

(d.) What are the three longest and the two shortest ?

(3.) Suppose that license plates consist of three letters followed by three digits. Find the number of possible plates.

(4.) At one time all telephone area codes in Canada and the U.S. consisted of three digits. The first was between 2 and 9, the second was either 0 or 1, and the third was between 1 and 9. How many such codes were possible?

(5.) How many 7 place licence plates are there if

(a.) the first three places are letters and the last four are digits?

(b.) How many if no symbol can be repeated?

(c.) How many if the symbols may appear in any order?

(d.) How many if the symbols may appear in any order, and no symbol may be repeated?

(6.) Find the number of distinct permutations *ABRACADABRA.*

(7.) In how many ways can a collection of 4 coins be chosen if all coins of the same denomination are regarded as identical?

(8.) How many permutations are there of the words (a.) *FLUKE* (b.) *PROPOSE*. (c.) TALLAHASSEE

(9.) (a.) Prove that

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

(*Hint.* Consider n White and m Black balls.)

(b.) Prove that

$$\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$$

(10.) (*Fermat's Identity*) Prove that for $n \geq r$,

$$\sum_{k=r}^{n} \binom{k-1}{r-1} = \binom{n}{r}.$$

(*Hint.* How many subsets of $1, 2, ..., n$ have k as their largest element?)

(11.) Find the sum

$$\sum_{k=0}^{n} k^2 \binom{n}{k}.$$

(12.) Expand $\left(2x^3 + y\right)^5$.

(13.) How many numbers less than a million are neither perfect squares nor perfect cubes?

(14.) Expand $(2x + y + 3z)^4$.

(15.) (a.) Use Stirling's approximation to approximate $\binom{n}{k}$ is n and k are both large.
(b.) Use this to find $\binom{50}{25}$

Section 1.7 Problems.

(16.) If all n_k are large, find an approximation to

$$\binom{n}{n_1\ n_2\ \cdots\ n_r}.$$

Chapter 2
Probability.

2.1 Introduction.

Classical mathematics - the mathematics that you have learned about so far - is built around two primitive intuitive notions. Arithmetic is based on the idea of the Number Sequence, while Geometry is based on the notion of Euclidean Space.

The objects of Arithmetic and Geometry are not physical things but objects of thought. The number Two and perfect lines and circles are not objects of the physical world. But these mental concepts enable us to think about the certain aspects of the world in extremely useful ways.

The theory of Probability is built around a new intuition - the concept of Independent Trials. As with Number and Space, the idea of Independent Trials also enables us to think in useful ways about many aspects of the world. It is used to understand *Gambling games*, such as poker, blackjack, craps, roulette and lotteries. It forms the basis of many *Statistical methods* for understanding such things as polls and drug trials, and of *Actuarial Science*, the theory of Insurance. It is used in *Biology* to study genetics and the dynamics of populations and in *Finance* to study fluctuations of stock prices. It contributes to the understanding of such *Physical processes* as radioactive decay, Brownian motion, chemical reaction rates, and the changes in phase of physical systems. In *Engineering*, it permits the study the transmission of messages over a noisy channel, a discipline known as *Information theory*. Last, and most spectacularly, it arises in a strange, but essential form in *Quantum Mechanics*, the fundamental theory of all microscopic physics.

A pretty good list, I think you will agree, and by no means exhaustive.

In contrast to the ancient notions of Number and Space, which go back to times about which we know very little, the idea of Independent Trials is comparatively recent, appearing, so far as we know, no earlier than the 16th century. If it has occurred to you that an idea so late in coming might not be exactly obvious, you are quite correct in your suspicions, and this brings us to a basic fact about the initial study of probability theory, on which we are about to embark. We are not primarily concerned - as we might be in, say, a course on differential equations - with a collection of computational techniques for solving certain kinds of problems, although there is a good bit of that. The difficult part is *to learn to think in terms of concepts like independence, probability, random variable, expectation*, etc. For this reason, most exercises are ”word problems”, in which a significant part of the problem is to interpret the question in terms of the concepts of probability theory.

So far as is known, the first appearance of the basic insight is in the writings of the Renaissance mathematician, physician, astrologer and gambler Gerolomo Cardano. His

book - published posthumously in 1663, but written around 1526 - is called *Liber de Ludo Aleae*, or *The Book of Games of Chance*. As is obvious from the title, it was concerned with gambling games, particularly those played with dice. It was in connection with such games that the theory originated, rather than with any of the other applications we have cited.

There is a good reason for this, one that underlies many - perhaps most - of the most important discoveries in Science. It is in the context of gambling games that *the essential idea occurs in a relatively pure and simple form.* Phenomena in which several different effects are combined are very difficult to unravel if one does not first understand their components. Thus, for probability, gambling games provide the model for concepts which are then exported to other situations. So we will begin our analysis with them.

Let us consider, then, the simplest game of chance: the *repeated tosses of a coin*. A player is betting on Heads and has won five times in a row. *How likely is it that he will win on the next toss*?

Three possible answers come to mind.

1. According to the "*Law of Averages*", "Tails is due"; thus, Heads is now *less likely* to turn up.

I think that perhaps the average person believes this to be true.

2. On the other hand, perhaps the player is on a "*Hot Streak*", and is therefore *more likely* to win on the next toss.

This view may be popular with some inveterate gamblers, although it is probably less common than the preceding one. I have actually known people who held this opinion. I was once advised by one of them to "bet small until you start to win - and then you're hot, and you bet big!". I am sorry to report that this individual, who eventually left large sums in Reno, had a degree in Mathematics from a very prominent institution.

3. A third idea is that the trials are *Independent and Identical*, that is, the chances of winning are the *same on every toss*, regardless of past history.

Probability theory is based on the third idea. The basic idea of probability theory is the concept of *a sequence of independent and identical repetitions of some game or experiment*. Once this idea is grasped, all of probability theory will eventually follow.

(Of course, whether this is so for an actual sequence of tosses of a particular coin is a matter for debate or perhaps experiment. What we are saying is that Probability theory is based on this idea.)

(*Note*: There are, surely, other possibilities that can be imagined.The game may, for example, be crooked. And, of course, the Statistician might say that Heads is more likely because we have evidence that coin is not fair. More interesting is the attitude of the ancients, who used dice, or something similar, for divination. The idea seemed to be that since the thrower had no control over the outcome, the deity in question had total control, and could communicate a favorable or unfavorable answer to whatever question was being

asked. It is not hard to see that such an attitude is not conducive to the development of the idea of Independence. See F. N. David. *"Games, Gods and Gambling."*)

2.2 Random Experiments and Probability.

The first thing to recognize about gambling games, such as dice or roulette, is that there is an *experiment* - the roll of the dice, the spin of the wheel - which may be repeated in principle as many times as we wish.

We are interested in certain things that may or may not occur on any given *trial* of the experiment - for example, the roll of a double six on the dice, or appearance of a Red number on the wheel. Such things are referred to as *events*.

In general, then, we consider an Experiment $\mathcal{E}$, which is subject indefinitely to identical repetitions or *trials*. An *event* E is something which may or may not occur on a given trial.

Logical Combinations of Events.

New events can be created from old ones by simple *logical operations*. Given two events E and F, we can create:

1. the event "E *and* F" - or just EF - which occurs iff *both* E *and* F *occur*;
2. the event "E *or* F" which occurs iff *either* E *or* F *(or both)* occur;.
3. the event "*not* E" or E^c, which occurs iff E does not occur.

There is also a *relation of inclusion*: if the event F *must* occur whenever the event E occurs, we will write $E \subset F$. For example, on the roll of two dice, the event *"double five"* is contained in the event *"Sum* 10*"*.

We will recognize a *null event* $\emptyset$ which *never* occurs and a *sure event* Ω which *always* occurs.

Probability.

The outcome of such an experiment in not the same on every trial. However, it is clear that some events are more likely to occur than others. For example, in the roll of two ordinary dice, the sum of seven is more likely than a double six.

Why do we say this? Because if we roll the dice a large number of times, we find more sevens occurring than double sixes. "*More likely*." therefore means "*occurs more frequently*."

In order to have a mathematical description of what is happening, we need to quantify the notion of "more likely." Can we find a number $P(E)$ associated with an event E that measures its relative likelihood of occurring?

The answer is readily at hand. *We take* $P(E)$ *to be the relative frequency of* E *in a large number of trials* - the fraction of trials on which E will occur in the long run. That is, if $\mathcal{E}$ is subject to a large number n of independent trials, and the event E occurs on m of them, then

$$P(E) \simeq \frac{m}{n}$$

if n is large. This number $P(E)$ is called the *probability* of the event E.

The following properties of $P(E)$ are clear.

Theorem 1. (Properties of Probability.)

(a.) *For every event,* $0 \leq P(E) \leq 1$.

(b.) $P(\emptyset) = 0$ *and* $P(\Omega) = 1$.

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

Proof. (a.) Since always $0 \leq m \leq n$, we must have $0 \leq m/n \leq 1$.

(b.) We always have $m = 0$ for the event $\emptyset$, and $m = n$ for the event Ω.

(c.) Let E occur k times and F occur m times. Since $E \subset F$, we must have $k \leq m$.□

The characteristic property of probability is expressed in the *Addition Law*. For this we need the concept of *mutually exclusive events*.

Definition 1. Two events are said to be *mutually exclusive* iff they cannot occur on the same trial; that is, iff the occurrence of one excludes the occurrence of the other. More generally, several events $E_1, E_2, \ldots, E_n$ are *mutually exclusive* iff at most one of these events can occur on a single trial.

The events $E_1, E_2, \ldots, E_n$, are *exhaustive* iff they exhaust the possibilities of outcomes, that is, iff at *least one event* E_k must occur on every trial. To say that $E_1, E_2, \ldots, E_n$, are both mutually exclusive and exhaustive is to say that exactly one of the events occurs on every trial.

Example 1. Examples of mutually exclusive sets of events are

(a.) Heads and Tails in the toss of a coin,

(b.) The numbers 1, 2, 3, 4, 5 and 6 on the roll of a die,

(c.) Red and Black on a roulette wheel.

Examples (a.) and (b.) are also exhaustive, but *(c.) is not* due the the presence of the zero, which is neither Red nor Black. ■

Theorem 2. (The Addition Law.)

(a.) *If* E *and* F *are mutually exclusive, then*

$$P(E \text{ or } F) = P(E) + P(F).$$

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

Proof. (a.) Let E occur k times and F occur m times. Since $E \subset F$ never occur together, the event E or F occurred exactly $k + m$ times. Thus,

$$P(E \text{ or } F) \simeq \frac{k+m}{n} = \frac{k}{n} + \frac{m}{n} \simeq P(E) + P(F).$$

(b.) The events E and E^c are clearly mutually exclusive, and E or $E^c = \Omega$. Hence,

$$P(E \text{ or } E^c) = P(E) + P(E^c) = P(\Omega) = 1.\square$$

Example 2. A die is rolled. What is the probability that a five or a six will show up?

Solution. We shall assume that all six numbers on the die are equally likely; that is, over a large number of rolls, each number will appear on average about one-sixth of the time, so that the probability of each number is $1/6$. Such a die is called a *fair die*.

The probability of a five or a six is then

$$P(5 \text{ or } 6) = P(5) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.\blacksquare$$

Example 3. Suppose that a card is drawn from a standard 52 card deck. What is the proability of an Ace or a King? Of an Ace or a Spade?

Solution. If we assume that all cards are equally likely to be drawn, then since there are 4 Aces, the probability of an *Ace* is $4/52 = 1/13$. Similarly, the probability of a *King* is also $1/13$. Since no card can be an Ace and a King at the same time, the events *Ace* and *King* are mutually exclusive, and we have

$$P(\textit{Ace} \text{ or } \textit{King}) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}.$$

In a similar way, the probability of a Spade is $13/52 = 1/4$. However, a card can very well be both an Ace and a Spade at the same time, so the events *Ace* and *Spade* are *not mutually exclusive*, so that $P(\text{Ace or Spade})$ is not the sum $1/13 + 1/4 = 17/52$. Rather, we have

$$P(\textit{Ace} \text{ or } \textit{Spade}) = \frac{16}{52} = \frac{4}{13}.$$

since there are 16 cards that are either Aces or Spades.$\blacksquare$

Remark. The perceptive reader may uneasy about such a definition. For if the trials are truly independent of one another, it seems to be - *and, in fact, is* - perfectly possible to run $10,000$ consecutive Heads with a fair coin. Nevertheless, we shall assume that there is such a number as $P(E)$. We shall return to this point later when discussing the *Law of Large Numbers*.

*The Odds Ratio.

Probabilities are sometimes expressed as *Odds ratios*. If we say that the odds on a certain event are *2:1*, this means that it will occur 2 times out of three, or with probability $2/3$. In general, odds of $a : b$ in favor of an event means that it has probability

$$p = \frac{a}{a+b}.$$

In the other direction, an event with probability p has odds $p : q$ where $q = 1 - p$.

Odds are a ratio, so both a and b may be multiplied by the same number.

Example 3. If E has probability $1/4$ it has odds of *1:3*. Conversely, if the Odds are *5:3* in favor of an event, its probability is

$$\frac{5}{5+3} = \frac{5}{8}.$$

2.3 Label Spaces and the Addition Theorem.

We shall now describe a useful formalism of probability problems.

A *label space* Ω for an experiment $\mathcal{E}$ is a set of objects which *label the possible - or better, the relevant - outcomes of* $\mathcal{E}$.

An event E of $\mathcal{E}$ then corresponds to a subset of Ω, namely the subset of all (labels of) outcomes for which the event occurs. The labels - the points of Ω - label a set of mutually exclusive and exhaustive events, and are sometimes called *elementary events*.

The logical operations on sets correspond to elementary set operations.

The event "E *or* F" corresponds the *union* $E \cup F$,

The event "E *and* F" to the *intersection* $E \cap F$, and

The event "*not* E" to the *complement* E^c of E.

The relation $E \subset F$ holds iff E is a subset of F.

Two events are *mutually exclusive* iff the corresponding sets are disjoint.

Label spaces are usually called *sample spaces*. We will use both terms interchangeably. The original term of their inventor, Richard von Mises, was label space - *Merkmalraum* in German. In the author's opinion, the term label space is more descriptive of the purpose of the space. It is a labeling of the result of the experiment for the purpose at hand, and may be different depending on the question to be asked.

Example 1.

(a.) A coin is tossed.
The sample space is the set $\{H, T\}$

(b.) Two coins are tossed.
The sample space is the set $\{HH, HT, TH, TT\}$.

(c.) A die is rolled.
The sample space is the set $\{1, 2, 3, 4, 5, 6\}$ .of the possible numbers that can appear.

(d.) A coin is tossed n times.

The sample space is the set of sequences of n symbols H or T. It has 2^n points.

(e.) Five cards are drawn from a deck without replacement.

If we are only interested in the *set* of cards drawn, the sample space may be taken as the collection of *all five card sets*. However, if we are interested in the *order* in which cards are drawn - for example, if we wish to know whether the third card is an Ace - then the sample space must be taken to be *all lists of five cards*. This illustrates the fact that the sample space may be different depending on the question to be asked.

Example 2. In Example 1, the sample space is finite. Here are some examples in which it is infinite.

(a.) A coin is tossed until the first head appears.
The sample space is the set $\{1, 2, 3, \ldots\}$ positive integers.

(b.) The time between two successive phone calls to a switchboard is recorded.
The sample space is the set of positive real numbers.

(c.) A pointer is spun.
The sample space is the interval from 0 to 2π, labeling the angle between the pointer and the horizontal.

(d.) A dart is thrown onto a circular dartboard.
The sample space is a circular disc in the plane, consisting of the coordinates of the point of the dart.

In (a.) the sample space is an *infinite discrete* set; in (b.) and (c.) it is a *continuous interval*; in (d.), it is a *continuous subset of the plane*.

Probability Measures.

The notion of probability may then be formalized in terms of sample spaces.

Definition 1. A *probability measure* on a sample space Ω is a function $P(E)$ on the subsets of Ω satisfying

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

(b.) $P(\emptyset) = 0$ and $P(\Omega) = 1$.

(c.) If E and F are disjoint sets, then

$$P(E \text{ or } F) = P(E) + P(F).$$

Corollary 1. *For any sets E and F, we have*

(a.) *If $E \subset F$, then*

$$P(E) \leq P(F)$$

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

Proof. For (a.), note that $F = (FE^c \text{ or } E)$, so that

$$P(F) = P(FE^c) + P(E) \geq P(E).$$

For (b.), E^c and E are mutually exclusive and "E or E^c"$= \Omega$. Thus

$$P(E) + P(E^c) = P(\Omega) = 1.\square$$

Thinking of events as subsets of a sample space has its advantages. Set theoretic identities may be applied to events. For example, *De Morgan's Laws*

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

for sets become

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

for events.

More generally, we have

$$\begin{aligned} (E_1 \text{ or } E_2 \text{ or} \cdots \text{or } E_n)^c &= E_1^c E_2^c \cdots E_n^c \\ (E_1 E_2 \cdots E_n)^c &= E_1^c \text{ or } E_2^c \text{ or} \cdots \text{or } E_n^c \end{aligned}$$

Venn diagrams can also aid in visualizing set identities. For example, Figure 3.1 shows the events $E\&F = EF$, E or F, $E \backslash F$ and $F \backslash E$.

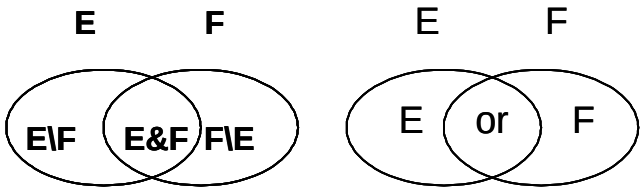

Figure 3.1 *Venn Diagrams.*

The Addition Theorem.

The diagrams of Figure 3.1 lead to the *Addition Theorem*, which generalizes the Addition Law to events which are not mutually exclusive.

Theorem 3. (The Addition Theorem.)

$$P(E \text{ or } F) = P(E) + P(F) - P(EF).$$

Proof. From the Venn diagram, and Addition Law

$$\begin{aligned} P(E \text{ or} F) &= P(E) + P(E^c F) = P(E) + [P(E^c F) + P(EF)] - P(EF) \\ &= P(E) + P(F) - P(EF). \square \end{aligned}$$

Example 1. In the roll of two dice, find the probability of at least one six.

Solution. Let S_1 and S_2 denote the events "six on first die" and "six on second". Then

$$\begin{aligned} P(S_1 \text{ or } S_2) &= P(S_1) + P(S_1) - P(S_1 S_2) \\ &= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}. \blacksquare \end{aligned}$$

Example 2. Of a given brand of ball point pens, 80 percent leak, 75 percent will not write, and 60 percent have both defects. What is the probability that a pen selected at random will write, but not leak?

Solution. Let L be the event that the pen leaks and W that it writes. We have

$$\begin{aligned} 1 - P(WL^c) &= P(W^c \text{ or } L) \\ &= P(W^c) + P(L) - P(W^c L) \\ &= 0.75 + 0.80 - 0.60 = 0.95 \end{aligned}$$

Thus, $P(WL^c) = 0.05$.$\blacksquare$

2.4 Classical Probability.

Suppose that $\Omega = \{e_1, e_2, \ldots, e_N\}$ is a finite sample space. Then the probability of an event E is the sum of the probabilities of all of the elementary events contained in E; in symbols

$$P(E) = \sum_{e_k \in E} p_k = \sum \{p_k : e_k \in E\}$$

If all the events e_k are equally likely, then every $p_k = 1/N$, and

$$P(E) = \frac{\#E}{N} = \frac{\#E}{\#\Omega}.$$

where $\#E$ is the number of elements of the set E. In Classical terminology, the total number of elementary events $\#\Omega$ is called the number of *total cases*, and the number $\#E$ of points of E, the number of *favorable cases*. Thus,

$$P(E) = \frac{\#E}{\#\Omega} = \frac{\text{Favorable Cases}}{\text{Total Cases}} = \frac{F}{T}.$$

This is the "*Classical definition of Probability*" of Laplace If one is able to set up a problem so that it has a sample space of elementary events *with equal probabilities*, then the computation of probabilities becomes a problem of Combinatorics: one merely has to count the total and favorable cases and take that quotient.

Example 1. For a very simple case, consider the roll of a single die. The sample space is

$$\{1, 2, 3, 4, 5, 6\}.$$

For a fair die, these numbers are equally likely. So, for example, the probability of rolling an odd prime - a 3 or 5 - is $2/6 = 1/3$.■

The trick here is to be sure you have set up the space so that *the points really are equally likely*. A classic blunder is due to no less a person than D'Alembert In in volume 5 of the famous *Encyclopedia* of 1754, he considered two tosses of a coin. According to D'Alembert, there are three possibilities 2 heads, 2 tails, and 1 head and 1 tail. So the sample space is

$$\Omega = \{HH, HT, TT\}$$

and the probability the probability of two heads is $1/3$. The problem is that in this space HT is twice as likely as HH and TT. The correct space for equal probabilities same as for two tosses:.

$$\{HH, HT, TH, TT\}.$$

Thus the correct probability of HT is $2/4 = 1/2$.

One is tricked into the error because the two results HT and TH look alike. If one thinks of tossing. say, a quarter and a dime, one is less libel to this error.■

Example 2. A similar problem occurs with two dice. There are two ways to roll a 10 - *six and four* or *two fives*. But *six and four* is twice as likely as *two fives*. As with the coins, one may be fooled because the dice look alike. Instead of two dice of the same color, think of one Red and one Green die. Then there are three ways to get the sum 10 - a five on both dice, 6 on the Red and 4 on the Green, and 4 on the Red and 6 on the Green. Two correspond to *six and four*, and only one to *two fives*

The correct sample space for equal probabilities consists of *all 36 ordered pairs* of numbers from 1 to 6. Thus e.g.

$$P(Sum = 10) = 3/36 = 1/12.,$$

$$P(Sum = 2) = 1/36 = 1/12;$$

$$P(Sum = 7) = 6/36.$$

D'Alembert was no mean mathematician; among other things, he discovered the wave equation. This is an indication of the subtlety that the ideas of probability may present initially. D'Alembert apparently persisted in his opinion. In a case like this, the best counter argument is an offer to gamble. Your opponent may remain unconvinced, but your winnings will provide consolation, and will increase in proportion to his obstinacy.

2.5 Combinatorial Probability Problems.

In this section, we shall consider some examples of Combinatorial methods.

Example 1. Three coins are tossed. Find the probability of exactly one Head.

Solution. The sample space consists of eight points, so the number of total cases is $T = 8$. The favorable cases are HTT, THT, TTH, so $F = 3$. Thus

$$P(\text{Exactly 3 Heads}) = 3/8.\blacksquare$$

Example 2: An urn contains 10 Red and 5 White balls. Two balls are drawn successively *without replacement*.

(a.) Find probability that both are Red.

Solution . The total cases consist of all possible sets of two balls selected from the urn. There are 15 balls in it, so the number of total cases is

$$T = \binom{15}{2}$$

The favorable cases are those sets of two balls selected from among the 10 Red balls. So the number of favorable cases is

$$F = \binom{10}{2}.$$

Thus,

$$P(\text{Two Reds}) = \frac{\binom{10}{2}}{\binom{15}{2}} = \frac{3}{7} = 0.43.$$

(b.) Find probability that the first Red ball is drawn on the third draw.

Solution Since the problem involves *order*, we must takes the total cases to consist of all *ordered* choices of three balls,

$$T = P(15, 3) = 15 \cdot 14 \cdot 13.$$

The favorable cases consist of all ordered choices with the first two balls from the Whites, and the last from the Reds. Thus, the number of favorable cases is

$$F = 5 \cdot 4 \cdot 10.$$

The probability is therefore

$$P = \frac{5 \cdot 4 \cdot 10}{15 \cdot 14 \cdot 13} = \frac{20}{273} = 0.073.\blacksquare$$

Example 3. An urn contains 3 Red, 2 White and 5 Blue balls. Three balls are drawn successively, and *replaced* in the urn after they are drawn.

(a.) Find the probability of balls of three different colors.

Solution. The Total cases are $T = 10^3$. To count the number of sequences of balls of different colors first choose one Red, one White and one Blue ball. There are $3 \cdot 2 \cdot 5$ ways to do so..Then arrange them into a sequence. There are $3!$ ways to do this. Hence, there are

$$F = 3 \cdot 2 \cdot 5 \cdot 3! = 180$$

favorable cases. Thus,

$$P(\text{Three colors}) = \frac{3 \cdot 2 \cdot 5 \cdot 3!}{10^3} = \frac{180}{100} = 0.180.$$

(b.) Find the probability of exactly one Red.

Solution. The Total cases are the same. For the Favorable, first decide which of 3 places the Red will come. Then choose successively the Red and the two nonreds - with repeats allowed. There are $3 \cdot 7 \cdot 7$ ways to do this, Thus,

$$F = 3 \cdot 3 \cdot 7 \cdot 7 = 441,$$

and

$$P(\text{ exactly one Red }) = \frac{3^2 \cdot 7^2}{10^3} = \frac{441}{1000} = 0.441.\blacksquare$$

Example 4. Answer the two questions of Example 3 if the balls are *not replaced* in the urn after they are drawn.

Solution. Take the sample space to consist of all $T = \binom{10}{3}$ sets of balls.
(a.) For exactly one Red, choose sets of one Red and two nonreds, so that

$$F = \binom{3}{1}\binom{7}{2}$$

and

$$P(\text{ exactly one Red }) = \frac{\binom{3}{1}\binom{7}{2}}{\binom{10}{3}} = \frac{63}{120} = 0.525.$$

(b.) For all three colors,

$$F = 3 \cdot 2 \cdot 5 = 30$$

and

$$P(\text{ Three colors }) = \frac{3 \cdot 2 \cdot 5}{\binom{10}{3}} = \frac{30}{120} = 0.25.$$

Examples 2 an 3 illustrate Sampling *with* and *without replacement*.■

The Birthday problem.

Example 5. (*The Birthday problem.*) In a room of r people, what is the probability $p(r)$ that *at least two have the same birthday*? Assume that all birthdays are equally likely, and ignore the existence of February 29.

Solution. We compute the probability $q(r) = 1 - p(r)$ that no two people have coincident birthdays. Consider a list of the r people, with their birthdays beside them. The number of total number of cases is $(365)^r$, since there are 365 choices of days to put by each name. The number of favorable cases is $P(365, r)$, since the list can have no repetitions. Thus

$$q(r) = \frac{P(365, r)}{(365)^r}$$

and

$$p(r) = 1 - q(r) = 1 - \frac{P(365, r)}{(365)^r}.$$

The probability $p(r)$ is surprisingly large. A brief table gives the idea.

$\mathbf{r}$	5	10	15	20	25	30	35	40	50	60
$\mathbf{p(r)}$	0.027	0.117	0.253	0.411	0.569	0.706	0.814	0.891	0.970	0.995

The author has tried this in many classes of size around of 30 or 40, and has only rarely failed to find coincident birthdays.■

Poker Hands.

Poker hands are a good source of combinatorial problems. A Poker Hand is a set of five cards drawn from a standard deck of 52 cards. Various combinations have certain rankings in the game of poker. The less likely a combination is to be drawn, the higher the ranking. The combinations are these, in order of increasing value.

Pair Two cards of the same rank; no other matches.

Two pair Two different pairs, and one unmatched card.

Straight Five cards in sequence, not all the same suit. (An Ace may be high or low.)

Flush Five cards of the same suit, not in sequence.

Full House Three of a kind and a pair.

Four of a kind Four of a kind.

Straight flush Five cards in sequence, all the same suit.

In each case, the sample space is the collection of five card sets from a 52 card deck. The number of Total cases is therefore

$$T = \binom{52}{5}.$$

We will compute the probabilities of two hands, and leave the rest as problems.

Example 6. Find the probability of a five card hand in which all cards are of the same suit.

Solution. Construct the hand in two steps.
First, select the suit: there are 4 choices.
Next, select five cards from the 13 cards of the suit: there are $\binom{13}{5}$ choices.
The number of favorable hands is therefore

$$F = 4 \cdot \binom{13}{5}$$

and the probability is

$$P = F/T = \frac{4 \cdot \binom{13}{5}}{\binom{52}{5}} = 0.0020$$

or about $1/500$.■

Example 7. Find the probability of two pair.

Solution. We construct the hand in three steps.

First: Select which two ranks the hand will have pairs of.(e.g. King and 10.).
There are$\binom{13}{2}$ choices.

Second: Select pairs from each of these two ranks.
There are $\binom{4}{2} \cdot \binom{4}{2}$ choices.

Third: Select a card from the remaining 44 cards not of these ranks.
There are 44 choices.

The total number of favorable hands is therefore

$$F = 44 \cdot \binom{13}{2} \cdot \binom{4}{2}^2 = 123,552$$

and the probability is

$$P = F/T = \frac{44\binom{13}{2}\binom{4}{2}^2}{\binom{52}{5}} = 0.0475.■$$

2.6 Sampling without Replacement.

The general formula for sampling *without replacement* is given by the Hypergeometric formula. We will formulate it in terms of the "*Ball and Urn*" model of *Jacob Bernoulli* (1712).

Theorem 4. (The Hypergeometric Formula.) *An urn has a Red and b White balls. If n balls are drawn without replacement, the probability of exactly k Red balls is*

$$\frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}.$$

Proof. The number of total cases is the number of ways to select m balls from the n available.

$$T = \binom{n}{m}.$$

The number of favorable cases is the number of ways to select k Red balls from the a available and $n-k$ White balls from the b available:

$$F = \binom{a}{k}\binom{b}{n-k}.\blacksquare$$

Example 1. Find the probably that a five card poker hand has exactly three face cards. The face cards are the King, Queen and Jack.

Solution. There are 12 face cards, so the probability is

$$\frac{\binom{12}{3}\binom{40}{2}}{\binom{52}{5}} = 0.06603.\blacksquare$$

Example 2. Find the probably that a 13 card bridge hand has exactly two aces.

Solution. The probability is

$$\frac{\binom{4}{2}\binom{48}{11}}{\binom{52}{13}} = 0.2135.\blacksquare$$

Generalized Hypergeometric formula.

For sampling without replacement when there are several possibilities on each draw, there is a generalization of the Hypergeometric formula.

Theorem 5. (The Generalized Hypergeometric Formula.) *An urn has n balls of r different colors, with n_1 balls of the first color, ..., and n_r balls of the r^{th} color, where*

$n_1 + \cdots + n_r = n$. *If* m *balls are drawn without replacement, the probability of exactly* k_i *balls of the* i^{th} *type, where* $k_1 + \cdots + k_r = m$.

$$\frac{\binom{n_1}{k_1} \cdots \binom{n_r}{k_r}}{\binom{n}{m}}.$$

Proof. The proof is a simple modification of the argument for two colors.■

Example 3. A urn contains 6 Red, 4 White and 2 Blue balls. If half of the balls are drawn at random, what is the probability of obtaining 3 Red, 2 White and 1 Blue?

Solution. The probability is

$$\frac{\binom{6}{3}\binom{4}{2}\binom{2}{1}}{\binom{12}{6}} = \frac{20}{77} = 0.25974.\blacksquare$$

Example 4. Find the probably that a 13 card Bridge hand has 5 spades, 4 hearts, 3 diamonds and 1 club.

Solution. The probability is

$$\frac{\binom{13}{5}\binom{13}{4}\binom{13}{3}\binom{13}{1}}{\binom{52}{13}} = 0.00539.\blacksquare$$

2.7 *Inclusion-Exclusion.

The probabilistic version of the Inclusion-Exclusion formula is a generalization of the Addition Theorem.

Theorem 6. (Inclusion-Exclusion.) *Let* E_1,E_2,...,E_n *be any events, and*

$$S_k = \sum_{i_1 < i_2 < \cdots < i_k} P(E_{i_1} E_{i_2} \cdots E_{i_k}).$$

Then

$$P(E_1 \text{ or } E_2 \text{ or } \cdots \text{ or } E_n) = S_1 - S_2 + S_3 - \cdots (-1)^{n+1} S_n. \tag{2.1}$$

Proof. The proof is by induction. Suppose first that there are only two sets. For $n = 2$, the formula reads,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 E_2).$$

This is just the Addition Theorem, so the formula holds for $n = 2$.

Suppose now that the formula holds for some integer n. Let $F = E_1 \cup E_2 \cup \cdots \cup E_n$. By the Addition Theorem,

$$\begin{aligned} & P(E_1 \cup E_2 \cup \cdots E_n \cup E_{n+1}) \qquad (2.2) \\ = \; & P(E_1 \cup E_2 \cup \cdots E_n) + P(E_{n+1}) - P((E_1 \cup E_2 \cup \cdots E_n) \cap E_{n+1}) \\ = \; & P(E_1 \cup E_2 \cup \cdots E_n) + P(E_{n+1}) - P(E_1E_{n+1} \cup E_2E_{n+1} \cup \cdots \cup E_nE_{n+1}) \end{aligned}$$

where we have used DeMorgan's Law..

By induction, we may apply the formula for n events to both terms. We have

$$\begin{aligned} & P(E_1 \cup E_2 \cup \cdots E_n) \\ = \; & \sum_{k=1}^{n} P(E_k) + \cdots + \sum_{i_1<i_2<\cdots<i_k} P(E_{i_1}E_{i_2}\cdots E_{i_k}) + \cdots + P(E_1E_2\cdots E_n) \end{aligned}$$

and

$$\begin{aligned} & P(E_1E_{n+1} \cup E_2E_{n+1} \cup \cdots \cup E_nE_{n+1}) \\ = \; & \sum_{k=1}^{n} P(E_kE_{n+1}) + \cdots + \sum_{i_1<i_2<\cdots<i_k} P(E_{i_1}E_{i_2}\cdots E_{i_{k-1}}E_{n+1}) \\ & + \cdots + P(E_1E_2\cdots E_nE_{n+1}). \end{aligned}$$

But

$$\begin{aligned} & \sum_{i_1<\cdots<i\le n_r} P(E_{i_1}E_{i_2}\cdots E_{i_r}). + \sum_{i_1<\cdots<i_{r-1}} P(E_{i_1}E_{i_2}\cdots E_{i_{r-1}}E_{n+1}). \\ = \; & \sum_{i_1<\cdots<i_r\le n+1} P(E_{i_1}E_{i_2}\cdots E_{i_r}) \end{aligned}$$

so the terms on the right side of (2.2) add up to the terms of (2.1) for $n+1$. □

The Inclusion-Exclusion formula of section 1.5 can also be proved by induction in this way.

Example 1. Urn contains 3 Red, 3 White and 3 Blue balls. Three balls are drawn *without* replacement. Find the probability of being void in at least one color.

Solution. Let $R =$"contains a Red.", etc. Then

$$\begin{aligned} P(R^c \text{ or } W^c \text{ or } B^c) &= P(R^c \text{ or } W^c \text{ or } B^c) \\ &= [P(R^c) + P(W^c) + P(B^c)] \\ &\quad - [P(R^cW^c) + P(R^cB^c) + P(W^cB^c)] + P(R^cW^cB^c) \\ &= 3P(R^c) - 3P(R^cW^c) + P(R^cW^cB^c) \\ &= 3 \cdot \frac{\binom{6}{3}}{\binom{9}{3}} - 3 \cdot \frac{\binom{3}{3}}{\binom{9}{3}} + 0 = \frac{19}{28} = 0.68. \end{aligned}$$

As a check, we have an alternate method::

$$P(R^c \text{ or } W^c \text{ or } B^c) = 1 - P(\text{ Three colors }) = 1 - \frac{3 \cdot 3 \cdot 3}{\binom{9}{3}} = 1 - \frac{9}{28} = \frac{19}{28}.\blacksquare$$

Example 2 A five card hand is called a *Happy Family* if it contains at least one King, at least one Queen and at least one Jack. Find the probability of a Happy Family.

Solution. We compute the probability of the event $H^c$the a hand is *not* a Happy Family. By Theorem 11, this is

$$\begin{aligned} P(H^c) &= P(K^c \text{ or } Q^c \text{ or } J^c) \\ &= P(K^c) + P(Q^c) + P(\ J^c) \\ &\quad -P(K^cQ^c) - P(K^cJ^c) - P(Q^cJ^c) + P(K^cQ^cJ^c) \\ &= 3\frac{\binom{48}{5}}{\binom{52}{5}} - 3\frac{\binom{44}{5}}{\binom{52}{5}} + \frac{\binom{40}{5}}{\binom{52}{5}} = 0.976. \end{aligned}$$

Thus,

$$P(H) = 1 - P(H^c) = 0.024.\blacksquare$$

2.8 *Rencontre.

A classic application of Inclusion-Exclusion is to the game of *Rencontre*, first analyzed by Montmort in his 1708 book *Essai d'Analyse sur les Jeux de Hasard*. The problem can be phrased in several ways. One of the most amusing is following.

Problem of the Hat Check Girl.

Gentlemen at a club check their hats in the cloakroom. As they leave, the young lady in charge of the cloakroom distributes the hats at random, with no regard to its proper owner. If there are n gentlemen, what is probability that at least one gentleman receives his own hat?

In particular, if n is large, does the probability get small, because it is very unlikely for any given man to receive his hat? Or does it get large because there are so many more opportunities for a hit?

Solution. Let H_k denote the event that the k^{th} gentleman gets his hat. We want to find $P(H_1 \text{ or } H_2 \text{ or} \cdots \text{ or } H_n)$. By Inclusion-Exclusion,.

$$P(H_1 \text{ or } H_2 \text{ or} \cdots \text{or } H_n) = S_1 - S_2 + S_3 - \cdots + (-1)^{n+1} S_n$$

By symmetry, the probabilities $P(E_{i_1}E_{i_2}\cdots E_{i_k})$ are all equal if $_1 < i_2 < \cdots < i_k$. Hence,

$$\begin{aligned} S_1 &= \sum_{i=1}^{n} P(H_i) = nP(H_1) = n \cdot \frac{1}{n} = 1 \\ S_2 &= \sum_{1\le i<j\le n} P(H_iH_j) = \binom{n}{2} P(H_1H_2) = \frac{n\,(n-1)}{2!}\frac{1}{n\,(n-1)} = \frac{1}{2!} \end{aligned}$$

.....

$$\begin{aligned} S_k &= \sum_{i_1<i_2<\cdots<i_k} P(E_{i_1}E_{i_2}\cdots E_{i_k}) = \binom{n}{k} P(E_1E_2\cdots E_k) \\ &= \frac{n\,(n-1)\cdots(n-k+1)}{k!}\frac{1}{n\,(n-1)\cdots(n-k+1)} = \frac{1}{k!} \end{aligned}$$

and we find that

$$p(n) = P(H_1 \text{ or } H_2 \text{ or} \cdots \text{or } H_n) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1}\,\frac{1}{n!}$$

However, from the exponential series,

$$\lim_{n\to\infty}\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n\,\frac{1}{n!}\right) = \frac{1}{e}$$

this probability converges to

$$1 - \frac{1}{e} = 0.632.$$

The convergence is very rapid, so this probability is *very nearly independent of* n *for large* n. The two possibilities we mentioned above compensate!

Moreover, since we have the terms of an alternating series, the error is less that the next term in the expansion, and has the same sign. In particular, it is *smaller for any even* n *than for any odd* n.

A brief table of $p(n)$ shows how quick the convergence is.

$\mathbf{n}$	2	3	4	5	6
$\mathbf{p(n)}$	0.500	0.667	0.625.	0.633	0.632

It is sometimes assumed that the Hat Check Girl was not very bright. It is the author's opinion, however, that the members of the club were notoriously poor tippers, and got what they deserved. ■

2.9 Problems.

(1.) A box contains 3 marbles, 1 red, 1 green, and 1 blue. Consider an experiment that consists of taking 1 marble from the box, then replacing it in the box and drawing a second marble from the box. Describe the sample space. Repeat when the second marble is drawn without first replacing the first marble.

(2.) Among the digits $1, 2, 3, 4, 5$, first one is chosen, then a second selection is made among the remaining four digits. Assume that all 20 possible results have the same probability. Find the probability that an odd digit will be selected (a.) the first time, (b.) the second time, (c.) both times.

(3.) Two cards are drawn (without replacement) from a deck. Find the probability that the *second* card is the ace of spades.

(4.) The numbers 1, 2, 3, 4, and 5 are written down on five cards. Three cards are drawn at random, in succession, and the digits thus obtained are written from left to right in the order in which they were drawn. What is the probability that the resulting 3-digit number is *even*.

(5.) You are playing stud poker. Your cards are: A of spades, A of hearts, K of diamonds, and K of hearts. The A of spades is concealed (the hole card), but the other three cards are visible to your opponent. Your opponents cards are: A of diamonds, Q of spades, Q of diamonds, and a hole card which you cannot see. You will each be dealt another card, your opponent receiving his first. What is the probability that your card will be an ace or a king?

(6.) There are three sticks. Both ends of the first are painted red and both ends of the second blue, while the third has one red end and one blue end. One stick is seized by the end, at random, and held out with only one end visible. It is blue. What is the probability that the *other* end is blue?

(7.) The King has one sibling. What is the probability that it is his sister?

(8.) (*Bertrand's Boxes. 1889.*) There are three identical boxes, one containing two gold coins, one containg two silver coins and one containing one gold and one silve coin. A box is selected, and one coin drawn form it. The coin is giold. What is the probability that the other coin in the box is also gold?

(9.) A tax evader estimates that he has probability 0.2 of being caught by the Feds, and 0.3 of being caught by the State. If the two agencies work independently, the probability of being caught by both is 0.06. What is the probability he gets away with it.

(10.) A student applies for a fellowship at two different universities. He estimates that the probability being awarded a fellowship to be $1/3$ by the first, $1/4$ by the second and $1/6$ by both. What is the probability that ha rfeceives at least one offer?

(11.) On a multiple choice exam with three possible answers for each of the five questions, what is the probability that a student would get four or more correct answers just by guessing?

(12.) (*Galileo, c.1620.*). In the roll of 3 dice, the sum 10 can appear in 6 ways (6-3-1, 6-2-2, 5-4-1, 5-3-2, 4-4-2, and 4-3-3), and the sum 12 also in 6 ways. Nevertheless, gamblers of Galileo's time observed that 10 appears more frequently than 12. Explain this and compute correctly the probabilities of 10 and 12. Note the magnitude of the difference in probabilities, which was, apparently, detected at the gambling tables.

(13.) (*Huygens' 3rd Problem. 1657.*) There are 40 cards forming 4 sets of 10 cards each (i.e., four 10-card suits). A plays with B and undertakes in drawing four cards to obtain one of each set. Find A's probability to win.

(14.) An elevator starts with 7 passengers and stops at 10 floors. What is the probability that no two passengers leave at the same floor? Assume that all arrangements of discharging the passengers have equal probability. Feller states: "When the event was once observed, the occurrence was deemed remarkable, and odds of 1000 to 1 were offered against a repetition."?

(15.) A deck of cards is dealt out. What is the probability that the fourteenth card dealt is an ace? What is the probability that the first ace occurs on the fourteenth card?

(16.) A blackjack is a two-card hand consisting of one ace and a second card which is a ten, jack, queen, or king?

(a.) If two cards are randomly selected from an ordinary playing deck, what is the probability that they form a blackjack?

(b.) Two two-card hands are dealt , one to the player and one to the dealer. What is the probability that neither is a blackjack?

(17.) A bus with 8 stops to make has 5 passengers. What is the probability that they all get off at different stops?

(18.) A *skeet* is a five card poker hand containing a nine, a five, a deuce, one card between the nine and the five, and another between the five and the deuce. Find the probability of a skeet.

(19.) (*Samuel Pepys.*) Which is easier, to throw at least one ace with six dice or to throw at least two aces with 12 dice? Justify your answer by computing both probabilities. This problem was proposed to Isaac Newton by the famous Samuel Pepys, who was "contemplating a wager, ten pounds deep", although he did not admit this to Newton.

(20.) A *Big Dog* is a five card poker hand with Ace high, 9 low, and no pair. Find the probability that five cards drawn at random form a *Big Dog*. (*Translation*: The hand contains an Ace, a 9, and three cards in rank between the ace and the nine, but no two cards of the same rank. *Example*: A, K, J, 10, 9.).

Section 2.9 Problems.

(21.) Four cards are drawn from a standard deck. What is the probability that they are of four different suits?

(22.) Find the probability of the following poker hands.

(a.) A *Pair:* Two cards of the same rank; no other matches.

(b.) A *Straight:* Five cards in sequence, not all the same suit. An Ace may be high or low.

(c.) A *Full House:* Three of a kind and a pair.

(d.) *Four of a kind*.

(e.) A *Straight flush:* Five cards in sequence, all the same suit.

(23.) (*Poker dice.*) Five dice are rolled. Find the probability of

(a.) One pair

(b.) Two pair

(c.) Three of a kind

(d.) No two alike

(e.) A full house; i.e. three of a kind and a pair

(f.) Four of a kind

(g.) Five of a kind

(h.) Low straight (1 - 2 - 3 - 4 - 5.)

(i.) High straight (2 - 3 - 4 - 5 - 6.)

(24.) (*Yarborough.*) A 13 card Bridge hand containing no card higher than a nine is referred to as a Yarborough, since the second Earl of Yarborough is said to have offered odds of 1000 to 1 against drawing such a hand. Find the probability of drawing a Yarborough. (The Ace is high, not low.)

(25.) Find the distribution of the number of aces in a 13-card bridge hand.

(26.) Twenty-six cards containing $n = 6$ spades are divided into two 13-card hands.

(a.) What is the probability of a $4 - 2$ split (3 spades in one hand and 2 in the other)?

(b.) Of $3 - 3$ split?

(27.) A 13-card Bridge hand is said to have a $5 - 4 - 3 - 1$ distribution if it contains 5 cards of one suit, 4 of another, 3 of a third and 1 card of the fourth.

(a.) What is the probability of a $5 - 4 - 3 - 1$ distribution?

(b.) Of a $5 - 3 - 3 - 2$ distribution?

(c.) Of a $4-3-3-3$ distribution?

(28.) A box contains 90 good and 10 defective screws. If 10 screws are used, what is the probability that none is defective?

(29.) *(Fisher's Tea Experiment.)* A lady states that she prefers milk to be added to her teacup before the tea is poured in, and claims to be able to tell if milk was added before or after the tea. A test of her abilities is proposed as follows: four cups of tea are prepared with milk added before the tea is poured, and four with it added afterwards The eight cups are arranged in random order, and she is asked to identify which is which. What is the probability she can get them all right merely by guessing?

Suppose eight cups are prepared, and for each cup a coin is tossed to determine whether milk will be added before or after the tea. Is the answer the same?

(30.) Compute the probability that a 13 card bridge hand contains the ace and king of some suit.

(31.) A 13 - card Bridge hand is *void* in a suit if it contains no cards of that suit. What is the probability that a Bridge hand is void in at least one suit?

(32.) Two players turn over cards one after the other from two shuffled decks. An *encounter* is said to occur if both players simultaneously. What is the probability that an encounter occurs before the decks are exhausted?

Chapter 3
Conditional Probability.

3.1 Conditional Probability.

Suppose that two cards are to be drawn in succession from a standard deck. What is probability that the second card is an ace?

Player A says:"It depends: it is either $4/51$ or $3/51$ depending on the first card."

Player B says: "This is just a more complicated way of picking a random card from the deck, so the probability $4/52$."

As the problem is stated, B is correct. *If bets are to be made before any cards are drawn,* then the correct odds are $12:1$, corresponding to a probability of $4/52$.

In a sense, however, A is also correct. The probabilities $3/51$ and $4/51$ are what are called *conditional probabilities*. They are *probabilities belonging to different experiments*.

On the one hand, we have the experiment $\mathcal{E}$: *shuffle and draw two successive cards.* The probability the second card is an ace is $4/52$.

On the other hand, we have the following experiment $\mathcal{E}'$: *shuffle and draw one card. If it is an ace, draw a second. If it is not an ace, reshuffle and repeat until an ace is drawn on the first draw and then draw a second card.*

If the first card is not an ace, *there is no trial of* $\mathcal{E}'$; *a trial of* $\mathcal{E}'$ *is a trial of* $\mathcal{E}$ *on which the first card is an ace*. The probability of an ace on the second draw in $\mathcal{E}'$ is $3/51$. We call this the *conditional probability* that the second card is an ace, given that the first card is an ace.

Similarly, $4/51$ is the conditional probability that the second card is an ace given that the first card is *not* an ace. It corresponds to the experiment $\mathcal{E}''$ whose trials are trials of $\mathcal{E}$ on which the first card is *not* an ace.

In general, let $\mathcal{E}$ be a random experiment, and A an event of $\mathcal{E}$. We define a new experiment $(\mathcal{E} \mid A)$ of $\mathcal{E}$ given A so that a trial of $(\mathcal{E} \mid A)$ is a trial of $\mathcal{E}$ on which A occurs. Any event E of $\mathcal{E}$ is alsoan event of $(\mathcal{E} \mid A)$ and the probability of E in the experiment $(\mathcal{E} \mid A)$ is written $P(E \mid A)$.

Thus, in the example above, we would write

$$P(\text{Ace on second draw} \mid \text{Ace on first draw}) = \frac{3}{51}.$$

Similarly,

$$P(\text{Ace on second draw} \mid \text{No ace on first draw}) = \frac{4}{51}.$$

We remark that it necessary that $P(A) > 0$, since otherwise, we have no guarantee that A will ever occur, and hence no guarantee that the experiment $(\mathcal{E} \mid A)$ is indefinitely repeatable.

The Law of Conditional Probability.

Consider now an experiment $\mathcal{E}$, and two events A and E of $\mathcal{E}$. Let $\mathcal{E}$ be subjected to a large number n of trials. Suppose that A occurs on m of these trials. Then

$$P(A) \simeq \frac{m}{n}.$$

Suppose further that the event $(E$ and $A)$ occurs on k trials. This is the same as saying that the event E occurs on k of the m trials on which A occurs. Thus, there have been m trials of $(\mathcal{E} \mid A)$ and the event E has occurred on k of these m trials. Hence,

$$P(E \mid A) \simeq \frac{k}{m}.$$

But

$$P(E \text{ and } A) \simeq \frac{k}{n} = \frac{k}{m}\frac{m}{n} \simeq P(E \mid A)\, P(A)$$

and so we obtain the

Law of Conditional Probability. *For any events E and A,*

$$P(E \text{ and } A) = P(E \mid A)\, P(A).$$

As a first application, we solve two problems that were solved by combinatorial methods in Example 2 of section 2.5.

Example 1. An urn contains 10 Red and 5 White balls. Two balls are drawn successively *without* replacement.

(a.) Find probability that both are Red.

Solution. Letting R_1 denote the event that the "*Red on the first draw*", etc. we have

$$P(R_1R_2) = P(R_1)P(R_2 \mid R_1) = \frac{10}{15} \cdot \frac{9}{14} = \frac{20}{273}.$$

(b.) Find probability that the first Red ball is drawn on the third draw.

Solution. We have

$$P(W_1W_2R_3) = P(W_1)P(W_2 \mid W_1)\, P(R_3 \mid W_1W_2) = \frac{5}{15} \cdot \frac{4}{14} \cdot \frac{10}{13} = \frac{3}{7}.$$

Note how much more naturally the computation flows with conditional probabilities.■

Example 2. *Polya's urn model*. An urn contains 1 Red and 1 White ball. Draw, replace and add one of the color drawn. Repeat two more times. What is the probability of three successive reds?

Solution. We have

$$\begin{aligned} P(R_1R_2R_3) &= P(R_1)P\left(R_2R_3 \mid R_1\right) \\ &= P(R_1)P\left(R_2 \mid R_1\right)P\left(R_3 \mid R_1R_2\right) \\ &= \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{3}{4} = \frac{1}{4}. \end{aligned}$$

This computation would be considerably more difficult by combinatorial methods. ■

The Law of Total Probability.

Frequently, a good way to compute probabilities is to sum over all possible contingencies when the conditional probabilities can be found in each case.

Example 3. There are two urns. Urn I contains 5 Red and 5 White balls, while Urn II contains 15 Red and 5 White balls. An urn is selected at random and a ball drawn. Find the probability that the ball is Red.

Solution. Let R be the event that the ball is Red, and I and II respectively be the events that Urns I and II are selected. Then

$$\begin{aligned} P(R) &= P(I \text{ and } R) + P(II \text{ and } R) \\ &= P(I)P\left(R \mid I\right) + P(II)P\left(R \mid II\right) \\ &= \frac{1}{2}\cdot\frac{5}{10} + \frac{1}{2}\cdot\frac{15}{20} = \frac{5}{8}. \end{aligned}$$ ■

Example 4. Consider the problem that we began with, the probability of an ace on second draw.

Solution. Let A_1be the event "*Ace on first draw*", and A_2 the event "*Ace on second draw*". Then

$$\begin{aligned} P(A_2) &= P(A_2 \text{ and } A_1) + P(A_2 \text{ and } A_1^c) \\ &= P(A_1)P\left(A_2 \mid A_1\right) + P(A_1^c)P\left(A_2 \mid A_1^c\right) \\ &= \frac{4}{52}\frac{3}{51} + \frac{48}{52}\frac{4}{51} = \frac{4}{52}. \end{aligned}$$

What we have done is to *average the two conditional probabilities, weighted according to the probabilities that the conditioning events occur*. We see that it all comes out in the wash, and the result is just the result of selecting a card at random. ■

To formulate the general result, let $E_1, E_2, \ldots, E_n$ be a set of *mutually exclusive and exhaustive events* for the experiment $\mathcal{E}$. That is to say, *one and only* one of the events $E_1, E_2, \ldots, E_n$ will occur on every trial of $\mathcal{E}$.

As an example, if $\mathcal{E}$ is the roll of a die, and E_k is the event that the number k shows up, then $E_1, E_2, \ldots, E_6$ is a set of mutually exclusive and exhaustive events for this experiment.

With this definition, we then have:

Theorem 1. (Law of Total Probability.) *Let $E_1, E_2, \ldots, E_n$ be a set of mutually exclusive and exhaustive events, then*

$$P(A) = \sum_{k=1}^{n} P(E_k) P(A \mid E_k).$$

Proof. The events $AE_1, AE_2, \ldots, AE_n$ are mutually exclusive and $A = AE_1 \cup AE_2 \cup \ldots \cup AE_n$. Hence,

$$P(A) = \sum_{k=1}^{n} P(AE_k) = \sum_{k=1}^{n} P(E_k) P(A \mid E_k). \square$$

Example 5. A player draws a card from a well shuffled deck, and then rolls a pair of fair dice. In order to win, if he draws an Ace, he must then roll any doubles; if he draws a face card (K, Q, J), he must roll a double $2, 3, 4$ or 5; if he draws any other card ($2 - 10$), he must roll a double six or a double one. What is his probability to win?

Solution. Averaging over the possibilities of the card drawn, we have

$$\begin{aligned} P(\text{Win}) &= P(\text{Ace}) P(\text{Win} \mid \text{Ace}) + P(\text{Face}) P(\text{Win} \mid \text{Face}) + P(\text{2 - 10}) P(\text{Win} \mid \text{2 - 10}) \\ &= P(\text{Ace}) P(\text{Doubles}) + P(\text{Face}) P(\text{Double 2 - 5}) + P(\text{2 - 10}) P(\text{Double six or one}) \\ &= \frac{4}{52}\frac{6}{36} + \frac{12}{52}\frac{4}{36} + \frac{36}{52}\frac{2}{36} = \frac{1}{13}. \blacksquare \end{aligned}$$

3.2 Bayes Theorem.

Let A and B be two events. *Suppose that we know $P(A \mid B)$ and that we want to find $P(B \mid A)$*. A trick attributed to Thomas Bayes is frequently useful. Expanding $P(AB)$ in two ways gives

$$P(A \mid B) P(B) = P(AB) = P(B \mid A) P(A)$$

so that

$$P(B \mid A) = P(A \mid B) \frac{P(B)}{P(A)} \tag{3.1}$$

A typical example is the following.

Example 1. There are three urns. Urn I contains 1 Red and 4 White balls, Urn II contains 10 Red and 5 White and Urn III contains 2 Red and 3 White balls.

An urn is selected at random and a ball drawn. It is Red. Find the probability that it came from Urn II.

Solution. We want to find $P(II \mid R)$. We know that $P(R \mid II) = 10/15 = 1/3$ By Bayes' trick,

$$P(II \mid R) = P(R \mid II) \frac{P(II)}{P(R)}.$$

From the Law of Total Probability, we compute that

$$\begin{aligned} P(R) &= P(I)P(R \mid I) + P(II)P(R \mid II) + P(III)P(R \mid III) \\ &= \frac{1}{3} \cdot \frac{1}{5} + \frac{1}{3} \cdot \frac{10}{15} + \frac{1}{3} \cdot \frac{2}{5} = \frac{19}{45} \end{aligned}$$

and hence we obtain that

$$P(II \mid R) = \frac{P(R \mid II) P(II)}{P(R)} = \frac{\frac{1}{3} \cdot \frac{10}{15}}{\frac{19}{45}} = \frac{10}{19}. \blacksquare$$

The general result for more than two outcomes is

Theorem 2. (Bayes Theorem.) *Let $E_1, E_2, \ldots, E_n$ be a set of mutually exclusive and exhaustive events and A an event. Then*

$$P(E_r \mid A) = \frac{P(A \mid E_r) P(E_r)}{\sum_{k=1}^{n} P(A \mid E_k) P(E_k)}.$$

Example 2. Of three machines - A, B, and C - for manufacturing bolts, machine A makes 40% of a company's total output, while B and C each make 30%. If 5% of A's bolts, 10% of B's and 2% of C's are defective, what is the probability that a given defective bolt was made by machine A?

Solution. Let D be the event that a bolt is defective, A the event that it is made by machine A, etc. Then

$$\begin{aligned} P(D) &= P(D \mid A) P(A) + P(D \mid B) P(B) + P(D \mid C) P(C) \\ &= (0.05)(0.40) + (0.10)(0.30) + (0.02)(0.30) = 0.056 \end{aligned}$$

Hence,

$$P(A \mid D) = \frac{P(D \mid A) P(A)}{P(D)} = \frac{(0.05)(0.40)}{0.056} = \frac{5}{14} = 0.357. \blacksquare$$

Remember: The key to knowing when to use (3.1)is this:

If you want $P(A \mid B)$ but only know $P(B \mid A)$, use Bayes formula.

Drug Testing. There is an interesting application of Bayes Theorem is to Drug Tests. Suppose you have a test for a drug which is "99% sensitive and 99% specific". In English, this says that a drug user will fail 99% of the time and, a nonuser will pass 99% of the time.

Suppose that only 0.5% of the population actually uses the drug. The probability that a randomly chosen person fails the test is

$$\begin{aligned} P(Failure) &= P(Failure \mid User)\, P(User) + P(Failure \mid NonUser)\, P(NonUser) \\ &= (0.99)(0.005) + (0.01)(0.995) = 0.0149 \end{aligned}$$

By Bayes Theorem, the probability that a person who fails is actually a user is

$$\begin{aligned} P(User \mid Failure) &= \frac{P(Failure \mid User)\, P(User)}{P(Failure)} \\ &= \frac{(0.99)(0.005)}{0.0149} = 0.3322 \end{aligned}$$

Thus, about $2/3$ of the failures are false positives.■

3.3 Independent Events.

We come now to one of the most important concepts of probability theory. As stated in the introduction to Chapter 2, Probability Theory is based on the intuitive notion that the results of successive trials of a random experiment are independent of one another. The notion of the independence of two events can be expressed precisely in terms of conditional probabilities. The idea is simply that A is independent of B if the occurrence of B does not affect the probability of the occurrence of A.

Definition 1. Two events A and B are *independent* iff

$$P(A \mid B) = P(A).$$

Example 1. Events associated with different trials of the same experiment are independent. For example, the event of Heads on the *first* toss of a coin and Heads on the *second* toss are independent.

The *Multiplication Law of Probability* is simply the Law of Conditional Probability in the case that the events are independent.

Theorem 3. **(The Multiplication Law.)** *A and B are independent iff*

$$P(A \,\&\, B) = P(A)\, P(B)$$

Hence, If A is independent of B, then B is independent of A.

Proof. By the Law of Conditional Probability, if A and B are independent,then

$$P(A \,\&\, B) = P\,(B)\,P(A \mid B) = P\,(B)\,P(A).$$

Conversely, if

$$P(A\&B) = P\,(A)\,P(B)$$

then

$$P(A \mid B) = \frac{P(A \,\&\, B)}{P\,(B)} = \frac{P\,(A)\,P(B)}{P(B)} = P\,(A)\,.\blacksquare$$

As a consequence, the definition of independence is symmetrical in A and B.

Corollary 1. *If A is independent of B, then B is independent of A.*

Example 2. Find the probability of rolling a double six with two dice.

Solution. Let S_1 and S_2 denote respectively, "six on the first die" and "six on the second die". These events are independent, so

$$P(S_1S_2) = P(S_1)P(S_2) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}.\blacksquare$$

Independent Sets of Events.

Definition 2. *Independent sets of events*. A set $E_1, E_2, \ldots, E_n$ of events is an *independent set of events* iff

$$P(E_{i_1}E_{i_2}\ \ldots E_{i_r}) = P(E_{i_1})P(E_{i_2}\,)\cdots P(E_{i_r})$$

for every subset of $E_1, E_2, \ldots, E_n$.

Again, if the events $E_1, E_2, \ldots, E_n$. belong to different trials of the same experiment, the set $E_1, E_2, \ldots, E_n$.is independent.

Example 3. Find the probability of tossing three successive Heads with a fair coin.

Solution. Let H_1, H_2 and H_3 denote respectively, Heads on the first, second and third tosses. These events are independent, so

$$P(H_1H_2H_3) = P(H_1)P(H_2)P(H_3) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}.\blacksquare$$

Example 4. Find the probability of rolling at least one six in three rolls of a die.

Solution. Let F_1, F_2 and F_3 denote respectively the *failure to roll a six* on the first second and third rolls respectively. These events are independent, so

$$P(\text{ No six }) = P(F_1F_2F_3) = P(F_1)P(F_2)P(F_3) = \frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6} = \left(\frac{5}{6}\right)^3.$$

Hence,

$$P(\text{ At least one six }) = 1 - P(F_1F_2F_3) = 1 - \left(\frac{5}{6}\right)^3 .\blacksquare$$

***Remark**. If $E_1, E_2, \ldots, E_n$. is an independent set of events, then, in particular, every pair E_{i_1} and E_{i_2} of events of the set, are independent. This property is called *pairwise independence*. Note that *pairwise independence is not sufficient for independence*. As an example, consider two rolls of a fair die. Let S_1 be a six on the first roll, S_2 a six on the second roll and E that the sum on the two rolls is an even number. Then the set S_1, S_3, E is *pairwise* independent. However, it is not an independent set, since if S_1 and S_3 both occur, then E is bound to occur. (See Problem 18.)

A Comment on the Addition and Multiplication Laws.

The two basic Laws of Probability - the Addition and Multiplication Laws - each hold only under certain conditions. The *Addition Law*

$$P(A \; or \; B) = P(A) + P(B)$$

holds if A and B are *mutually exclusive* events, and the *Multiplication Law*

$$P(A \;\&\; B) = P(A)P(B)$$

holds if A and B are *independent* events.

In each case, there is a more general rule that holds in genera for all events: the *Addition Theorem*

$$P(A \; or \; B) = P(A) + P(B) - P(A \;\&\; B)$$

and the *Law of Conditional Probability*

$$P(A \;\&\; B) = P(A \mid B)P(B).$$

It would be well to get straight now, if you have not already, the difference between independent and mutually exclusive. They are *entirely different*. In fact, the only way two events can be *both* mutually exclusive and independent is if at least one of them has probability zero. For in such a case, we have

$$P(A)P(B) = P(AB) = 0$$

so at least one factor must be zero.

3.4 Problems.

(1.) The probabilities of a person selected at random having the four blood types O, A, B, and AB are given by the following table.

Type	**O**	**A**	**B**	**AB**
Probability	0.45	0.42	0.10	0.03

These probabilities are independent of the sex of the person chosen. What is the probability that a married couple chosen at random have the same blood type?

(2.) (*Polya's Urn Scheme.*) An urn contains b black and r red balls. A ball is drawn at random, replaced, and then c balls of the color drawn are added. The process is repeated.

(a.) Find the probability of drawing two black balls on the first two draws.

(b.) Given that the second ball is black, what is the probability that the first was black?

(c.) What is the probability that the n^{th} ball is black?

(3.) (*Polya's Urn Scheme.*) Show that the probability of drawing a given sequence of n Black and m Red balls is independent of the order in which they are drawn. Find the probability of drawing n Black and m Red balls in whatever order.

(4.) A offers to bet at even odds that the toss of a coin, which A asserts is fair, will come up Heads. B suggests that the game be changed: The coin will be tossed twice, with A winning if the two tosses result in one head and one tail, while B will win if there are two heads or two tails. A declines the bet. Explain why.

(5.) There are15 tennis balls in a box, of which 9 have been used. Three balls are randomly chosen and used, and returned to the box. Later, three balls are chosen from the box. What is the probability that all three have never been used?

(6.) A, B, and C are playing a game which equal chance to win. The first to win a certain number of games receives the stake. A and B each need just one more game to win, but C needs 2 more games. Compute the probabilities that each contestant wins the stake.

(7.) Let A and B be two events which are both independent and mutually exclusive. Prove that at least one of these events has probability zero.

(8.) Three dice are rolled. If they are all different, what is the probability that none shows a six?

(9.) There are two urns, a Red urn and a White urn. The Red urn contains 6 Red and 4 White balls, and the White urn contains 3 Red and 7 White balls. A ball is drawn from the Red urn, replaced, and a second ball drawn from the urn of the of the same color as the ball just drawn The process is repeated. (That is, one changes urns according to the color of the ball drawn.). What is the probability that the third ball is drawn from the Red urn?

(10.) (*Craps.*) The game of craps is played as follows. Two dice are rolled. The player loses on $2, 3$ or 12, and wins on 7 or11. If none of these appears, the number thrown becomes the players "point", and he continues to roll until either he rolls his point and wins or rolls a 7 and loses. Find his probability to win.

(11.) (*Blackjack.*) A Blackjack is a set of two cards, one of which is an Ace and the other a King, Queen or Jack.

(a.) Find the probability that two randomly drawn cards form a Blackjack.

(b.) Two hands of two cards are dealt. What is the probability that neither is a Blackjack?

(12.) (*Quinquenove.*) A and B play with two dice. A rolls first, and wins with a seven. Otherwise it is B's roll, who wins with a 5 or a 9. If B fails, the play reverts to A, and continues until someone wins. Find probability that A wins.

(13.) Let $P(A) = 0.25$ and $P(B) = 0.4$. Find the probability of "A or B", AB and "A but not B" if

(a.) A and B are mutually exclusive.

(b.) A and B are independent.

(14.) A cat, when its tail is pulled will either growl or purr, with a probability depending on the species of cat. For example, a lion will growl with probability 50%, a tiger with 80%, and a leopard with only 20%. A youth, on passing a cage containing 4 lions, 6 tigers and 10 leopards, and seeing a tail hanging from it, grasps the tail with both hands and pulls. He is rewarded with a low growl.

What is the probability he has a tiger by the tail?

(15.) Educational research indicates that 40% of the students in Calculus I received a grade of C, but that 70% of those are incompetent in the subject. The corresponding percentages for the other grades are;

Grade	**A**	**B**	**C**	**F**
% with grade	20%	30%	40%	10%
% incompetents	20%	50%	70%	100%

(a.) What is the probability that a student who has completed Calculus I is incompetent?

(b.) Jones is a known incompetent. What is the probability he got an A?

Section 3.4 Problems.

(16.) There are two species of Jubjub bird. 75% are *j. virginianus* which flies and the rest are *j. carolinus* which does not. Both may have either the orange or the blue plumage. *Virginius* is 50% blue, while *Carolinus* is 90% blue.

(a.) What is the probability that a randomly sighted Jubjub bird is blue?

(b.) You spot an orange Jubjub bird. What is the probability that it can fly?

(17.) Term papers in an English course with 200 students are graded by three TA's. Alan and Babs are half-time and each grades 50 papers, while Clyde is full-time and grades 100 papers. Alan will assign a failing grade to 10% of his papers, Babs to 20% of hers and Clyde to 40% of his.

(a.) What is the probability that a randomly chosen student will fail?

(b.) If a given student has failed, what is the probability that his paper was graded by Clyde?

(18.) A rare subspecies of moth conststuts only.0.1% of the species population. 98% of the the rare subspecies have a distinctive wing pattern, while only 5% of the common species have it. If a moth with the pattern is spotted, how likely is it to be of the rare subspecies?

(19.) A medical test for a certain disease will show positive for 99% of the people who have the condition, and positive for only 5% of those who do not. If only 0.1% of the population has the condition, what is the probability that a person who tests positive will actually have the disease?

(20.) For two rolls of a fair die, let S_1 be six on the first roll, S_2 six on the second roll and E be the event that the sum is even number. Prove that set S_1, S_3, E is pairwise independent but not independent.

(21.) Solve *Polya's urn model* (Example 2 of section 1) by combinatorial methods.

(22.) Let E and F be independent events. If E has probability p and F has probability q, what is the probability of the event "E or F" ?

(23.) There is a 70% chance that Professor K will be on leave next year, and a while 50% chance Professor L will be likewise. If their decisions are made independently, what is the probability that exactly one will be on leave?

(24.) A tax evader estimates that he has probability 0.2 of being caught by the Feds, and 0.3 of being caught by the State. If the two agencies work independently, what is the probability he gets away with it.

(25.) A student applies for a fellowship at two different universities. He estimates that the probability being awarded a fellowship to be $1/3$ by the first, $1/4$ by the second and $1/6$ by both.

(a.) What is the probability of receiving at least one offer?

(b.) If he receives an offer from the first, what is the probability he also receives one from the second?

(26.) An urn contains 3 red and 7 black balls. Players A and B alternately withdraw balls from the urn without replacement, until a red ball is selected. A draws the first ball, then B, and so on. Find the probability that A selects the red ball.

Chapter 4
Independent Trials.

4.1 Bernoulli Trials.

By a sequence of *Bernoulli trials* we shall mean a sequence of independent trials of an experiment $\mathcal{E}$, in which we focus only on the occurrence or non-occurrence of a single event S. We shall refer to the event S as "*Success"* and to its complementary event F, as "*Failure*". Thus, $p = P(S)$ is the probability of Success and $q = 1 - p = P(F)$ is the probability of failure.

As an example, think ot the toss of a coin having the probability p of Heads; or of the roll of a die, where *Success* is a Six, so that $p = 1/6$ and $q = 5/6$ for a fair die,

Cardano's Law.

In the case of the die, a natural first question to ask is "*What is the probability of a run of n straight sixes in n rolls?*"; that is, of n successive Successes in n trials?

Theorem 1. (Cardano's Law.) *The probability of n Successes in n trials is p^n.*

Proof. Let S_k be the event that there is a Success on the k^{th} trial. Then

$$P(n \text{ Successes }) = P(S_1 S_2 \cdots S_n).$$

But the events $S_1, S_2, \cdots, S_n$ are an independent set since they all correspond to different trials. Hence,

$$\begin{aligned} P(n \text{ Successes }) &= P(S_1 S_2 \cdots S_n) \\ &= P(S_1)P(S_2)\cdots P(S_n) \\ &= pp\cdots p = p^n. \blacksquare \end{aligned}$$

Example 1. Find the probability of rolling four straight sixes with a die.

Solution: By Cardano's Law, the probability is

$$1/6^4 = 1/1296. \blacksquare$$

Example 2. Find the probability of 10 straight Heads in 10 tosses of a fair coin.

Solution: By Cardano's Law, the probability is.

$$1/2^{10} = 1/1024.\blacksquare$$

Corollary 1. *The probability of at least one Success in n trials is*

$$1 - q^n$$

Proof. The probability of not obtaining at least one Success in n trials is the probability n successive Failures. By Cardano's Law, the probability of that is q^n, so the probability of at least one Success is $1 - q^n$.$\blacksquare$

Example 3. Find the probability of *at least one* six in four rolls of a die.

Solution: The probability of six straight failures to roll a six is

$$\left(\frac{5}{6}\right)^4$$

by Cardano's Law. The probability of at least one six is therefore.

$$1 - \left(\frac{5}{6}\right)^4 = 0.42.\blacksquare$$

Example 4. (*Cardano's Problem of Equality*.) Another interesting problem is the following. Suppose that you are playing a game which you have a probability p of winning. How many times must you play to have at least a 50% chance of winning at least once? Say, for example, that you have only one chance in 100 to win on any one play, so that $p = 1/100$. Since you will win on average about once in 100 plays, it may seem reasonable that you should have a 50% chance to winning at least once in 50 plays. This "*argument on the mean*" is plausible, but erroneous.

What is needed is that *the probability of n successive losses is less that or equal to* $1/2$. That is,

$$q^n \leq 1/2.$$

Equivalently,

$$n \log q \leq \log (1/2)$$

or

$$n \geq \log (2) / \log (1/q) .$$

Thus, if $p = 1/100$, we need

$$n \geq \log 2 / \log(100/99) = 68.97 \simeq 69 \text{ plays.}$$

If p is small, then because

$$\lim_{p\to 0} \frac{\log(1-p)}{p} = -1,$$

we have

$$-\log(1-p) \simeq p$$

for small p, so that the condition for equality is approximately,

$$n > \frac{\log 2}{p} \simeq \frac{0.69}{p}. \blacksquare$$

Remark. The Problem of Equality is what led Cardano to his formula. He first used the argument on the mean, realized it was wrong, and then found the correct answer. The book contains both the correct and incorrect solutions. It remained unpublished in Cardano's lifetime, and is probably a draft. See Ore: *Cardano, the Gambling Scholar*, which has a full translation of Cardano's book.

4.2 The Binomial Formula.

Consider next the probability of *exactly k Successes in n trials.*

Theorem 2. **(Binomial Formula.)** *The probability of exactly k Successes in n independent trials is*

$$b(k;n,p) = \binom{n}{k} p^k q^{n-k}.$$

Proof. Let S_k and F_k be be respectively Success and Failure on the k^{th} trial. One possible way to get exactly k Successes in n trials is to start off with k Successes followed by $n-k$ Failures. This is the compound event

$$S_1 \cdots S_k F_{k+1} \cdots F_n$$

which, as above, has probability

$$\begin{aligned} P(S_1 \cdots S_k F_{k+1} \cdots F_n) &= P(S_1)\cdots P(S_k)P(F_{k+1})\cdots P(F_n) \\ &= p \cdots pq \cdots q = p^k q^{n-k}. \end{aligned}$$

There are, however, many other ways to do get k Successes. The k Successes may come at any k positions in the n trials - the corresponding event is denoted by a distinct rearrangement of k letters S and $n-k$ letters F. There are

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

such rearrangements, so the probability is

$$\binom{n}{k} p^k q^{n-k}. \blacksquare$$

The Binomial formula is one of the most basic formulas of probability theory. Note that it gives the probability of *exactly* k Successes, not *at least* k Successes.

Example 1. Find the probability of *exactly* 2 sixes on 5 rolls.

Solution. The probability is

$$\binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3. \blacksquare$$

Example 2. Find the probability of *at least* 2 sixes on 5 rolls.

Solution. The probability is

$$1 - \binom{5}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^5 - \binom{5}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^4 = \frac{1526}{7776} = 0.196. \blacksquare$$

*Banach's Match Boxes.

Stefan Banach was a pioneer of the branch of mathematics known as Functional Analysis. He is the subject, rather than the originator, of the following famous problem.

Banach's Match Box Problem. A mathematician carries a matchbox in each pocket of his coat. When he needs a match, he randomly selects a pocket and withdraws a match from the corresponding box, which he returns to his pocket without observing whether or not it is empty. Eventually, of course, he will select a box that turns up empty.

If the boxes originally contained n matches, what is the probability that when he selects an empty box, the other box is also empty?

Solution. The desired event will occur if after $2n$ trials, he has selected the left pocket exactly n times; for then all matches are gone, but no box has been selected more than n times and thus found empty. Thus if S is the event that he that he selects the left pocket, we want the probability of exactly n Successes in $2n$ trials, where $p = 1/2$. This is

$$\binom{2n}{n} \left(\frac{1}{2}\right)^{2n}. \blacksquare$$

*Weldon's Dice Data.

A famous sequence of Bernoulli trials took place in 1894, when the zoologist *W. F. R. Weldon (1860 - 1906)* performed the experiment of rolling 12 dice at a time from a cup a total of $26,306$ times, the equivalent of $12 \times 26,306 = 315,672$ rolls of a single die. He

recorded the number of times that either a five or a six appeared. His data are given in the following table, where n_k is the number of times that k five or six appeared.

k	0	1	2	3	4	5	6	7	8	9	10	11	12
$\mathbf{n}_k$	185	1149	3265	5475	6114	5194	3067	1331	403	105	14	4	0

The total number of fives and sixes is $106,602$. If the dice were "fair", the number of fives or sixes should have a Binomial distribution with $p = 1/3$ and $n = 315,672$. However, the estimate for p here is

$$p = \frac{106,602}{315,672} = 0.33770.$$

This is close, but there are a lot of trials. Is it reasonable to assume fairness? This is a Hypothesis Testing question to which we shall return later?

4.3 The Multinomial Formula.

Consider next the case of a sequence of n independent trials in which we focus - not just on two events S and F - but on a set of several mutually exclusive and exhaustive events $S_1, .., S_r$. Thus, instead of *Success* S and *Failure* F, we have r different types of Success $S_1, .., S_r$, a sort of *Multiple Bernoulli* - or *"Multinoulli"* - trials.

We are interested in the probability of exactly n_1 occurrences of S_1, n_2 occurrences of S_2,..., and n_r occurrences of S_r, where necessarily $n = n_1 + n_2 + \cdots + n_r$.

To illustrate the argument, we shall first consider some examples.

Example 1. Suppose an urn contains 5 Red, 3 White and 2 Blue balls. If balls are drawn with replacement, we have a sequence of independent trials on which we are interested in the three events *Red*, *White* and *Blue*. If we draw 10 times what is the probability of getting exactly 5 Red, 3 White and 2 Blue balls?

Solution. Let $R, W,$and B be respectively the events "*Red" "White"* and *"Blue"* Let R_k be "*Red on the* k^{th} *trial*", etc. Proceeding as with the Binomial case, one way to win is to roll

$$R_1R_2R_3R_4R_5W_6W_7W_8B_9B_{10}.$$

This has probability

$$\begin{aligned} & P\left(R_1R_2R_3R_4R_5W_6W_7W_8B_9B_{10}\right) \\ = \; & P(R_1)P(R_2)P(R_3)P(R_4)P(R_5)P(W_6)P(W_7)P(W_8)P(B_9)P(B_{10}) \\ = \; & \left(\frac{5}{10}\right)^5\left(\frac{3}{10}\right)^3\left(\frac{2}{10}\right)^2. \end{aligned}$$

The number of ways to win is the number of distinct permutations of the word

$$RRRRRWWWBB,$$

which is

$$\binom{10}{5\ 3\ 2} = \frac{10!}{5!3!2!} = 2520.$$

The desired probability is therefore

$$\begin{aligned} & \binom{10}{5\ 3\ 2} P(R)^5 P(W)^3 P(B)^2 \\ = \quad & \frac{10!}{5!3!2!} \left(\frac{5}{10}\right)^5 \left(\frac{3}{10}\right)^3 \left(\frac{2}{10}\right)^2 = 0.08505. \blacksquare \end{aligned}$$

Example 2. Consider the roll of *six* dice. What is the probability of rolling *exactly* 2 twos and 3 fours? We will call this event a *Win*.

Solution. We have six independent trials of rolling a die. The three events of interest are "*Two" "Four"* and *"None of the above* which, we denote by T, F and N. We want two T's, three F's and one N. Let T_k be "*Two on the first trial*", etc. Proceeding as with the Binomial case, one way to win is to roll

$$T_1 T_2 F_3 F_4 F_5 N_6.$$

This has probability

$$\begin{aligned} P\left(T_1 T_2 F_3 F_4 F_5 N_6\right) &= P(T_1) P(T_2) P(F_3) P(F_4) P(F_5) P(N_6) \\ &= P(T_1)^2 P(F_3)^3 P(N_6)^1 = \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^3 \left(\frac{4}{6}\right)^1. \end{aligned}$$

The number of ways to win is the number of distinct permutations of $TTFFFN$, which is

$$\binom{6}{2\ 3\ 1} = \frac{6!}{2!3!1!} = 60.$$

Thus,

$$\begin{aligned} P(\text{Win}) &= \binom{6}{2\ 3\ 1} P(T_1)^2 P(F_3)^3 P(N_6)^1 \\ &= \frac{6!}{2!3!1!} \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^3 \left(\frac{4}{6}\right)^1 = \frac{5}{972} = 0.0005. \blacksquare \end{aligned}$$

In general, consider events a set of events $E_1, .., E_r$ which are *mutually exclusive and exhaustive*, with probabilities $P(E_k) = p_k$ satisfying

$$p_1 + \cdots + p_r = 1.$$

By the same argument, we obtain

Theorem 3. (The Multinomial formula.) *The probability of obtaining exactly n_1 occurrences of E_1, n_2 occurrences of E_2,..., and n_r occurrences of E_r in $n = n_1 + n_2 + \cdots + n_r$ trials of E is*

$$\binom{n}{n_1 \ \cdots \ n_r} p_1^{n_1} \cdots p_r^{n_r}$$

Example 3. A die with 3 Red, 2 White and 1 Blue side is rolled six times. What is the probability of getting 3 Red, 2 White and 1 Blue?

Solution: The probability is

$$\begin{aligned} & \binom{6}{3\,2\,1} \left(\frac{1}{2}\right)^3 \left(\frac{1}{3}\right)^2 \left(\frac{1}{6}\right)^1 \\ = \quad & \frac{6!}{3!2!1!} \frac{1}{2^3 3^2 6} = \frac{5}{36}. \blacksquare \end{aligned}$$

Example 4. What is the probability that a sequence of 10 random digits $(0, 1, ..., 9)$ contains exactly 1 one, 2 twos, 3 threes and 4 fours?

Solution: The probability is

$$\begin{aligned} & \binom{10}{1\,2\,3\,4} \left(\frac{1}{10}\right)^1 \left(\frac{1}{10}\right)^2 \left(\frac{1}{10}\right)^3 \left(\frac{1}{10}\right)^4 \\ = \quad & \frac{10!}{1!2!3!4!} \frac{1}{10^{10}} = 1.26 \times 10^{-6} = 0.00000126. \blacksquare \end{aligned}$$

4.4 Waiting Times.

The Wait for the First Success.

How long does one have to wait for the first Success? Let T be the number of trials it takes; i.e. $T = n$ iff the first S occurs one the n^{th} trial. Then

$$\begin{aligned} P(T = n) &= P(F_1 \cdots F_{n-1} S_n) \\ &= P(F_1) \cdots P(F_{n-1}) P(S_n) = q \cdots qp = q^{n-1} p. \end{aligned}$$

We therefore have:

Theorem 4. (Geometric Formula.) *The probability that the first Success occurs on the n^{th} trial is*

$$q^{n-1} p.$$

Note also that

$$P(T > m) = P(F_1 \cdots F_m) = q^m.$$

Example 1 In the roll of a die, the probability that the *first six occurs on the third roll* is

$$\left(\frac{5}{6}\right)^2 \frac{1}{6} = 0.116.$$

The Wait for the r^{th}Success.

The wait for the second or third, or in general, the r^{th} Success can be treated in the same way.

Theorem 5. (Pascal's Formula.) *If T_r is the wait for the r^{th} Success, then*

$$P(T_r = n) = \binom{n-1}{r-1} p^r q^{n-r}.$$

Proof.

$$\begin{aligned} P(T_r = n) &= P(\text{Exactly } r-1 \text{ S's in the first } n-1 \text{ rolls \& S on the } n^{th} \text{ roll}) \\ &= P(\text{Exactly } r-1 \text{ S's in the first } n-1 \text{ rolls})\, P(\text{ S on the } n^{th} \text{ roll}) \\ &= b(r-1, n-1; p) P(S_n) = \left[\binom{n-1}{r-1} p^{r-1} q^{n-r}\right] p \\ &= \binom{n-1}{r-1} p^r q^{n-r}. \square \end{aligned}$$

Example 2. In the roll of a die, the probability that third six occurs on the 10^{th} roll is

$$\binom{9}{2} \left(\frac{5}{6}\right)^7 \left(\frac{1}{6}\right)^3. \blacksquare$$

*The Problem of Points.

Although a few calculations of probabilities are known from an earlier date - notably by Cardano and Galileo - the main activity in the subject dates from a correspondence in 1654 between Pascal and Fermat concerning the "*Problem of Points*".

In essence, the problem is this. Two gamblers, A and B, are playing a series of games in each of which A has probability p to win and B has probability $q = 1 - p$ to win. The wager, however, is not on the individual games, but on whether A can win a certain number

of games before B wins a certain - generally different - number. Each player contributes a certain amount of money to the stakes, and the winner takes the whole amount.

The problem arises if the players decide to stop playing after a certain number of games have been played, but before either player has won enough games to claim the stakes. The question then is: "*How are the stakes to be divided*?"

The key to the solution is to realize that the fairest way is to divide them is *in proportion to the probabilites that each of the two players would win if the game were to continue.* On any other assumption, it would benefit one or the other of the players to drop out without finishing.

Example 1. As an example, suppose that A undertakes to roll a *Six* three times in 15 rolls of a single die. He has rolled a *Six* twice and has three rolls remaining when play ends. Since A needs to win once more in his last three rolls, his probability to win is

$$1 - \left(\frac{5}{6}\right)^3 = 0.42.$$

He should therefore receive 42% of the stakes and his opponent B should receive 58%. ■

This leads to the following general problem:

Problem of Points. *Suppose that A has probability p to win a single game and must win a games to win the stake, while B has probability $q = 1 - p$ to win a single game, and must win b games to win the stake. Find the probability that A will win the stake.*

Solution. The solution follows easily from Pascal's formula. If A wins, he will win in n games, where $a \leq n < b + a - 1$. He will win in n games if his a^{th} Success occurs on the n^{th} game. The probability that this occurs is, by Pascal's formula,

$$P(A \text{ wins in } n \text{ games}) = \binom{n-1}{n-a} q^{n-a} p^a.$$

The probability that A wins is this quantity summed over all possible values of n :

$$P(A \text{ wins}) = \sum_{n=a}^{b+a-1} P(A \text{ wins in } n \text{ games}) = \sum_{n=a}^{b+a-1} \binom{n-1}{n-a} q^{n-a} p^a. \blacksquare$$

Example 2. (*Pacioli 1494.*) Two players A and B, with equal chances to win a single game, play for 6 wins, with equal chances to win, but stop with the score at 5 games to 3. What is the probability that A will win if the game continue? How should the stakes be divided if the game is terminated at this point?

Solution. We have $a = 1, b = 3$ and $p = q = 1/2$. Thus

$$P(A \text{ wins}) = \sum_{n=1}^{3} \binom{n-1}{n-1} \left(\frac{1}{2}\right)^n = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}.$$

The stakes should therefore be divided in the ratio $7:1$. ■

Pacioli treated this problem incorrectly as a question of proportions and gave the answer as $5:3$. Pacioli's method can be used to advantage by an astute player.For instance, it is clearly to B's advantage to quit at this point if the stakes are divided in the ratio $5:3$. (See Problem 6 of Chapter 5.)

*Playoff Series.

In some sports - for example, in Baseball's World Series or the NBA playoffs - the two competing teams play a series of games. The theory is (aside from the consideration of increased revenue) that this improves the chances of the superior team winning the championship. To gauge to what extent this is true, consider a 7 game series in which team A has probability $p = 0.6$ to win any single game. What is $A's$ probability to win the series? Treating the games as independent trials of the same experiment, we have from the formula above,

$$\begin{aligned} P(A \text{ Wins}) &= \sum_{n=4}^{7} \binom{n-1}{3} p^4 q^{n-4} = p^4 \left[1 + \binom{4}{3} q + \binom{5}{3} q^2 + \binom{6}{3} q^3\right] \\ &= p^4 \left[1 + 4q + 10q^2 + 20q^3\right] = (0.6)^4 (5.48) = 0.71. \end{aligned}$$

Thus, *on these assumptions*, playing a series of 7 games improves the superior team's chances of winning from 60% to 71%. ■

We stress the phrase "*on these assumptions*". We have assumed in this calculation that team A has a fixed probability to win each game and that the outcome of any game is independent of the outcome of any others. This neglects, for example, the alleged "home court advantage". The reader may wish to consider the reasonableness of these assumptions in practice.

The point here is that in any application of probability - any probability model of a real situation - certain assumptions are made. The success of the model will depend on how closely these assumptions are satisfied in reality. *It is therefore always important to know what assumptions are being made.* Then, if the model does not work , it means that some of the assumptions fail and must be changed. If you are unsure of your assumptions, you will not be in a position to suggest appropriate changes.

4.5 Problems.

(1.) What is the probability of rolling Doubles three successive times with two dice?

(2.) A die is rolled 4 times. Find the probability of at least one six.

Section 4.5 Problems.

(3.) If the chances of at least one success in 30 trials are 50%, in how many trials are the chances 75%?

(4.) If the chances of at least one success in 30 trials are 50%, what are the chances in 60 trials?

(5.) Find a formula for minimum number of trials to make the probability of at least one Success equal to p_0.

(6.) (*Huygens. 1657.*) (a.) In how many throws is it safe to wager that one can throw at least one six with one die?

(b.) A double six with three dice? ('double six' means '*exactly* two sixes').

(c.) With how many dice may one safely wager to throw at least one six on the first roll?

(7.) (*Problem of the Chevalier de Méré.*) In 1654, Pascal wrote to Fermat, "The Chevalier de Méré said to me that he has found falsehood in the theory of numbers for the following reason. If I undertake to throw a six with one die, there is an *advantage* in undertaking to do it in 4 throws. If I undertake to throw *two* sixes with *two* dice, there is a *disadvantage* in undertaking to do it in 24 throws. But 24 is to 36 as 4 is to 6. This is his '*grande scandale*' which makes him say loftily that the propositions are not constant and that arithmetic is self-contradictory."

Resolve the difficulty by computing the probabilities of

(a.) throwing at least one six in four throws of a single die, and

(b.) throwing at least one double six in 24 throws of two dice.

Compare the results of (a.) and (b.). The difference was presumably noted at the tables.

(8.) Show that for $0 < p < 1/2$, the number obtained by "*reasoning on the mean*" is always too low; i.e. show that

$$\frac{\log 2}{\log(1/q)} > \frac{1}{2p}.$$

where $q = 1 - p$.

(9.) Two fair dice are thrown n times in succession. Compute the probability that a double 6 appears at least once. How large need n be to make this probability at least $\frac{1}{2}$.

(10.) A roulette wheel has the numbers 0 through 36 and a double 0 (38 numbers in all). A player bets on the numbers 1 through 12. What is the probability that he loses five consecutive times?

(11.) Five dice are rolled. Given that there are exactly 2 sixes what is the probability that there are also exactly 2 fives?

(12.) Show that the probability of an *even number* of successes in n Bernoulli trials is

$$\frac{1}{2}[1 + (q - p)^n].$$

(13.) Two people toss a fair coin n times each. Find the probability that they will score the same number of heads.

(14.) A fair coin is flipped $2n$ times. Find the probability P_n that exactly n heads and n tails are obtained. Find an asymptotic formula for P_n and show that $\lim P_n = 0$.

(15.) A workman attends 12 machines of the same type. The probability that any given machine will require his attention during a period of an hour is $1/3$. What is the probability that exactly four machines will require his attention in an hour?

(16.) Only 1 bureaucrat in 5 is competent. Assume that bureaucrats are placed in offices independently of one another. Find the probability that in an office with 5 bureaucrats:

(a.) all are incompetent.

(b.) exactly 3 are incompetent.

(c.) In an office of 15, at least two are competent.

(d.) For success of a project, it is necessary that there be at least one competent bureaucrat in each of two independent offices, of 5 and 6 people respectively. What is the probability of success?

(17.) Smith undertakes to throw three even numbers in five throws of a single die. Since there are the same number of evens and odds on a die, Jones proposes odds of *3:2*. Would you play with him?

(18.) Solve Banach's match box problem if there are initially a different number of matches in each box.

(19.) Solve Banach's match box problem if the left pocket is favored.

(20.) Six dice are painted with three sides Red, two sides White and one side Blue. When all six are rolled what is the probability of obtaining three Reds, two Whites, and one Blue?

(21.) Three fair coins are tossed together 10 times. What is the probability of obtaining three heads twice, three tails twice, one head four times, and two heads twice?

(22.) According to Barnum, "*There's a sucker born every minute and two to take him.*" If five people are born every minute, what is the probability that a group of six people contains three suckers but no one to take them?

(23.) An *astragalus* is a bone from the ankle of a sheep. When rolled, an astragalus can come to rest on any one of four sides, which are numbered as 1, 3, 4 and 6. Sides 1 and 6 are smaller, and so occur less often than 3 and 4. Experimentally, the probabilities of landing on each side are as follows

Side	1	3	4	6
Probability	0.1	0.4	0.4	0.1

Astragali were used by the Romans like dice for gambling and divination. Compute the probability of the following throws with four or five Astragali, as indicated.

(a.) Throw of Venus: 1-3-4-6.

(b.) The dogs: 1-1-1-1.

(c.) Throw of the Savior Zeus: 1-3-3-4-4.

(d.) Throw of the Good Chronos: 6-3-3-3-3.

(e.) Throw of Child-eating Chronos: 4-4-4-6-6.

(f.) Throw of Poisidon: 6-4-4-4-4.

The Venus throw was considered the most favorable, although it was not the least likely, perhaps indicating a certain optimism on the part of the ancients.

(24.) A lake contains 30% bass, 20% catfish, and 50% perch. What is the probability that a random catch of nine fish contains two bass, three catfish, and four perch?

(25.) A target has three concentric regions: the bull's-eye, the inner ring and the outer ring. When a certain archer shoots arrows at the target, his probabilities of hitting each region are as follows: *Bull's-eye* $1/9$; *Inner Ring* $3/9$; *Outer ring* $5/9$.

Assuming that his shots are independent, find the following:

(a.) The probability of exactly 2 bull's-eyes in 9 shots.

(b.) The probability of at least 2 bull's-eyes in 9 shots.

(c.) The probability that the first bull's-eye occurs on the 5^{th} shot.

(d.) The probability that in 9 shots, he hits exactly 1 bull's-eye, 3 inner rings, and 5 outer rings.

(26.) An urn contains four white and four black balls. Four balls are drawn. If two are white and two black, we stop; if not, the balls are replaced and the experiment repeated. What is the probability that this takes no more than three draws?

(27.) There are three urns, each containing two balls. An urn is selected at random and a ball removed. The process is repeated until eventually an empty urn is selected. What is the probability that when this happens, the other two urns contain the same number of balls?

(28.) Three players toss three coins. If the coins are different then odd man wins; otherwise, no decision is reached.

(a.) Find the probability that no decision is reached after 10 throws.

(b.) If all the coins are *identical, but not fair*, does the probability of reaching a decision increase, decrease, or remain the same?

(29.) (*A Traffic problem.*) A pedestrian is waiting to cross the street. The probability of a car passing in a given second is p, and the pedestrian can cross in a gap of 3 seconds. Assume that the events of a car appearing is a given second are independent. If T is the wait for a 3 second gap, find $P(T = k)$ for $k = 0, 1, 2, 3, 4$.

(30.) (*Problem of Points with 3 players.*) Suppose A has probability a to win a single game, and needs α games to win the stakes, and let b, β and c, γ be the corresponding numbers for B and C, so that $a + b + c = 1$. Find A's probability to win the stakes.

(31.) Let E and F be independent events. If E has probability p and F has probability q, what is the probability of the event "E or F" ?

(32.) A tax evader estimates that he has probability 0.2 of being caught by the Feds, and 0.3 of being caught by the State. If the two agencies work **independently**, what is the probability he gets away with it.

(33.) A student applies for a fellowship at two different universities. He estimates that the probability being awarded a fellowship to be $1/3$ by the first, $1/4$ by the second and $1/6$ by both.

(a.) What is the probability of receiving at least one offer?

(b.) If he receives an offer from the first, what is the probability he also receives one from the second?

(34.) A fair coin is tossed 100 times. Use Stirling's approximation to approximate the probability of 50 Heads.

Chapter 5
Discrete Random Variables.

5.1 Random Variables.

When discussing sequences of independent trials in the preceding chapter, we encountered the number N of Successes in n independent trials and the wait T for the first Success. These quantities are *numbers whose values depend on chance*; that is, they *vary randomly* with each trial of the corresponding experiment.

A number whose value depends on the outcome of a random experiment $\mathcal{E}$ is called a *random variable* (r.v.).

Example 1. Some examples of r.v. are the following:

1. The *number of heads in* 3 *tosses of a coin*.

2. More generally, the *number of Successes in* n *independent trials* of an experiment.

3. The *wait for the first Success at Bernoulli trials.*

4. The *number appearing uppermost* when a die is rolled.

5. The *sum of the of numbers on two dice.*

6. A *gambler's gain* - that is, the amount won or lost - in the play of a game.

7. The *wait for the first call to a switchboard*.

8. The *height of an inductee* into the French army.

9. The *chest measurement* of a Scottish soldier.

10. The *annual number of deaths from mule kicks* in a Prussian army corps.

The last three examples were the subject of famous statistical studies by the 19^{th} century statisticians *Quetelet* and *von Bortkewitz*.

The first thing to ask about a r.v. is "*What are the possible values that the r.v. can assume*?" The set of possible values that the r.v. X can assume is referred to as the *range* of X.

For example:

1. The range the number of heads in 3 tosses is the set $\{0, 1, 2, 3\}$,

2. The range of the wait for first Success is the set of positive integers $\{1, 2, 3, \ldots\}$, and

3. The range of the wait for the first call to a switchboard is the set of non-negative reals $[0, \infty)$.

A random variables may be classified as *discrete* or *continuous.* The range of a *discrete random variable* is a finite or countably infinite set (such as the integers). The range of a *continuous random variable* consists of all values in some interval (or intervals) of the real numbers.

Thus, the number of heads in 3 tosses of a coin is a *finite discrete r.v.*; the wait for first Success is an *infinite discrete r.v.*, while the wait for first call is a *continuous r.v.*

There are naturally occurring r.v. that are neither discrete nor continuous, but rather a mixture of the two. One such r.v. is discussed in section 7.7.

We shall begin our discussion in this chapter with discrete r.v. Most of the basic concepts will already appear in this context. The transition to continuous r.v. later chapters is mostly a technical matter of replacing finite sums by integrals

Random Variables and Sample Spaces.

As we have noted, a r.v. is a number whose value is determined by the outcome of a random experiment $\mathcal{E}$. Since the points of the sample space of $\mathcal{E}$ correspond to the possible outcomes of $\mathcal{E}$, it follows that, in the language of sample spaces, the r.v. is a *function on the sample space* Ω.

5.2 Distribution of a Random Variable.

How do we describe a random variable X from the point of view of probability theory? As far as probability theory goes, all we can ask about a *single* event E is *"What is its probability?"* It follows that for a r.v. X, all we can ask is "*What are the probabilities that, on a given trial, the r.v. will assume the various values of its range*?". This information is known as the *distribution* of the r.v.

Let X be a finite discrete r.v. with values $x_1, x_2, \ldots, x_n$. The *distribution* of X is determined by giving the probabilities

$$p_i = P(X = x_i) \qquad i = 1, \ldots, n$$

that X takes on its various values. Now since X will assume *exactly one* of these values, the events $\{X = x_i\}$ ($i = 1, \ldots, n$) are mutually exclusive and exhaustive, so we must have

$$\sum_{i=1}^{n} P\{X = x_i\} = \sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots + p_n = 1.$$

Example 1. Let X be the number of Heads in three tosses of a fair coin. The distribution of X is given by the following table.

$\mathbf{x}_i$	0	1	2	3
$\mathbf{p}_i$	1/8	3/8	3/8	1/8

.

This distribution can be represented graphically as shown in Figure 2.1.

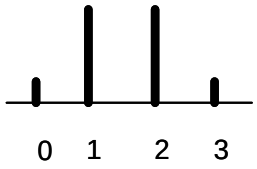

Figure 2.1. *Binomial Distribution* $Bin(1/2,3)$**.** ■

Example 2. If X is the number uppermost on a fair die, then its distribution is given by

$\mathbf{x}_i$	1	2	3	4	5	6
$\mathbf{p}_i$	1/6	1/6	1/6	1/6	1/6	1/6

.■

Example 3. If 2 balls are drawn from an urn containing 3 Red and 2 Black balls, the number of R Reds drawn has distribution

$\mathbf{r}_i$	0	1	2
$\mathbf{p}_i$	1/10	6/10	3/10

.■

Examples 1, 2 and 3 are special cases of the following general distributions.

Example 4. *Binomial distribution.* The number X of successes in n Bernoulli trials has the *Binomial distribution*

$$P(X=k)=\binom{n}{k}p^k q^{n-k} \qquad k=0,1,\ldots,n.$$

This is a very common and important distribution. As a convenient shorthand, we will refer to it as $Bin(n,p)$. The r.v. X in Example 1 has the distribution $Bin(\frac{1}{2},3)$.■

Example 5. *Discrete uniform*. If all values of a r.v. are equally likely then, $p_i = 1/n$ for $i = 1,\ldots,n$. The r.v. X in Example 2 is uniformly distributed on the set $\{1,2,3,4,5,6\}$.■

Example 6. *The Hypergeometric distribution.* If n balls are drawn *without replacement* from an urn containing r Red and b Black balls, then the number X of Red balls drawn has the distribution

$$P(X=k)=\frac{\binom{r}{k}\binom{b}{n-k}}{\binom{r+b}{n}} \qquad k=1,\ldots,n.$$■

The r.v. R in Example 3 has the hypergeometric distribution with $n=2, r=3$ and $b=2$.

5.3 Joint Distributions.

Just as the probability of an event tells us all that we can know about the occurrence of the event, so the distribution of a r.v. tells us all we can know about the r.v., at least from the point of view of probability theory. In the case of two or more events, however, there is a question not only of how frequently the events occur *separately*, but of how frequently they occur *together*.

Similarly, we may ask how frequently the various values of two r.v. X and Y occur together on the same trial. This information is known as the *joint distribution* of X and Y. More precisely, we have

Definition 1. The *joint distribution* of two discrete r.v. X and Y is given by

$$p_{ij} = P(X = x_i \,\&\, Y = y_j) \qquad i = 1, \ldots, n \; j = 1, \ldots, m.$$

Example 1. For a very simple example, let two balls be drawn *without replacement* from an urn with 2 Red and 3 Black balls. Let X be the number of Reds drawn on the first draw and Y the number of Reds drawn on the second draw. Thus, both X and Y take on only the values 0 and 1.

Clearly, it is somewhat unlikely that one will draw two Reds; that is, it is unlikely that the r.v. X and Y will both equal 1. This is reflected in the joint distribution, which we shall now compute. We have

$$p_{11} = P(X = 1 \,\&\, Y = 1) = \frac{2}{5} \cdot \frac{1}{4} = 0.1$$

and similarly

$$p_{00} = p_{01} = p_{10} = 0.3.$$

The joint distribution can therefore be expressed in a table as follows:

	$\mathbf{X = 0}$	$\mathbf{X = 1}$
$\mathbf{Y = 0}$	0.3	0.3
$\mathbf{Y = 1}$	0.3	0.1

■

In general, one can have the joint distribution of any number of r.v. $X_1, X_2, ..., X_n$

$$p(x_1, x_2, ..., x_n) = P(X_1 = x_1 \,\&\, X_2 = x_2 \,\&....\&\, X_n = x_n).$$

Example 2. Consider an urn containing 5 Red, 3 White and 2 Blue balls. If 5 balls are drawn *without replacement*, let R be the number of Red balls drawn, W the number of

white and B the number of Blue. The joint distribution of R, W and B is then given by the *Generalized Hypergeometric formula* of section2.6 as

$$P(R = r \;\&\; W = w \;\&\; B = b) = \frac{\binom{5}{r}\binom{3}{w}\binom{2}{b}}{\binom{10}{5}}.\blacksquare$$

Example 3. If the experiment of Example 2 is performed *with replacement*, the joint distribution is given by the *Multinomial formula* as

$$P(R = r \;\&\; W = w \;\&\; B = b) = \binom{5}{r \quad w \quad b} \left(\frac{5}{10}\right)^r \left(\frac{3}{10}\right)^w \left(\frac{2}{10}\right)^b .\blacksquare$$

We see, therefore, that we may refer to the *Multinomial* and *Generalized Hypergeometric distributions*, since these two formulas are actually joint distributions of appropriate r.v.

Marginal Distributions.

The distribution of X can be obtained from the joint distribution by summing over all values of Y.

$$P(X = x_i) = \sum_{j=1}^{m} P(X = x_i \;\&\; Y = y_j) = \sum_{j=1}^{m} p_{ij}.$$

Example 4. In Example 1 above, we see that - by the Law of Total Probability - the distribution of X is

$$\begin{aligned} P(X &= 0) = P(X = 0 \;\&\; Y = 0) + P(X = 0 \;\&\; Y = 1) = 0.3 + 0.3 = 0.6 \\ P(X &= 1) = P(X = 1 \;\&\; Y = 0) + P(X = 1 \;\&\; Y = 1) = 0.3 + 0.1 = 0.4 \end{aligned}$$

This corresponds to summing the columns of the table and writing the result in the *lower margin*.

	$\mathbf{X = 0}$	$\mathbf{X = 1}$
$\mathbf{Y = 0}$	0.3	0.3
$\mathbf{Y = 1}$	0.3	0.1
	0.6	0.4

Similarly, the distribution of Y is found by summing over all X values, only this time one sums over rows and writes the values in the *side margin*.

	$\mathbf{X = 0}$	$\mathbf{X = 1}$	
$\mathbf{Y = 0}$	0.3	0.3	0.6
$\mathbf{Y = 1}$	0.3	0.1	0.4

For this reason, the distributions of X and Y are called *marginal distributions*. (The term has nothing to do with the use of the term '*marginal*' in Economics.) ■

Independent Random Variables.

Definition 2. Two discrete r.v. X and Y are *independent* iff the events $\{X = x_i\}$ and $\{Y = y_j\}$ are independent events for all values x_i and y_j. This is the case iff

$$P(X = x_i \,\&\, Y = y_j) = P(X = x_i)P(Y = y_j)$$

that is, *the joint distribution is the product of the marginal distributions*. In this way, one can check for independence of two r.v. from their joint distribution.

Example 5. In the example above , X and Y are not independent, since e.g.

$$0.1 = P(X = 1 \,\&\, Y = 1) \neq P(X = 1)P(Y = 1) = (0.4) \cdot (0.4) = 0.16$$

Actually, of course, it is obvious the X and Y are not independent, since it is clear that drawing a Red on the first draw will influence the probability of drawing one on the second. In most applications, it is not necessary to resort to this procedure; the r.v. will either be clearly independent, or - more likely - assumed to be independent. ■

More generally, we have

Definition 3. A set $X_1, X_2, ..., X_n$ of r.v. is an *independent set* iff..

$$P(X_1 = x_1 \,\&\, X_2 = x_2 \,\&....\&\, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n)$$

for all values of $x_1, x_2, ..., x_n$.

5.4 Expectation.

The *mean* or *expectation* of a r.v. is intuitively the average (in the sense of arithmetic mean) of the values of over a large number of trials. We run a large number n of trials, measure X on each trial, add up the values and divide by n.

Let n trials of $\mathcal{E}$ be run and assume that the value x_i of X is assumed n_i times, so that $n_1 + \cdots + n_r = n$. The sum of all values assumed by X is

$$\sum_{i=1}^{r} x_i n_i$$

and so the mean is

$$\frac{1}{n}\sum_{i=1}^{r} x_i n_i = \sum_{i=1}^{r} x_i \left(\frac{n_i}{n}\right) \simeq \sum_{i=1}^{r} x_i p_i$$

where $p_i = P(X = x_i)$. If n is large , we therefore expect that the average will be given by the sum

$$\sum_{i=1}^{r} x_i p_i.$$

of the values of X, *weighted according to the probabilities that they occur*. This quantity is called the *expectation* of X.

Definition 4. *The mean or expectation $E(X)$ of a discrete r.v. X is defined to be.*

$$E(X) = \sum_{i=1}^{r} x_i p_i.$$

Example 1. The number X of heads in three tosses has the distribution

$\mathbf{x}_i$	0	1	2	3
$\mathbf{p}_i$	1/8	3/8	3/8	1/8

Thus

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}.\blacksquare$$

Example 2. The value appearing on a fair die has the distribution

$\mathbf{x}_i$	1	2	3	4	5	6
$\mathbf{p}_i$	1/6	1/6	1/6	1/6	1/6	1/6

and hence the expectation

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3\frac{1}{2}.\blacksquare$$

Important Remarks.

(1.) These examples illustrate that *the mean need not be an actual possible value X*. In particular, *the mean need not be the most likely value of X*. For this reason, the term "*expected value*", which is sometimes used for the mean, is misleading. We will not use this term

(2.) It is also *not true that X has the same probability to be greater than $E(X)$ as it does to be less than $E(X)$*. A simple example is the following:

Example 3. Let N be the number of Heads in three tosses of a fair coin, and $X = N^2$. Then X has the distribution

$\mathbf{x}_i$	0	1	4
$\mathbf{p}_i$	1/4	1/2	1/4

The mean is

$$\mu = E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = 1.5$$

but

$$P(X > \mu) = 1/4 \text{ and } P(X < \mu) = 3/4.\blacksquare$$

Properties of Expectation.

Theorem 1. (**Properties of Expectation.)**

(a.) If $X \geq 0$, then $E(X) \geq 0$.

(b.) $E(1) = 1$.

(c.) $E(X + Y) = E(X) + E(Y)$

(d.) *If X and Y are independent, then* $E(XY) = E(X)E(Y)$.

Proof. Parts (a.) and (b.) are clear. For (c.), note that the values of $X + Y$ are $(x_i + y_j)$, which have the probabilities $P(X = x_i \ \& \ Y = y_j)$. Thus

$$\begin{aligned} E(X+Y) &= \sum_i \sum_j (x_i + y_j)\, P(X = x_i \ \& \ Y = y_j) \\ &= \sum_i x_i \left(\sum_j P(X = x_i \ \& \ Y = y_j) \right) + \sum_j y_j \left(\sum_i P(X = x_i \ \& \ Y = y_j) \right) \\ &= \sum_i x_i P(X = x_i) + \sum_j y_j P(Y = y_j) = E(X) + E(Y). \end{aligned}$$

(d.) The values of XY are $(x_i y_j)$, which have the probabilities

$$P(X = x_i \ \& \ Y = y_j) = P(X = x_i) P(Y = y_j).$$

Thus

$$\begin{aligned} E(XY) &= \sum_i \sum_j (x_i y_j)\, P(X = x_i \ \& \ Y = y_j) = \sum_i \sum_j (x_i y_j)\, P(X = x_i) P(Y = y_j) \\ &= \left(\sum_i x_i P(X = x_i) \right) \left(\sum_j y_j P(Y = y_j) \right) = E(X)E(Y). \square \end{aligned}$$

Note that *(d.) is definitely not true in general*. For example, it is emphatically *not true that* $E(X^2) = E(X)^2$.

Moments.

The number $\mu_p = E(X^p)$ is called the p^{th} *moment of* X. The following theorem is useful in computing moments.

Theorem 2. (The Law of the Unconscious Statistician.) *If $\phi(X)$ is a function of the r.v. X, then*

$$E(\phi(X)) = \sum_i \phi(x_i)\, p_i.$$

Proof. Let $S(y) = \{x_i : \phi(x_i) = y\}$ Then

$$P(\phi(X) = y) = \sum_{x_i \in S(y)} P(X = x_i) = \sum_{x_i \in S(y)} p_i$$

Thus

$$\begin{aligned} E(\phi(X)) &= \sum_{y \,\in\, ran(\phi(X))} yP(\phi(X) = y) = \sum_{y \,\in\, ran(\phi(X))} y \left(\sum_{x_i \in S(y)} p_i \right) \\ &= \sum_{y \,\in\, ran(\phi(X))} \left(\sum_{x_i \in S(y)} yp_i \right) = \sum_{y \,\in\, ran(\phi(X))} \left(\sum_{x_i \in S(y)} \phi(x_i) yp_i \right) \\ &= \sum_i \phi(x_i) p_i . \square \end{aligned}$$

Example 5. If X is the number of Heads in three tosses, then, using the distribution of X obtained in Example 1,

$$\mu_4 = E(X^2) = 0^4 \cdot 1/8 + 1^4 \cdot 3/8 + 2^4 \cdot 3/8 + 3^4 \cdot 1/8 = 132. \blacksquare$$

The Schwarz Inequality.

Theorem 3. (Schwarz Inequality.)

$$|E(X,Y)| \le \sqrt{E(X^2)}\sqrt{E(Y^2)}$$

Proof. Expanding, we have, for every t,

$$0 \le E\left(X + tY, X + tY\right) = t^2 E(Y^2) + 2tE(XY) + E(X^2) = p(t).$$

This says that the quadratic polynomial $p(t)$ has either no real roots or only one. That means the the discriminant

$$B^2 - 4AC = 4(X,Y)^2 - 4(Y,Y)(X,X) \le 0.$$

which gives the desired result.$\square$

Corollary 1. *If*

$$|E(X,Y)|^2 = E(X^2)E(Y^2),$$

then X and Y are linearly related, that is

$$aX + bY = 0$$

for some constants a and b.

Proof. Equality holds iff the discriminant is zero, in which case $p(t)$ has a real root c, so that

$$p(c) = E(X + cY)^2 = 0.$$

Since $(X + cY)^2 \geq 0$, this implies that $X + cY = 0$. □

5.5 Variance.

The mean is an estimate of the center of a distribution - what the statisticians call a *"measure of central tendency."*. As such it summarizes one aspect of a distribution. Suppose now that we want to measure of the ***fluctuations*** of a r.v., or in other words, the spread of its distribution about the mean. One measure of fluctuation is the variance.

Definition 5. The *variance* of a r.v. X is defined by

$$var(X) = \sigma^2 = E(X - \mu)^2.$$

where $\mu = E(X)$.

Example 1. The number X of heads in three tosses has the distribution

$\mathbf{x}_i$	0	1	2	3
$\mathbf{p}_i$	$1/8$	$3/8$	$3/8$	$1/8$

We have already computed that $\mu = 3/2$. Thus, from the definition,

$$var(X) = \left(0 - \frac{3}{2}\right)^2 \cdot \frac{1}{8} + \left(1 - \frac{3}{2}\right)^2 \cdot \frac{3}{8} + \left(2 - \frac{3}{2}\right)^2 \cdot \frac{3}{8} + \left(3 - \frac{3}{2}\right)^2 \cdot \frac{1}{8} = \frac{3}{4}. \blacksquare$$

Figure 5.1 shows three r.v. with medium, small and large variances.

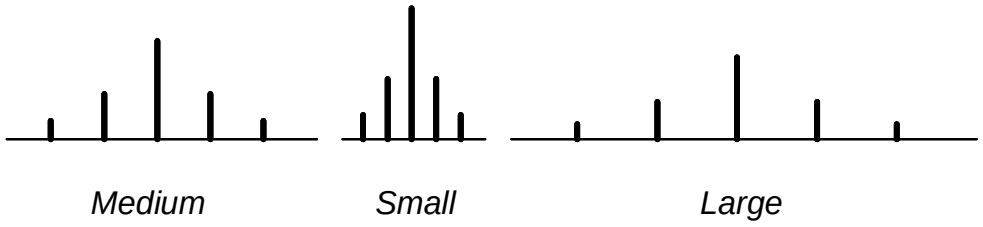

Figure 5.1.

Theorem 4. (Properties of Variance.) *The variance satisfies*

(a.). $var(X) \geq 0$.

(b.) $var(X+a) = var(X)$.

(c.) $var(cX) = c^2 var(X)$.

(d.) $\sigma^2 = \mu_2 - \mu^2$ *where* $\mu_2 = E(X^2)$ *is the second moment.*

(e.) *If* X *and* Y *are independent, then*

$$var(X+Y) = var(X) + var(Y)$$

Proof. Parts (a.) is clear. For (b.), we have

$$E(X+a) = \mu + a$$

So

$$var(X+a) = E\left([(X+a)-(\mu+a)]^2\right) = E\left([X-\mu]^2\right) = var(X)$$

For (c.), we have

$$E(cX) = c\mu$$

So

$$var(cX) = E\left((cX-c\mu)^2\right) = E\left(c^2(X-\mu)^2\right) = c^2 var(X)$$

For (d.), expanding gives

$$\begin{aligned} \sigma^2 &= E\left((X-\mu)^2\right) = E\left(X^2 - 2\mu X + \mu^2\right) = E\left(X^2\right) - 2\mu E(X) + \mu^2 \\ &= E\left(X^2\right) - 2\mu\mu + \mu^2 = \mu_2 - \mu^2. \end{aligned}$$

For (e.), replacing X by $X - \mu_x$ and Y by $Y - \mu_y$, we may assume that X and Y have zero mean. Then

$$\begin{aligned} var(X+Y) &= E(X+Y)^2 = E\left(X^2\right) + E(Y^2) + 2E(XY) \\ &= E\left(X^2\right) + E(Y^2) + 2E(X)E(Y) = E\left(X^2\right) + E(Y^2) \\ &= var(X) + var(Y). \square \end{aligned}$$

Equation (d.) is useful in computing variances.

Example 2. For the number X of Heads in three tosses, we have

$$\mu_2 = E(X^2) = 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{8} = 3$$

and so

$$var(X) = \sigma^2 = \mu_2 - \mu^2 = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4}. \blacksquare$$

5.6 Why Do We Use Mean and Variance?

As we have remarked, the mean μ is an estimate of the center of a distribution; as the statisticians say, it is a "*measure of central tendency*". There are, however, other ways to estimate the center. For example, the *median* of a r.v. is the value m (if there is one) such that

$$P(X > m) = P(X < m).$$

There are problems for which the median is a much more useful quantity than the mean.

Example 1. As an example, some years ago - the numbers are much higher now - the mean salary of Basketball players in the NBA was around $\$\ 1$ million. Nevertheless, about half of the NBA players made the minimum salary of $\$250,000$. The mean was skewed higher by the very high salaries of a small number of star players.

It is indeed a feature of the arithmetic mean that one eccentric measurement can have a very large effect on the value of the mean. This problem of "*outliers*" is a matter of considerable concern to Statisticians.

The median does not have this property. In the case above, the median would be close to $\$250,000$

So why the emphasis on the mean? It is usually stated that the mean is a "*convenient measure*" of the center. That is true, but *what makes it convenient*?

The answer is *Additivity*: The mean satisfies

$$E(X + Y) = E(X) + E(Y)$$

for any r.v.; the median does not. Additivity makes it easy to compute the means of a variety of distributions for which the median is mush less accessible.

Similarly, the variance measures the *fluctuations* of the r.v. about the mean. There are other measures of this; for example why do we take the mean *square*

$$E(X - \mu)^2$$

of the fluctuations? True, this makes the variance positive, but what is wrong with taking, say,

$$E(X - \mu)^4 \text{ or } E\,|X - \mu| \text{ or even } E\,|X - m|$$

particularly when some of these may exist when the variance does not?

Again, the answer lies with *Additivity*. The variance satisfies

$$var(X + Y) = var(X) + var(Y)$$

if X and Y are independent, which makes the variance much easier to compute in many cases.

Nevertheless, in some problems - for example, when means or variances do not exist - other measures may be useful.

5.7 Mean and Variance of the Binomial Distribution.

The number of successes in n trials has the *Binomial distribution*

$$P(X=k)=\binom{n}{k}p^k q^{n-k}$$

so called because of the binomial coefficient in the formula.

If X has this distribution, we say that X has the distribution $Bin(n,p)$.

The mean and variance of the distribution $Bin(p,n)$ can be computed with the aid of the Binomial Theorem.

Theorem 6. *A r.v. X with distribution $Bin(p,n)$ has mean np and variance npq.*

Proof. By definition,.

$$E(X)=\sum_{k=0}^{n} kP(X=k)=\sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k}$$

This can be summed by direct computation, using the Binomial Theorem. We have

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

Thus

$$\sum_{k=0}^{n} k\binom{n}{k}x^k y^{n-k}=x\frac{\partial}{\partial x}(x+y)^n=xn(x+y)^{n-1}$$

Let $x=p$ and $y=q$ to get

$$E(X)=\sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k}=pn(p+q)^{n-1}=np.\blacksquare$$

Similarly, by differentiating twice, we can get $var(X)=npq$.$\blacksquare$

Use of Indicators.

There is a much easier way to compute the mean and variance of the Binomial distribution. For this, we introduce the useful device of the *indicator of an event*.

Let S be an event with $P(S)=p$ and $q=1-p=P(F)$. We define the *Indicator 1_S of S* to be the r.v.

$$1_S=\begin{cases}1 & \text{if } S \text{ occurs}\\ 0 & \text{if } S \text{ does not occur}\end{cases}$$

Another way of saying this is the 1_S is the number of times S occurs on the given trial. This is, of course, either 0 times or 1 times. The indicator has a Binomial distribution with the number of trials $n=1$. Indicators look strange at first glance, but they are frequently quite useful. To illustrate, we compute the mean and variance of the Binomial distribution.

Theorem 7. *The r.v.* 1_S *has mean* p *and variance* pq.

Proof. We have

$$E(1_S) = 1 \cdot P(S) + 0 \cdot P(F) = 1 \cdot p + 0 \cdot q = p.$$

If we note that $1_S^2 = 1_S$, then we have also

$$var(1_S) = E(1_S^2) - (E(1_S))^2 = E(1_S) - (E(1_S))^2 = p - p^2 = p\,(1-p) = pq.\square$$

Let X_n be the number of Successes in n Bernoulli trials, and let S_k be Success on the k^{th} trial. Then the number of Successes in n trials is

$$X_n = 1_{S_1} + \cdots + 1_{S_n}.$$

The means add so $E(X) = np$,

Observe next that events are independent iff their indicators are independent r.v. Hence, the variances add as well, so that

$$var(X) = npq.$$

5.8 The Sample Mean and Variance.

Let X be a r.v. with mean μ, and suppose that we would like to know what μ is.

Suppose, for example, that we wish to know the average weight of professional Statisticians in the U.S. We could, of course, simply weigh all Statisticians and calculate the mean, but no one - except possibly the Census Bureau - would actually do such a thing. What one would do is to weigh some subset of Statisticians, calculate the mean weight of that subset and hope that the result would be close to the actual mean μ. We call this an *estimate* of the mean μ. *Estimation* of a parameter of a distribution, such as the mean μ or the variance σ^2 is one of the chief problems of Statistics.

A sequence $X_1, X_2, ..., X_n$ of n independent measurements of a r.v. X, is called a *sample* of X of size n. The problem of estimation is to estimate the value of the parameter from the sample values; that is, we need to find a *function of the sample values* alone that we have good reason to believe will be close to the parameter. Such a function is called a *Statistic*.

The Sample Mean.

We will consider briefly the problem of estimation for the mean μ. Let X be a r.v. with mean μ and variance σ^2, and let $X_1, X_2, ..., X_n$ be n independent measurements of X. The natural thing to do is to estimate the mean μ of X by the average

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

of the measured values of X. The number $\bar{X}_n$ is called the *sample mean*. It is a *statistic* - a function of the sample values.

This may be the natural thing, but exactly *why is this a good way to estimate* μ ? The first thing to realize about $\bar{X}_n$ is that *it is a random variable*. That is, if a second sample of size n is taken, $\bar{X}_n$ will probably *take a different value*. Second, the value obtained from your sample *may* be widely different from the true mean: you may, by an unlucky chance, have weighed the 100 fattest Statisticians in the entire country.

Now admittedly, *this is quite unlikely* - and that is the point. To pursue this further, let us compute the mean and variance of $\bar{X}_n$.

Theorem 8. *$\bar{X}_n$ has mean μ and variance σ^2/n.*

Proof. For the mean, we have

$$\begin{aligned} E(\bar{X}_n) &= \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n}(E(X_1) + E(X_2) + \cdots + E(X_n)) = \frac{1}{n} \cdot n\mu = \mu. \end{aligned}$$

and for the variance,

$$\begin{aligned} var\left(\bar{X}_n\right) &= var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \left(\frac{1}{n}\right)^2 var\left(X_1 + X_2 + \cdots + X_n\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.\square \end{aligned}$$

A special case of the sample mean is the *frequency ratio of a sequence of Bernoulli trials*. The frequency ratio ν_n is defined to be the number of Successes X_n divided by the total number n of trials

$$\nu_n = \frac{X_n}{n} = \frac{1_{S_1} + \cdots + 1_{S_n}}{n}$$

in the notation of the preceding section.

Corollary 4. *The frequency ratio ν_n has mean p and variance pq/n.*

Thus, the sample mean $\bar{X}_n$ has expectation equal to the mean μ, and if n is large, the variance σ^2/n of $\bar{X}_n$ is small. This means intuitively that the values of $\bar{X}_n$ do not fluctuate very much away from its mean μ. So $\bar{X}_n$ *should*, with high probability, be a good estimate of μ. That this is so is the content of the *Law of Large Numbers*, which is discussed the next section.

*The Sample Variance.

Suppose now that we also wish to find a statistic to estimate the *variance* σ^2 of X. A natural choice is the *sample variance*

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right).$$

One divides by $n-1$ rather than by n, as one might think, to make the expectation of S_n^2 equal to σ^2. For the proof of this fact, the following formula is quite convenient.

Lemma 1. *For any sequence $X_1, X_2, ..., X_n$ of r.v.,*

$$\sum_i \left(X_i - \bar{X}\right)^2 = \sum_i X_i^2 - n\bar{X}^2$$

Proof.

$$\begin{aligned}\sum_i \left(X_i - \bar{X}\right)^2 &= \sum_i \left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right) = \sum_i X_i^2 - 2\bar{X}\sum_i X_i + \sum_i \bar{X}^2 \\ &= \sum_i X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_i X_i^2 - n\bar{X}^2.\square\end{aligned}$$

Theorem 9. *S_n^2 has mean σ^2.*

Proof. Since

$$\sum_i E\left(X_i - \bar{X}_n\right)^2 = \sum_i E\left((X_i - \mu) - (\bar{X}_n - \mu)\right)^2$$

we may assume that $\mu = 0$. For the mean, we have

$$\begin{aligned}(n-1)\,E(S_n^2) &= \sum_i E\left(X_i - \bar{X}_n\right)^2 = \sum_i E(X_i^2) - nE(\bar{X}^2) \\ &= n\sigma^2 - n\cdot\frac{\sigma^2}{n} = (n-1)\,\sigma^2.\square\end{aligned}$$

We can also compute the variance of S_n^2.

Theorem 10. *S_n^2 has variance*

$$var(S_n^2) = \frac{1}{n}\left[\tilde{\mu}_4 - \frac{n-3}{n-1}\sigma^4\right]$$

Here $\tilde{\mu}_4$ is the 4^{th} *central* moment:$\tilde{\mu}_4 = E(X-\mu)^4$.

The proof is a rather dreadful computation, but here it is.

Proof. Again, we may assume that $\mu = 0$. We have

$$\begin{aligned}(n-1)^2\,S_n^4 &= \left(\sum_i X_i^2 - n\bar{X}^2\right)^2 = \sum_{i\ j} X_i^2X_j^2 - 2n\bar{X}^2\sum_i X_i^2 + n^2\bar{X}^4 \\ &= \sum_{i\ j} X_i^2X_j^2 - \frac{2}{n}\left(\sum_i X_i\right)^2\sum_i X_i^2 + n^2\left(\sum_i X_i\right)^4 \\ &= \sum_{i\ j} X_i^2X_j^2 - \frac{2}{n}\sum_i X_i^2X_jX_k + \frac{1}{n^2}\sum_{ijkl} X_iX_jX_kX_l\end{aligned}$$

Thus

$$(n-1)^2 E(S_n^4) = \sum_{i\ j} E(X_i^2 X_j^2) - \frac{2}{n}\sum_{i\,j\,k} E(X_i^2 X_j X_k) + \frac{1}{n^2}\sum_{i\,j\,k\,l} E(X_i X_j X_k X_l) \quad (5.1)$$

It remains to evaluate the expectations under the summation signs.

(1.) Since the X_i are independent, and $E(X_i) = 0$, $E(X_i X_j X_k X_l)$ is zero unless the X_i's are paired; for example,

$$E(X_1^2 X_2 X_3) = E(X_1^2)(E(X_2)E(X_3) = 0.$$

Thus, $E(X_i^2 X_j X_k) = 0$ unless $j = k$, so

$$\sum_{i\,j\,k} E(X_i^2 X_j X_k) = \sum_{i\ j} E(X_i^2 X_j^2)$$

(2.) We have

$$E(X_i^2 X_j^2) = \begin{cases} \tilde{\mu}_4 & \text{if } i = j \quad n \text{ terms} \\ \sigma^4 & \text{if } i \neq j \quad n^2 - n \text{ terms} \end{cases}$$

(3.) Also,

$$E(X_i X_j X_k X_l) = \begin{cases} \tilde{\mu}_4 & \text{if } i = j = k = l \quad n \text{ terms} \\ \sigma^4 & \text{if indices occur in pairs} \quad 3n(n-1) \text{ terms} \\ 0 & \text{otherwise} \end{cases}$$

The count on the σ^4 terms is obtained as follows. First choose two indices, say 1 and 2: there are $\binom{n}{2}$ ways. There are then 6 ways to arrange 1122 in sequence. This gives

$$6\binom{n}{2} = 3n(n-1)$$

terms.

(4.) Substituting these values into (5.1) gives

$$\begin{aligned} (n-1)^2\, var(S_n^2) &= (n-1)^2 \left[E(S_n^4) - \sigma^4\right] \\ &= \left[1 - \frac{2}{n}\right]\sum_{i\ j} E(X_i^2 X_j^2) + \frac{1}{n^2}\sum_{i\,j\,k\,l} E(X_i X_j X_k X_l) - (n-1)^2\sigma^4 \\ &= \left[1 - \frac{2}{n}\right]\left(n\tilde{\mu}_4 + n(n-1)\sigma^4\right) \\ &\quad + \frac{1}{n^2}\left[n\tilde{\mu}_4 + 3n(n-1)\sigma^4\right] - (n-1)^2\sigma^4 \\ &= \frac{(n-1)^2}{n}\tilde{\mu}_4 - (n-1)(n-3)\sigma^4. \end{aligned}$$

The result follows.□

Remark. The statistic S_n^2 can also be written in the form

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_i \left(X_i - \bar{X}_n\right)^2 = \frac{1}{n-1}\left[\sum_i X_i^2 - n\bar{X}_n^2\right] \\ &= \frac{1}{n-1}\left[\sum_i X_i^2 - \frac{1}{n}\left(\sum_i X_i\right)^2\right] \end{aligned}$$

which is better for calculations, since it reduces round off error. (See problem.2.)

5.9 The Law of Large Numbers.

The Law of Large Numbers was historically the first Limit Theorem of Probability to be discovered. It is due, in a special case, to James Bernoulli, and first appears his book *Ars Conjectandi - The Art of Guessing* - published posthumously in 1712.

It is concerned with a very basic question at the heart of the definition of probability. We have defined, intuitively, the probability of an event S to be the "asymptotic frequency of occurrence" of the event. That is, if a large number n of trials is run, and the event occurs on m of these trials, then we expect that the probability of S is close to the frequency ratio:

$$\nu_n = \frac{m}{n} \simeq p = P(S).$$

Now this seems to be in direct contradiction to the notion of independence, because if the trials are independent in the sense that the outcome of any trial is not influenced by the results of any previous trials, then it would seem that it is entirely possible, in principle, to obtain absolutely any sequence of results whatever, so that m could, in principle, be any number between 0 and n, and consequently m/n could be very far from p. *And this is true: according the theory*. For by the Binomial Formula - except in the trivial cases that p is zero or one - there is, for any $m \leq n$, a positive probability equal to

$$\binom{n}{m} p^m q^{n-m}$$

of coming up with exactly m Successes.

The resolution of this seeming paradox is that while the frequency ratio m/n *may* differ substantially from $P(S)$, for large n it is *very unlikely to do so*. For any positive number ϵ, the probability that ν_n differs from p by more than ϵ is small if n is large enough. To be exact,

$$\lim_{n\to\infty} P(|\nu_n - p| \geq \epsilon) = 0.$$

Chebychev's Inequality.

In order to prepare for the proof, we must first prove two very simple - and therefore very general - inequalities of the Russian school.

Theorem 11. (Markov's Inequality.) *Let $X \geq 0$ be a positive r.v. Then for $t > 0$,*

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

Proof. We have

$$X \geq t1_{\{X \geq t\}}$$

and so

$$E(X) \geq tE(1_{\{X \geq t\}}) = tP(X \geq t).\blacksquare$$

Corollary 5. (Chebychev's Inequality.) *Let X be a r.v. with mean μ and variance σ^2. Then for $\epsilon > 0$,*

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Proof. By Markov's inequality,

$$P(|X - \mu| \geq \epsilon) = P(|X - \mu|^2 \geq \epsilon^2) \leq \frac{E\left(|X - \mu|^2\right)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}.\blacksquare$$

Remark. The importance of Chebyshev's inequality lies in its great generality, not in its accuracy. That is to say, in most cases of interest, the left side is quite likely to be not only *less* than the right side, but a *lot less*. It is therefore largely a theoretical device. We will return to this point when we discuss the Central Limit Thoerem.

The Weak Law of Large Numbers.

We can now state the Law of Large Numbers. Actually, we will prove a more general result that applies to means of samples rather than frequency ratios, which, as we saw in section 5.8 are a special case of sample means.

Let X be a r.v. with mean μ and variance σ^2. Consider n independent measurements of X. Let X_k be the measured value of X on the k^{th} trial. As we proved in section 5.8, the *sample mean*

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is a r.v. with *mean μ and variance σ^2/n.*

Theorem 12. (Weak Law of Large Numbers.) *For every $\epsilon > 0$,*

$$\lim P(\left|\bar{X}_n - \mu\right| \geq \epsilon) = 0.$$

Proof. By Chebychev's inequality, applied to $\bar{X}_n$,

$$P(\left|\bar{X}_n - \mu\right| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0.\blacksquare$$

Applied to Bernoulli trials, this gives the

Corollary 6. (Weak Law of Large Numbers.) *Let S be an event with probability $p = P(S)$, and the frequency ratio after n trials. Then for every $\epsilon > 0$,*

$$\lim P(|\nu_n - p| \geq \epsilon) = 0.$$

5.10Variance of Sums.

In general, if $Y_n = X_1 + \cdots + X_n$ where the X_i are *not independent*, the variances do not add. However, we have

Theorem 13.

$$E(Y_n^2) = E\left(X_1 + \cdots + X_n\right)^2 = \sum_{i=1}^{n} E\left(X_i^2\right) + 2 \sum_{1 \leq i < j \leq n} E\left(X_i X_j\right).$$

Proof. Expand the square to get

$$\left(X_1 + \cdots + X_n\right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j = \sum_{i=1}^{n} X_i^2 + 2 \sum_{1 \leq i < j \leq n} X_i X_j.$$

Taking expectations gives the result.□

Example 1. (*Polya's urn model*.) For a very simple example, consider Example 2 of section 3.1 An urn contains 1 Red and 1 White ball. A ball is drawn, replaced and an additional ball of the color drawn is added. A second ball is then drawn. Find the mean and variance of the number X of reds balls drawn.

Solution. We have $X = 1_{R_1} + 1_{R_2}$ where R_k is the event '*Red on the k^{th} draw*'. Thus, referring to the previous example,

$$\begin{aligned}
E(X) &= E(1_{R_1}) + E(1_{R_2}) = \frac{1}{2} + \frac{1}{2} = 1. \\
E(X^2) &= E(1_{R_1}) + E(1_{R_2}) + 2E(1_{R_1 R_2}) \\
&= P(R_1) + P(R_2) + 2P(R_1 R_2) \\
&= \frac{1}{2} + \frac{1}{2} + 2 \cdot \frac{1}{2} \cdot \frac{2}{3} = \frac{5}{3} \\
var(X) &= E(X^2) - E(X)^2 = \frac{5}{3} - 1 = \frac{2}{3}.\blacksquare
\end{aligned}$$

***Example 2.** (*Rencontre.*) In the Hat Check problem, find the mean and variance of the number N of those receiving their own hats.

Solution. Let H_k be the event that the k^{th} gentleman gets his own hat. The number of those receiving their own hats is

$$N = 1_{H_1} + \cdots + 1_{H_n}$$

For every k,

$$P(H_k) = P(H_1) = \frac{1}{n}$$

Hence, the mean is

$$\begin{aligned} E(N) &= E(1_{H_1} + \cdots + 1_{H_n}) \\ &= E(1_{H_1}) + \cdots + E(1_{H_n}) = nP(H_1) = n \cdot \frac{1}{n} = 1 \end{aligned}$$

For the variance, we have

$$P(H_i H_j) = P(H_1 H_2) = \frac{1}{n} \cdot \frac{1}{n-1}$$

for every $i \neq j$, so

$$\begin{aligned} E(N^2) &= E(1_{H_1} + \cdots + 1_{H_n})^2 \\ &= \sum_{j=1}^{n} E(1_{H_j}^2) + 2 \sum_{1 \leq i < j \leq n} E(1_{H_i} 1_{H_j}) = nP(H_1) + 2\binom{n}{2} P(H_1 H_2) \\ &= n \cdot \frac{1}{n} + 2\frac{n\,(n-1)}{1 \cdot 2} \frac{1}{n} \frac{1}{n-1} = 2. \end{aligned}$$

and hence

$$var\,(N) = E(N^2) - (E(N))^2 = 2 - 1 = 1.$$

Note that the mean and variance are *independent of the number of hats.* ■

5.11The Hypergeometric Distribution.

The Hypergeometric formula of Theorem 4 of section 2.6 is the distribution of a r.v. obtained by sampling without replacement.

Suppose an urn has a Red and b White balls. If n balls are drawn without replacement, then the number X of Red balls drawn has the *Hypergeometric distribution*

$$P(X = k) = \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}.$$

The mean and variance can be computed by the method of the previous section.

Theorem 14 *The hypergeometric distribution has mean and variance*

$$\mu = np \text{ and } \sigma^2 = npq\left[\frac{N-n}{N-1}\right]$$

where $N = a + b$, $p = a/N$ *and* $q = 1 - p$.

Proof. Letting R_k be the event of Red on the k^{th} draw, the number of Reds in n draws is

$$X = 1_{R_1} + \cdots + 1_{R_n}.$$

The expectation is

$$\begin{aligned} E(X) &= E\left(1_{R_1} + \cdots + 1_{R_n}\right) = E\left(1_{R_1}\right) + \cdots + E\left(1_{R_n}\right) = nE\left(1_{R_1}\right) \\ &= n\frac{a}{a+b} = np. \end{aligned}$$

The variances do not add since $1_{R_1}, \cdots, 1_{R_n}$ are not independent. However,by Theorem 13,

$$\begin{aligned} \mu_2 &= E(X)^2 = E\left(1_{R_1} + \cdots + 1_{R_n}\right)^2 = \sum_i E\left(1_{R_i}^2\right) + 2\sum_{i<j} E\left(1_{R_i}1_{R_j}\right) \\ &= nE\left(1_{R_1}\right) + 2\binom{n}{2}E\left(1_{R_1}1_{R_2}\right) = n\frac{a}{a+b} + n\left(n-1\right)\frac{a}{a+b}\frac{a-1}{a+b-1}. \end{aligned}$$

A computation shows that

$$\sigma^2 = \mu_2 - \mu^2 = npq\left[\frac{N-n}{N-1}\right].$$

where $N = a + b$, $p = a/N$ and $q = 1 - p$.□

*Symmetries of the Hypergeometric Distribution.

The Hypergeometric distribution has three notable symmetries, all of which have probabilistic meanings.

Let $N = a + b$ be the total number of balls, n the number drawn, a be the number of Red balls drawn, and $b = N - a$ the number of Whites drawn. Write the distribution as

$$P(k; N, a, n) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}.$$

(1.) If n balls are selected, selecting k Reds from a Reds is same event as selecting $n - k$ Whites from $N - a$ Whites. So we must have

$$P(n - k; N, N - a, n) = P(k; N, a, n).$$

(2.) Selecting n balls to *take out* is the same as selecting $N - n$ balls to *stay in*; selecting k Reds to take out is the same as selecting $a - k$ Reds to stay in, so we must have

$$P(a - k; N, a, N - n) = P(k; N, a, n).$$

(3.) Both of the symmetries (1.) and (2.) above are evident from looking at the hypergeometric formula. The third symmetry also becomes clear if the binomial coefficients are written out in terms of factorials, but it is not immediately clear intuitively. It is the following:

$$P(k; N, n, a) = P(k; N, a, n).$$

Start with N all White balls. Select a of them and paint them Red, leaving the remainder White. Put them in an urn and then select n balls and paint a black stripe on them. The probability of exactly k Red balls having stripes is clearly $P(k; N, n, a)$.

Now start with N all White balls. Select n balls and paint a Black stripe on them. Put them in an urn and then select a balls and paint them Red, (being careful not to paint over the stripe if there is one). The probability of exactly k striped balls being Red is $P(k; N, a, n)$.

But these numbers are the same. We have simply assigned randomly and independently two different classifications to the balls: Color and Stripedness. In the Hypergeometric case, we assigned the properties of Color and Selectedness. It does not matter in what order we do this.

5.12Covariance and Correlation.

Definition 6. The *covariance* of two r.v. is defined to be

$$cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

The following properties of covariance are quite evident.

Theorem 15. (Properties of Covariance.)

(a.) $cov(X, X) = var(X)$

(b.) $cov(X, Y) = cov(X, Y)$

(c.) $cov(X + Y, Z) = cov(X, Z) + cov(Y, Z)$

(d.) $cov(aX, Y) = a\, cov(X, Y)$

(e.) *If X and Y are independent, then ,* $cov(X,Y) = 0$.

Proof. Parts (a.) - (d.) are clear. For (e.), we have

$$cov(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right] = E(X-\mu_x)E(Y-\mu_y) = 0.\square$$

By Schwarz's inequality, we have

Corollary. 3.

$$|cov(X,Y)|^2 \leq var(X)var(Y)$$

The variance of a sum can be expressed in terms of the covariance.

Theorem 16. *If $S_n = X_1 + \cdots + X_n$, then*

$$\sigma^2(S_n) = \sum_i var\left(X_i\right) + 2\sum_{i<j} cov\left(X_iX_j\right)$$

Proof. Let $Y_i = X_i - \mu$. By Theorem 13,

$$\begin{aligned}\sigma^2(S_n) &= E(S_n - n\mu)^2 = E\left(Y_1 + \cdots + Y_2\right)^2 = \sum_i E\left(Y_i^2\right) + 2\sum_{i<j} E\left(Y_iY_j\right) \\ &= \sum_i var\left(X_i\right) + 2\sum_{i<j} cov\left(X_iX_j\right).\square\end{aligned}$$

Correlation.

Definition 7. The *Correlation Coefficient* of two r.v. X and Y is defined to be

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}.$$

From Corollary 3, applied to $X - \mu_x$ and $Y - \mu_y$, we see that $-1 \leq \rho \leq 1$.

The correlation coefficient ρ is a measure of the relationship between X and Y which is *scale invariant*; that is, if X or Y is multiplied by a constant, the correlation remains the same, while the covariance changes. In the physicists terms, ρ is a *dimensionless quantity*. For example, if X and Y are lengths, ρ is the same whether we measure X and Y in feet or inches or meters or light years, whereas the covariance is different in all these cases.

What the correlation coefficient measures is *the extent to which Y is a linear function of X*. In the extreme case $\rho = \pm 1$, Corollary 1 of section 5.4 implies that X and Y are linearly related. On the other hand, if X and Y are independent - the opposite of being related - then by (e.), $\rho = 0$.

In general, however, $\rho = 0$ *does not imply that* X *and* Y *are independent.* The correlation measures only *linear* dependence. We must stress *linear,* since, for example, if Z is $N(0,1)$, then

$$\rho(Z, Z^2) = 0$$

so that Z and Z^2 are uncorrelated, although Z^2 is clearly a function of Z.

Example 3. Let X_1 and X_2 be r.v. with the same distribution, for example, the numbers on two dice. Then $X = X_1 + X_2$ and $Y = X_1 - X_2$ have zero correlation. For

$$\begin{aligned} E(Y) &= E(X_1) - E(X_2) = 0 \\ E(XY) &= E(X_1^2 - X_2^2) = E(X_1^2) - E(X_2^2) = 0 \end{aligned}$$

and hence

$$cov(X, Y) = E(XY) - E(X)E(Y) = 0.$$

However, X and Y may well be dependent. For example, if X_1 and X_2 are the numbers on two dice, then X and Y. are either both even or both odd, and so not independent. ■

Example 4. For a second example, if X is a r.v. with $E(X) = E(X^3) = 0$, and $Y = X^2$, then

$$cov(X, Y) = E(X \cdot X^2) - E(X)E(X^2) = 0.$$

Thus $\rho = 0$ although Y is actually a function - but not a *linear* function - of X. ■

5.13*Inclusion-Exclusion.

Using expectations, we can now present a third proof of the Inclusion-Exclusion formula.

Theorem 17. (Inclusion-Exclusion.) *Let* X_1,X_2,...,X_n *be any r.v., and*

$$S_k = \sum_{i_1 < i_2 < \cdots < i_k} E(X_{i_1} X_{i_2} \cdots X_{i_k})$$

Then

$$E\left(\prod_{i=1}^{n} (1 - X_i)\right) = 1 - S_1 + S_2 - S_3 + \cdots (-1)^n S_n.$$

Proof. As in the proof of the Multinomial Theorem,

$$\prod_{i=1}^{n} (x_i + y_i) = \sum u_1 u_2 \cdots u_n$$

over all possible products where u_i is either x_i or y_i.Applying this to the product

$$\prod_{i=1}^{n} (1 - X_i)$$

where each u_i is either $-X_i$ or 1, we see that sum consists of all products of the form

$$(-1)^r X_{i_1} X_{i_2} \cdots X_{i_r}$$

where $i_1 < i_2 < \cdots < i_r$ and $0 \leq r \leq n$. That is,

$$\begin{aligned} & \prod_{i=1}^{n} (1 - X_i) \\ = \; & 1 - (X_1 + X_2 + \cdots + X_n) + \cdots \\ & + (-1)^k \sum_{i_1 < i_2 < \cdots < i_k} X_{i_1} X_{i_2} \cdots X_{i_k} + \cdots + (-1)^n (X_1 X_2 \cdots X_n) . \end{aligned}$$

Taking expectations gives the result. □

Now let E_i be any sequence of events, and set $X_i = 1_{E_i}$. Then $X_{i_1} X_{i_2} \cdots X_{i_r}$ is the indicator of the event $E_{i_1} E_{i_2} \cdots E_{i_k}$ so that

$$E[X_{i_1} X_{i_2} \cdots X_{i_r}] = P(E_{i_1} E_{i_2} \cdots E_{i_r})$$

Theorem 6 of Chapter II now follows.

5.14 Problems.

(1.) Show that the variance of the Hypergeometric distribution can be written as

$$\sigma^2 = npq \left[\frac{N-n}{N-1} \right] .$$

where $N = a + b$, $p = a/N$ and $q = 1 - p$.

(2.) Show that the sample variance

$$S_n^2 = \frac{1}{n-1} \sum_i \left(X_i - \bar{X}_n \right)^2$$

can be written as

$$S_n^2 = \frac{1}{n-1} \left[\sum_i X_i^2 - \frac{1}{n} \left(\sum_i X_i \right)^2 \right]$$

(3.) By differentiating the Binomial theorem twice, show that the variance of the Binomial distribution is npq.

(4.) In *Banach's Match Box problem* of section 4.2, find the distribution of the number of matches left in other box when one box is found to be empty.

(5.) Prove that the Geometric distribution is the *only* discrete Markov waiting time.

(6.) (*Problem of Points*.) The Problem of Points, which we solved above was first solved successfully by Pascal and Fermat in a series of letters in 1654. Prior to that, it was treated by Pacioli (1494), Tartaglia.(1556) and Peverone (1558), largely as a problem in proportions. The specific problem treated by these authors was as follows: Two players A and B play for $n = 6$ wins, with equal chances to win, but stop at $a = 5$ games to $b = 3$. How are stakes divided?. Paciolli says $5 : 3$. Tartaglia's method was to divide the stakes in the ratio

$$n + (a - b) : n - (a - b)$$

which in this case is $2 : 1$.

The modern approach is to divide according to the probability of wining, which is $7 : 1$ in this case. Why is the modern method correct? The logic behind it is that if B is permitted to quit with this payoff, it is to his advantage to do so.

For a simple case, let A and B play to three games, but assume that B plays on at all times, but that A drops out at 1 game to 2 against him. Compute A's expected gain per dollar bet.

(7.) A roulette wheel has the numbers 0 through 36 and a double 0 (38 numbers in all). A player bets on the numbers 1 through 12. What is the probability that he loses five consecutive times? What is his average wait between wins?

(8.) Find the range of the random variables Example 1 of section 1.

(9.) Classify the random variables in Example 1 of section 1as *finite discrete*, *infinite discrete* or *continuous.*

(10.) Compute variance of the Binomial distribution $Bin(n, p)$ directly by differentiating Binomial Theorem twice.

(11.) Two fair dice are rolled. Find the range, distribution, mean and variance of the following random variables:

(a.) the maximum of the two numbers,

(b.) the minimum of the two numbers,

(c.) the sum,

(d.) the product,

(e.) the value of the first die minus the value of the second.

(12.) A basketball player shoots a "one-and-one". (This means that he will shoot one shot and if he makes it, will shoot another shot.) Each shot counts one point. Assume that his shots are independent and that his probability of making a shot if 0.8 (an 80% shooter). Let X be the number of points scored.

Find the mean and distribution of X.

(13.) A basketball player shoots foul shots with probability 80%. Find the distribution and the expectation of the points scored on

(a.) a two-shot foul,

(b.) a one-and-one,

(c.) three to make two.

(14.) A fair die is rolled, giving a number X_1, and then a second is rolled until the number X_2 on the second die is *no greater than* X_1. Find the distribution and mean of the sum $X_1 + X_2$.

(15.) (*Chuck-a-luck*) A player pays an entrance fee of $\$1$ selects a number between 1 and 6 and rolls three dice. If his number appears, he receives his fee back, plus $\$1$ for each time it appears. Find the player's expected gain. Also called *Wheel-of-Fortune*, this game is sometimes played at carnivals by spinning a wheel with 216 slots each labeled with 3 numbers from 1 to 6.

(16.) (*The blood-testing problem*) A large number, N, of people are subject to a blood test. This can be administered in two ways.

(*i.*) Each person can be tested separately. In this case N tests are required.

(*ii.*) The blood samples of k people can be pooled and analyzed together. If the test is negative, this one test suffices for the k people. If the test is positive, each of the k persons must be tested separately, and in all $k + 1$ tests are required for the k people.

Assume the probability p that the test is positive is the same for all people and that people are stochastically independent.

(a.) What is the probability that the test for a pooled sample of k people will be positive?

(b.) What is the expected value of the number, x, of tests necessary under plan (*ii.*)?

(c.) Find an equation for the value of k which will minimize the expected number of tests under the second plan.

(d.) Show that this k is close to $1/\sqrt{p}$, and hence that the minimum expected number of tests is about $2n\sqrt{p}$.

(17.) Let X have mean μ and variance σ^2. Find the mean and variance of

$$Z = \frac{X - \mu}{\sigma}.$$

(18.) Let Z have mean 0 and variance 1. Find the mean and variance of

$$X = \sigma Z + \mu.$$

Section 5.14 Problems.

(19.) Find the mean and variance of the sum of n random digits.

(20.) Two fair dice are rolled. Find the distribution, mean and variance of the *maximum* of the two numbers rolled.

(21.) A jar contains 135 bills, of the following denominations: 100 \$1 bills, 20 \$5 bills, 10 \$10 bills, and 5 \$20 bills. For an entry fee of \$3, a player may draw at random (and keep!) a bill from the jar.

Compute the distribution, mean and variance of his gain.

(22.) Let X_1, X_2, X_3 and X_4 be independent random variables with mean μ and variance σ^2. Express the mean and variance of $Y = X_1X_2 + X_3X_4$ in terms of μ and σ^2.

(23.) Let X_1 X_2 and X_3 be independent random variables with mean $\mu = 1$ and variance $\sigma^2 = 3$. Find, in terms of μ and σ^2, the mean and variance of the random variable $Y = X_1X_2 + X_3$.

(24.) Let $X_1, X_2, ..., X_n$ be a sequence of independent r.v. with the same mean μ, variance σ^2, and 4^{th} moment $\mu_4 = E(X^4)$.

In terms of n, μ, σ^2, and μ_4, find the mean and variance of

$$Y = X_1^2 + X_2^2 + ... + X_n^2.$$

(25.) Let X_1 X_2 and X_3 be independent random variables with mean μ and variance σ^2. Find, in terms of μ and σ^2, the mean and variance of the following random variables:

(a.) $S = 2X_1 - X_2$.

(b.) $Y = X_1X_2$.

(26.) Teams A and B play a series in which each team has equal probability to win each game. The first team to win three games wins the series. (A "best of five series".)

(a.) What is the probability that the series goes the full five games?

(27.) (a.) Prove that

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

(b.) Use the result to find the mean of the Binomial distribution.

*(c.)Show that a similar trick also works for the multinomial distribution.

(28.) If X is $Bin(n, p)$, show that

$$E(\frac{1}{X+1}) = \frac{1 - q^{n+1}}{(n+1)\,p}$$

(29.) Compute the marginal distributions of the *Multinomial* distribution.

(30.) Compute the marginal distributions of the *Generalized Hypergeometric* distribution.

(31.) Prove that the mean is the number a that minimizes $f(a) = E(X - a)^2$.

(32.) (*A Shoe Problem.*) A closet contains n pairs of shoes. If r shoes are removed at random, find the (a.) distribution, (b.) mean and (c.) variance of the number X of complete pairs removed.

(33.) (*Polya's Urn Scheme.*) An urn contains b black and r red balls. A ball is drawn at random, replaced, and then c balls of the color drawn are added. The process is repeated.

(a.) Let $b = c = r = 1$. Find the distribution of the number of black balls in n draws.

(b.) Solve (a.) for general b, c, and r.

(c.) Find the expected number of black balls in n draws.

(34.) The following are Quetelet's data on the chest measurements, in inches, of 5758 Scotish soldiers. (Stigler, p. 207.)

Circumference	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
Number	3	18	81	185	420	740	1075	1079
Circumference	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**
Number	934	658	370	92	50	21	4	1

Compute the sample mean and variance.

Chapter 6
Infinite Discrete Distributions.

6.1 Introduction.

There are r.v which may take on an infinite list of values. We have already met with some. The value of the wait for the first - or more generally - the r^{th} success at Bernoulli trials may, in principle, be any positive integer. The *expectation* of such a r.v. is given by the same formula as for a finite number of values,

$$\mu = E(X) = \sum_{n=1}^{\infty} x_n p_n$$

where $p_n = P(X = x_n)$. The only difference is that that the sum is now an infinite series.

The *variance* is again given by

$$\sigma^2 = E(X - \mu)^2.$$

The properties of the expectation and variance are the same as before.

There is one caveat, however. Since the sum for the mean is now an *infinite series*, there is a question of *convergence*. Of course, if the values x_n of X are bounded, say by a positive number M, then the series converges to a number μ with $|\mu| \leq M$. However, if X can take on arbitrarily large values, then the series may well diverge, and indeed, there are naturally occurring and useful r.v. for which this actually happens.

Such r.v. do not have expectations. Moreover, the series for $E(X)$ may converge,.while that for $E(X^2)$ diverges. In this case, X has a mean but no variance. (Note, however, that bySchwarz's inequality, a finite variance implies a finite mean.)

This is *not just a mathematical subltety*. These r.v. behave differently in some important respects. For example, for such r.v., the arithmetic mean over a large number of trials may not be a useful quantity.

One needs to be aware of this possibility, and *always stop to ask the question whether some expectation is finite or not*.

We shall meet some r.v. without expectations in section 6.4.

6.2 The Geometric and Pascal Distributions.

The wait T for the first Success at Bernoulli trials has the *geometric distribution*

$$P(T = n) = pq^{n-1} \text{ for } n = 1, 2, 3...$$

Theorem 1. *The geometric distribution has mean* $1/p$ *and variance* q/p^2.

Proof. Since T has an infinite number of values, the expectation is an infinite sum:

$$E(T) = \sum_{n=1}^{\infty} nP(T = n) = \sum_{n=1}^{\infty} npq^{n-1} = p\sum_{n=1}^{\infty} nq^{n-1}$$

This series can be summed by the same method we used for the Binomial mean. Consider the geometric series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

Differentiating gives

$$\sum_{n=1}^{\infty} nx^{n-1} = \frac{d}{dx}\sum_{n=0}^{\infty} x^n = \frac{d}{dx}\frac{1}{1-x} = \frac{1}{(1-x)^2}$$

(The sum is from 1 to n, not 0 to n, because differentiation kills the first term, which is a constant.) Differentiating a second time gives

$$\sum_{n=1}^{\infty} n^2x^{n-1} = \frac{d}{dx}\sum_{n=1}^{\infty} nx^n = \frac{d}{dx}\frac{x}{(1-x)^2} = \frac{1+x}{(1-x)^3}.$$

Hence, the mean is

$$E(T) = p\sum_{n=1}^{\infty} nq^{n-1} = p\frac{1}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}.$$

For the variance, we have

$$E(T^2) = \sum_{n=1}^{\infty} n^2P(T = n) = \sum_{n=1}^{\infty} n^2pq^{n-1} = p\sum_{n=1}^{\infty} n^2q^{n-1} = p\frac{1+q}{(1-q)^3} = \frac{1+q}{p^2}$$

and hence

$$var(T) = \frac{1+q}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{q}{p^2}.\blacksquare$$

Example 1. The expected wait for the first six on successive rolls of a die is $1/p = 6$, a very intuitive result.

A similar calculation can be made for the wait T_r for the r^{th} Success, which has the *Pascal, or negative binomial, distribution* given by

$$P(T_r = n) = \binom{n-1}{r-1} p^r q^{n-r}.$$

However, a simpler method is available. We have

$$T_r = T^{(1)} + T^{(2)} + \cdots + T^{(r)}$$

where $T^{(k)}$ is the wait between the $(k-1)^{th}$ and the k^{th} Success. This is no different probabilistically from the wait T for the first Success. Thus these r.v. are independent and all have the same geometric distribution as T. Hence, the means and variances add, and we have

Theorem 2 *The Pascal distribution has mean r/p and variance rq/p^2.*

*Coupon collecting.

Consider the case where one is collecting coupons or prizes (for example. from cereal boxes), with the goal of collecting a full set of n coupons. We shall suppose that each box is equally likely to contain any of the n coupons of the set. the wait T for the n^{th} coupon (measured in units of the number of boxes purchased) is equal to

$$T = T_1 + T_2 + \cdots + T_n$$

where T_k is the wait between the $(k-1)^{st}$ and the k^{th} coupon. After $k-1$ coupons have been collected, there are n possible coupons in the next box, of which $n-k+1$ are ones still needed. Thus, T_k has geometric distribution with probability of Success

$$p_k = \frac{n-k+1}{n}$$

and hence

$$E(T_k) = \frac{n}{n-k+1}$$

Thus,

$$\begin{aligned} E(T) &= E(T_1) + E(T_2) + \cdots + E(T_n) \\ &= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{2} + \frac{1}{n} = n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + 1\right) \\ &\approx n \log n. \end{aligned}$$

For example, if there are six coupons to collect, the expected number of boxes required is

$$E(T) = 6\left(\frac{1}{6} + \frac{1}{5} + \frac{1}{4} + \frac{1}{3} + \frac{1}{2} + 1\right) = 14.7.\blacksquare$$

*Buffon's coin tossing data.

Georges-Louis LeClerc, Comte de Buffon (1707-1788) was a prominent naturalist, considered by some to be a forerunner of Darwin. In 1777, he conducted the following experiment. A coin was tossed repeatedly until the first Head appeared, and the number of tosses was needed recorded. There were a total of 2048 measurements, and a grand total of 4040 tosses.An estimate of the probability of Heads is thus

$$\frac{2048}{4040} = 0.507.$$

The complete data are shown in the following table. The number m_k of runs of length k is recorded in the first row.

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**≥10**	**Total**
m_k	1061	494	232	137	56	29	25	8	6	0	2048
ε_k	1024	512	256	128	64	32	16	8	4	4	2048

We would expect a geometric distribution with probability of Heads $p = 1/2$. Thus the expected number of runs of length k is $\varepsilon_k = 2048/2^k$. This is recorded in the second row.

The Markov Property.

Suppose that we have experienced a run of n Failures. What is the probability that we experience an additional m Failures before the first Success? Since the trials are independent, so that *the past has no effect on the future*, this probability is the same as the probability of m successive Failures. It is as if the sequence of trials were "*starting afresh*", as is sometimes said. We describe this by saying that *the wait for the first Success is memoryless*.

To put this down numerically, what we are saying is that

$$P\left(T > n + m \mid T > n\right) = P(T > m).$$

This is called the *Markov Property*.

To see that the Geometric distribution actually has this property, note first that

$$P(T > n) = q^n.$$

Thus

$$\begin{aligned} P\left(T > n + m \mid T > n\right) &= \frac{P\left(T > n + m \;\&\; T > n\right)}{P\left(T > n\right)} = \frac{P\left(T > n + m\right)}{P\left(T > n\right)} \\ &= \frac{q^{n+m}}{q^n} = q^m = P(T > m).\blacksquare \end{aligned}$$

The geometric distribution is the *only distribution on the positive integers with the Markov Property*. (See problem 2.)

6.3 The Poisson Distribution.

The Poisson distribution is one of the most important distributions of probability theory. It reflects randomness of a certain type. First introduced by *C. Poisson* in 1838, it was rather neglected until the 1898 book "*The Law of Small Numbers*" of *von Bortkewicz*. There are several ways of coming up with the Poisson distribution, and you will need to understand them all. The first is Poisson's way, as an approximation to the Binomial distribution.

1. Poisson's Approximation to the Binomial.

Consider a lottery, with a probability of winning equal to $1/1,000,000$. If $500,000$ people play, what is the probability that there are multiple winners?

If each player's success or failure is regarded as an independent trial, then the number N of winners has a Binomial distribution with $p = 1/10^6$ and $n = 500,000$. The expected number of winners is $np = \frac{1}{2}$. In situations like this, where the *number of trials is large, and the probability of Success is small, with a moderate mean*, Poisson derived a convenient approximation to the Binomial probabilities.

To obtain this, we *let the mean $\lambda = np$ remain fixed, and let the number n of trials tend to infinity*. It then automatic that the probability of Success $p = \lambda/n$ will tend to zero. We have then have, for k fixed,

$$\begin{aligned} P(N &= k) = \binom{n}{k} p^k q^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \frac{n(n-1)\cdots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^{-k} \\ &= \left[\frac{n}{n}\frac{n-1}{n}\cdots\frac{(n-k+1)}{n}\left(1-\frac{\lambda}{n}\right)^{-k}\right]\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \end{aligned}$$

The quantity in brackets consists of $2k$ factors all of which tend to 1. Since k is fixed, this factor tends to 1. The second factor is independent of n, and the last tends to $e^{-\lambda}$ by the compound interest formula

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n.$$

Thus the distribution of N tends to the *Poisson distribution*

$$P(N=k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

The solution to the Lottery Problem is therefore approximately

$$\begin{aligned} P(N &> 1) = 1 - P(N=0) - P(N=1) \\ &= 1 - e^{-\lambda} - \lambda e^{-\lambda} = 1 - \frac{3}{2}e^{-1/2} = 0.09. \end{aligned}$$

The *mean and variance of the Poisson distribution* can be obtained as the limit of the mean and variance of the approximating Binomial distribution. This gives $\mu = \lambda$, and

$$\sigma^2 = \lim npq = \lim \lambda \left(1 - \frac{\lambda}{n}\right) = \lambda.$$

Thus, *the mean and variance of the Poisson distribution are both equal ti* λ.

The same result can also be obtained by computation with series. (See problem 1.)

Example 1 . A book of 200 pages contains 150 misprints. What is the probability that there are no misprints on page 25 ?

Solution. We shall assume that the 150 misprints are independently distributed throughout the book, so that the probability that the n^{th} misprint appears on any given page is $1/200$, and the appearance of a misprint on a page is independent of whether or not any of the other other misprints appear there. The number of misprints on page 25 therefore has a Binomial distribution with $n = 150$ and $p = 1/200$. This is approximately Poisson with $\lambda = np = 3/4$. Thus, approximately,

$$P(N = 0) = e^{-\lambda} = e^{-3/4} = 0.472.\blacksquare$$

Example 2. If is p small, how many trials are needed to have probability at least $1/2$ of at least one Sucess?

Solution. By Poisson's approximation, we need,

$$P(\text{no Success}) = P(N = 0) \simeq e^{-np} < \frac{1}{2}$$

or

$$n > \frac{\log 2}{p} \cong \frac{0.69}{p}.\blacksquare$$

This is the same result that was obtained by an approximation in Section 4.1

*Death by Mule.

As physicists have their classic experiments and mathematicians their classic theorems, so statisticians have their classic data sets. One such is the Mule Kick Data of von Bortkewicz. In his 1898 book *The Law of Small Numbers*, he considered the number of deaths from mule and horse kicks in 10 Prussian Army Corps over a period of 20 years, a total of 200 corps years of data points. Besides their value as a caution against too romantic a view of the pre-automotive era, these data illustrate the Poisson distribution.

The table below gives, in the first and second columns, the frequencies of the number of such deaths. For example, 109 times a corps experienced no deaths that year. As will be seen from the table, the total number of deaths was 122, for an average of $\lambda = 122/200 = 0.610$. Assuming that the number of deaths in a corps has a Poisson distribution with mean

$\lambda = 0.610$, we may compute the Poisson probabilities of n deaths for $n = 0, 1, 2, 3, 4...$ The results are listed in the third column. For example,

$$P(n = 2; \lambda = 0.61) = e^{-0.61} \frac{(0.61)^2}{2!} = 0.101.$$

The expected number of occurrences is this number times 200, which appears in the last column. It seems to match fairly well with the observed results.

# Deaths	**Frequency**	$P(n; \lambda = 0.61)$	**Expected Occurrences**
0	109	0.543	108.6
1	65	0.331	66.3
2	22	0.101	20.2
3	3	0.021	4.2
4	1	0.003	0.6

As a check, since the mean and variance of the Poisson distribution are equal, we should have $S^2 \simeq \bar{X}$. Tthe sample variance is

$$S^2 = \left(\frac{1}{n-1}\right) \sum_i \left(X_i - \bar{X}\right)^2 = 0.611.$$

which is good agreement. Of course, the Statisticians have much better ways of measuring goodness of fit, such as the χ^2-test, but this is presently beyond our scope.

We now want to consider some more complicated models in which the Poisson distribution occurs. The Poisson distribution models a certain kind of randomness, and one needs to develop a feeling for when a r.v. might have a Poisson distribution.

2. Calls to a Switchboard.

We shall consider incoming calls to a switchboard. Let $N(t)$ be the number of calls arriving in the interval $[0, t]$. To approximate the distribution of $N(t)$, divide the interval $[0, t]$ into n equal intervals each of length t/n. Let C_k be the event that there is a call in the k^{th} interval. If the interval is short enough, we may neglect the probability that there is more than one call in any interval. We shall assume

(a.) that the events C_k are independent, and.

(b.) that for short time intervals, the probability of C_k is approximately *proportional to the length of the time interval* C_k, and is

(c.) independent of k, which is to say that it is *independent of time.*

This means that

$$p_n = P(C_k) \simeq a\frac{t}{n}$$

where a is a parameter independent of k. The parameter a has units of *calls per unit time*.

If Success is regarded as a call coming in, then in this approximation, we have a sequence of Bernoulli trials, the k^{th} trial being the existence or nonexistence of a call in the interval C_k. The number of calls $N(t)$ is thus the number of Successes in n trials, and therefore has a Binomial distribution, with n trials and probability $p = at/n$. The mean $np = at$ is fixed and independent of n. As n increases, the number of intervals increases, and by Poisson's approximation, the distribution approaches a Poisson distribution with mean at.

$$P(N(t) = k) = \frac{(at)^k}{k!} e^{-at}.$$

The mean of $N(t)$ is therefore $\lambda = at$, and hence the parameter a has the interpretation of the *mean number of calls per unit time*.

The variance of $N(t)$ is $\sigma^2 = at$, since the mean and variance of the Poisson distribution are equal.

The fact that we are speaking about phone calls is not material. The calls could be any sort of events taking place randomly in time. The family of r.v. $N(t)$ is called a Poisson process. It will be discussed more fully in Chapter 10.

The Wait for the First Call.

Let T be the wait for the first call. We have then

$$P(T > t) = P(N(t) = 0) = e^{-at}.$$

Alternatively, T may be thought of as the limit of geometric waiting times

$$P(T > t) = \lim q_n^n = \lim \left(1 - \frac{at}{n}\right)^n = e^{-at}.$$

6.4 Poisson Ensembles of Points.

A Poisson ensemble of points is a model of *an infinite number of points distributed randomly in an infinite space with a finite density*. This is constructed in the following way.

Suppose first that n points are "*independently and randomly distributed*" throughout the interval $[0, L]$. Let $I = [a, b]$ be a subinterval of $[0, L]$ of length $|I| = b - a$. What is the probability that I is empty; i.e. contains none of the points. More generally, what is the distribution of the number $N(I)$ of points in I ?

In order to discuss this, we need to assign a precise meaning to the phrase "*randomly distributed*".

Consider first the case $n = 1$ of a single point X. What might we mean by saying that its position is random ? One interpretation which comes to mind is that if I and I' are two

subintervals $[0, L]$ of the equal length, then the point is as likely to be in one as it is in the other. This implies that

(1.) *The probability of the point being in I is*

$$p(I) = P(X \in I) = \frac{|I|}{L}.$$

Now suppose that we have n points $X_1, X_2, \ldots, X_n$. A reasonable interpretation of "independently' distributed is that:

(2.) *The events*

$$\{X_1 \in I\}, \{X_2 \in I\}, \ldots, \{X_n \in I\}$$

are an independent set of events.

Given these assumptions, we have a sequence of Bernoulli trials, with the k^{th} Success being that the k^{th} point is in the interval I.

Hence, *the number of points $N(I)$ in I has the Binomial distribution $Bin(n, p)$, where $p = |I|/L$. In particular, the mean of $N(I)$* is

$$\lambda = np = |I| \frac{n}{L} = |I| d$$

where

$$d = n/L$$

is the density of points - that is, the number of points per unit length.

Now suppose that length L of the interval and the number n of points both increase to infinity in such a way that the density $d = n/L$ remains constant. In this way, we obtain a model of an infinite number of points distributed randomly on an infinite interval with a finite density.

In this limit, by Poisson's approximation, *$N(I)$ has a Poisson distribution with $\lambda = |I| d$.*

One can do the same thing for the whole line by considering the interval $[-L/2, L/2]$.

Example 1. One hundred turtle nests are distributed randomly along a beach 10 km long. What is the probability of at least three nests in a 100 m stretch of beach?

Solution. The density is $d = 100/10 = 10$ nests per km. The number N of nests in a 0.1 km stretch is Poisson with mean $\lambda = 10d = 1$. Thus

$$P(n \geq 3) = 1 - e^{-1}(1 + 1 + \frac{1}{2}) = 1 - \frac{2.5}{e} = 0.0803.\blacksquare$$

*Independence.

Theorem 3. *In the limit, if I and J are disjoint intervals, $N(I)$ and $N(J)$ are independent.*

This statement is not obvious, since f*or finite n, $N(I)$* and $N(J)$ *are clearly not independent*. For if, to take an extreme case, all n points are in I, none can appear in J. Nevertheless, in the limit, $N(I)$ and $N(J)$ *are* independent for disjoint I and J.

Proof. For the proof, observe that if $G = [0, L] \setminus (I \cup J)$ is the complement of $I \cup J$ in $[0, L]$, then $N(I)$, $N(J)$ and $N(G)$ have a Multinomial distribution. Thus

$$\begin{aligned} P(N(I) &= k, N(J) = l) \\ &= P(N(I) = k, N(J) = l, N(G) = n - k - l) = \\ &\quad \frac{n!}{k!l!(n-k-l)!} \left(\frac{|I|}{L}\right)^k \left(\frac{|J|}{L}\right)^l \left(1 - \frac{|I| + |J|}{L}\right)^{n-k-l} \end{aligned}$$

Setting $d = n/L$, we get

$$\begin{aligned} P(N(I) &= k, N(J) = l) \\ &= \frac{n!}{k!l!(n-k-l)!} \left(\frac{d\,|I|}{n}\right)^k \left(\frac{d\,|J|}{n}\right)^l \left(1 - \frac{d\,|I| + d\,|J|}{n}\right)^{n-k-l} \\ &= \frac{(d\,|I|)^k\,(d\,|J|)^l}{k!l!} \left(1 - \frac{d\,|I| + d\,|J|}{n}\right)^n \\ &\quad \times \frac{n(n-1)(n-k-l+1)}{n^{k+L}} \left(1 - \frac{d\,|I| + d\,|J|}{n}\right)^{-(k+l)} \end{aligned}$$

The last two factors tend to 1, and so in the limit,

$$\begin{aligned} P(N(I) &= k, N(J) = l) = \frac{(d\,|I|)^k\,(d\,|J|)^l}{k!l!} \lim_{n\to\infty} \left(1 - \frac{d\,|I| + d\,|J|}{n}\right)^n \\ &= \frac{(d\,|I|)^k\,(d\,|J|)^l}{k!l!} e^{-d|I|+d|J|} = \frac{(d\,|I|)^k}{k!} e^{-d|I|} \frac{(d\,|J|)^l}{l!} e^{-d|J|} \\ &= P(N(I) = k) P(N(J) = l). \square \end{aligned}$$

Poisson Point Ensembles in Higher dimensions.

Poisson point ensembles are not limited to points on the line, but also apply to points in space or in the plane.

For let G be a region in two dimensional space of area $|G|$, which is contained in a large region - for example, a rectangle - of area A. Suppose that n points are distributed independently and randomly with respect area, so that for any such point X_i

$$p = P(X_i \in G) = \frac{|G|}{A}.$$

The events$\{X_i \in G\}$ are independent, so the number of points in G has the distribution $Bin(n,p)$. Now let A and n tend to infinity, with the ***density*** $d = n/A$ ***fixed***. Then the distribution of $N(G)$ is $Bin(n, d\,|G|)$ which tends to the Poisson distribution with mean $\lambda = d\,|G|$ as A tends to infinity.

Theorem 3 holds in this case as well.

The results is the same if area in the plane is replaced by the volume in space.

According to Feller, "Stars in space, raisins in cakes, weed seeds among grass seeds, flaws in materials, animal litters in fields are all distributed according to the Poisson law." Having no data on animal litters, we will content ourselves with Feller's example of bomb strikes in London during World War II.

Example 2. An area of 144 square kilometers in south London was partitioned into $N = 576$ cells of area $1/4$ square kilometers each, and the number of "flying bomb" strikes in each area recorded. The following chart shows the number $N(k)$ of areas with exactly k hits.

$\mathbf{k}$	0	1	2	3	4	≥ 5
$\mathbf{N(k)}$	229	211	93	35	7	1
Predicted	226.74	211.39	98.54	30.62	7.14	1.57

The total ***number of hits*** was 537. If the bombs are falling randomly and independently, the strikes will form a Poisson point ensemble of density

$$d = 537/576 = 0.9323$$

strikes per cell. The number of strikes in a cell I is therefore Poisson with mean $\lambda = d$.

Since the cells are independent, we have, by Theorem 3 above, $N = 576$ independent measurements of a Poisson r.v. X with mean $d = 0.9323$. The expected number of occurrences of the event $\{X = k\}$ is therefore

$$Ne^{-d}\frac{d^k}{k!}.$$

third line of the table. The agreement seems quite good, as is confirmed by a χ^2-test. ■

Example 3. An interesting application of a Poisson point process is the composition *Atlas Eclipticalis* of John Cage. The *Atlas Eclipticalis* is a collection of star charts. Cage took transparencies lined with music staffs, slapped them down on star charts and marked off notes wherever big stars appeared. The present author was privileged to experience a performance of this work in the early 1960's, conducted - if that is the word - by Cage himself. One was rather listening to an auditory representation of a Poisson process.

Remarks.

This seems to be a good time to discuss a point that is very important to anyone who intends to apply a probabilistic model, such as the Poisson distribution, to some phenomenon

in the real world. The point is this: *It is very important to pay strict attention to the assumptions that go into the derivation of the formula that one intends to use*.

This is not a question of mathematical rigor, but goes to the validity and ultimate success of the application of the model to the problem one has in mind. For if the assumptions of the model are not satisfied in the case at hand, the model is most unlikely to produce valid results. Moreover, if one finds that the model does not work, some one of the assumptions used in its derivation must not be true. If you know what these assumptions are, you have a key to finding what you must change to get a model that will work. If you do not know what they are, you will be at a loss as to what to do.

6.5 Infinite Expectations.

As we pointed out in section 1, the sum for the mean of a r.v. with an infinite discrete distribution is an *infinite series*, which may not converge. When it does not, the *r.v. does not an have expectation.*

We will now consider a few naturally occurring r.v. that do not have expectations, and to discuss briefly what is signified by this fact.

1. Polya's urn model.

An urn contains 1 Red and 1 White ball. Draw, replace and add one ball of the color drawn. Repeat. Let T be the *wait for the first White ball*. What are the mean and distribution of T ? We have, as in Section 3.1,

$$P(T > n) = \frac{1}{2} \cdot \frac{2}{3} \cdots \frac{n}{n+1} = \frac{1}{n+1}$$

and hence

$$P(T = n) = P(T > n-1) - P(T > n) = \frac{1}{2} \cdot \frac{2}{3} \cdots = \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}.$$

for $n = 1, 2, \ldots$.

The expectation is therefore

$$E(T) = \sum_{n=1}^{\infty} n \frac{1}{n(n+1)} = \sum_{n=1}^{\infty} \frac{1}{(n+1)} = \infty.$$

What is the significance of an infinite expectation? In Polya's model after a long run of Reds, White becomes very unlikely, and increasingly so. This results in an occasional very long wait. So one is averaging some really large numbers into the sample mean, and the average length blows up as number of trials tends to infinity.

One might think that since the "wait" for the first White ball is infinite, it means that there is a good chance that the game might go on forever, that is that a White will *never* show up. *This is not true.* The probability that a White ball will never show up is less than the probability that it fails to show up on the first n trials. That is,

$$P(\text{White never shows up}) \leq P(T > n) = \frac{1}{n+1}.$$

But this tends to zero as $n \to \infty$, so

$$P(\text{White never shows up}) = 0.$$

The behavior of the sample mean $\bar{X}_n$ for r.v. with infinite expectation is a too technical subject to enter into here**.**

Although the arithmetic mean is problematic here, the median is well behaved. For since

$$P(T = 1) = P(T \geq 2) = 1/2$$

the median is $3/2$.

*2. The Petersburg Game.

The following game was invented by Nicholas Bernoulli, and was first discussed by his cousin Daniel Bernoulli.

A single trial of the game consists of tossing a fair coin repeatedly until Heads appears. If Heads appears on the n^{th} toss, the player wins $\$\ 2^n$ dollars.

What is a fair entry fee for this game? This is usually taken to be equal to the expectation of the players winnings X. The probability of winning 2^n dollars is $1/2^n$, i.e.

$$P(X = 2^n) = 1/2^n.$$

Therefore

$$E(X) = \sum_{n=1}^{\infty} 2^n \cdot \frac{1}{2^n} = \sum_{n=1}^{\infty} 1 = \infty.$$

Thus, the "*fair entry fee*" is infinite, and it should be a bargain to pay $\$\ 1,000,000$ to play! However, no one would actually pay this amount to play this game.

The game has been widely discussed. Daniel Bernoulli introduced a notion of the *utility* of money into his discussion. Others have pointed out that no house would actually have the infinite resources necessary to offer such a game.

We refer the reader to the book of Feller for a serious discussion of "fair" games. (Feller, I , p. 251 - 3.) Note that on Buffon's data in section 6.2, the player would have won $\$10,057$ dollars.in 2084 games, or about $\$4.83$ per game.

*3. The Persistence of Bad Luck.

Everyone knows that no matter which line one gets in at the bank or grocery, it is the wrong one. We close this chapter with a mathematical meditation on this subject from the book of Feller.

Let X_0 be your wait, and $X_1, X_2, ..., X_n, ...$ the waits of various other people. *We shall assume that $X_0, X_1, ..., X_n, ...$ are independent r.v. with the same continuous distribution*. Let N be the wait for someone to have worse luck than you; that is, to have a longer wait. Then $N \geq n$, iff the largest number in the list

$$X_0, X_1, ..., X_{n-1}.$$

occurs at the first spot. By symmetry, the largest value occurs at each spot with equal probability, so

$$P(N \geq n) = \frac{1}{n}.$$

But then

$$P(N = n) = P(N \geq n) - P(N \geq n+1) = \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n\,(n+1)}.$$

he probability that the wait for the first head and so

$$E(N) = \sum_{n=1}^{\infty} nP(N = n) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty.$$

The expected wait for someone to have worse luck than you is infinite.

The reader may wish to reflect on the significance of this calculation, and perhaps, in the words of the poet, "*depart a sadder but a wiser man*."

6.6 Problems.

(1.) Compute the mean and variance of the Pascal distribution directly by summing the appropriate series.

(2.) Prove that the geometric distribution is the only distribution on the positive integers that has the Markov property.

(3.) If X and Y are independent and have geometric distribution with probability p, find the probability that $Y = X$.

(4.) (*Logarithmic series distribution*.) Find mean and variance of the distribution ,

$$p_n = \frac{1}{\log(1/q)} \frac{p^n}{n} \qquad n \geq 1.$$

where $0 < p < 1$ and $q = 1 - p$.

Section 6.6 Problems.

(5.) Coin A is a fair coin. Coin B is biased with the probability of heads equal to $\frac{2}{3}$. The coins are tossed alternately, starting with coin A, until a Head is obtained. Let T be the number of tosses needed to obtain the first Head. What are the distribution and expectation of T ?

(6.) A student of probability has little talent but works hard. When he attacks a problem he has probability 0.01 of solving it. Since he never learns from experience, all his attempts are independent of one another. If he tries 250 problems in a semester, what is his probability of getting at least three right?

(7.) The probability of winning a certain lottery is 1 in $10,000,000$. If $2,000,000$ people play, what is the probability that there is more than one winner?

(8.) A galaxy contains two trillion (2×10^{12}) stars. Each star has one chance in 5 trillion (5×10^{12}) to have a planetary system that supports Life. What is the probability that there is Life in the galaxy?

(9.) If X is $Poi(\lambda)$, show that

$$E(X^n) = \lambda E\left[(X+1)^{n-1}\right]$$

Use this to compute $E(X^3)$.

(10.) Compute the mean and variance of the Poisson distribution directly by evaluating the appropriate infinite series.

(11.) Rain is falling at a rate of three drops per square foot per minute. A board of area 0.25 square feet is placed outside.

(a.) Find the probability that after five minutes the board has been hit by no more than two drops.

(b.) What is the distribution of the waiting time T_3 for the third drop to hit the board?

(12.) 200 turtles nest at random on a 20 mile stretch of beach. Peach, a member of the turtle patrol starts at one end of the beach and drives north for quarter of a mile.

(a.) What is the probability that she finds no nests?

(b.) What is the probability she finds no nests if she begins at a random point along the beach instead of at the end?

(c.) If she finds no nest in the first quarter of mile, what is the probability she finds none in the second quarter of a mile?

(13.) Flying insects are distributed along a highway as a Poisson point ensemble with a density of a bugs per mile. Starting with a clean windshield, a motorist drives at a speed of v miles per hour until his cell phone rings, at which time he stops. If his wait for a cell call is exponential with mean $1/b$, what is the expected number of spots on his windshield when he stops?

(14.) A baker is making 1000 almond cookies. How many almonds are needed in order that only 1 in 100 cookies will have no almond?

(15.) The following data for the number of yeast cells in each of 400 unit volumes of a hemacytometer were obtained by the famous statistician and Guiness brewer W S Gossett.

cell count	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
# times observed	0	20	43	53	86	70	54	37	18	10	5	2	2	0

(a.) Compute the sample mean and compare with predicted Poisson probabilities.

(b.) The mean and variance of the Poisson distribution are equal. Compare the sample mean and variance for these data.

(16.) The following data give the statistics for number of goals scored in the 420 NHL hocky games in the 1966-7 season.

Goals scored	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
# of games	29	71	82	89	65	45	24	7	4	1	3

Compute the sample mean and variance. Compare with predicted Poisson probabilitiese same mean..(*Ans.* $\bar{X} - 2.98$; $S^2 = 3.53$ JSH)

(17.) (*The Cow Pie problem.*) According to Feller, "Stars in space, raisins in cakes, weed seeds among grass seeds, flaws in materials, animal litters in fields are all distributed according to the Poisson law." To escape his pursuers, a fugitive must cross a pasture in the dark of night, in which circular cow pies of diameter 8 inches are randomly distributed, with a density of one in every 40 square feet. He is wearing square-toed boots one foot long and four inches wide, and must take 20 steps to cross the field. What is the probability that in crossing he will step in a cow pie?

(18.) In the Petersburg problem, suppose that the House has only sufficient capital $B = 2^N$ to offer N games.

(a.) Compute the expected winnings in this case. Assume that if no Heads has appeared in N tosses, the player wins the entire capital of the bank.

(b.) What is the probability that the player comes out ahead if he pays this entry fee?

Chapter 7
Continuous Random Variables.

7.1 Continuous Random Variables.

At the end of the last chapter, we encountered two random variables which could take on any values in some interval of the real line. The wait T for the first call can be any positive number. The location of a random point on an interval can be any point in that interval.

Here are a few additional examples of continuous random variables:

1. The daily amount of rainfall at a given location.
2. The angle of the pointer on a spinner makes with the horizontal.
3. The chest measurements of Scottish soldiers.
4. The heights of inductees into the French army.

Examples 3 and 4 refer to classic data sets studied by the pioneer statistician Quetelet in the 19th century.

Some new techniques are needed to describe such random variables. To see the problem, recall that for the *wait T for the first call*, we found that

$$P(T > t) = e^{-at} \text{ for } t > 0.$$

Thus, for any fixed number c, we have

$$\begin{aligned} 0 &\leq P(T = 0) \leq P((c - h < T \leq c + h) - P) \\ &= P(T > c + h) - P(T > c - h) \\ &= e^{-a(c+h)} - e^{-a(c-h)} \to 0 \quad \text{as } h \to 0. \end{aligned}$$

Hence,

$$P(T = c) = 0.$$

The probability that T is *actually equal* to any fixed number c is zero!

How can the distribution of such a r.v. be described ? Clearly, we cannot do this as we did for discrete variables simply by giving $P(T = c)$ for every c, since these numbers are all just zero.

The general method, which *applies to any random variable X without exception*, is to give, for every interval $(a, b]$, the probability

$$P(a < X \leq b)$$

that X lies in that interval.

For example, for the *wait T for the first call*, we found that

$$P(T > t) = e^{-at} \text{ for } t > 0.$$

So we have

$$P(t < T \leq s) = P(T > t) - P(T > s) = e^{-at} - e^{-as}.$$

Similarly, for the position of a random point in the interval $[0, L]$, we found that for $0 \leq a < b \leq L$,

$$P(a < X \leq b) = \frac{b-a}{L}.$$

Independence of two r.v. can then be defined in terms of the interval probabilities.

Definition 1. *(Independent random variables.)* We say that two r.v. X and Y are *independent* iff the events $\{a < X \leq b\}$ and $\{c < Y \leq d\}$ are independent events for every pair of intervals $(a, b]$ and $(c, d]$.

This is equivalent to saying that

$$P(a < X \leq b \ \& \ c < Y \leq d) = P(a < X \leq b)P(c < Y \leq d)$$

Similarly, $X_1, X_2, \ldots, X_n$ is an *independent set of random variables.* iff the events

$$\{a_1 < X_1 \leq b_1\}, \{a_2 < X_2 \leq b_2\}, \cdots, \{a_n < X_n \leq b_n\}$$

is an independent set of events for every set of intervals $(a_1, b_1], (a_2, b_2], ..., (a_n, b_n]$.

7.2 Density Functions.

The distribution of a continuous random variable can often be described by means of a *density function*. A random variable X is said to have the *density* $f(x)$ if

$$P(a < X \leq b) = \int_a^b f(x)dx.$$

In order for a function to be the density of a r.v. X, it must have two basic properties:

(1.) It must be non-negative, since otherwise $P(a < X \leq b)$ could turn out to be negative; and

(2.) It must integrate to 1, since we want

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

Any such function can be used as a density.

Definition 2. (*Density Function.*) *A function $f(x)$ is a density function iff*

(a.) $f(x) \geq 0$ and

(b.) $\int_{-\infty}^{\infty} f(x)dx = 1$.

If X has the density $f(x)$, then

$$P(X = c) = \int_c^c f(x)dx = 0.$$

Hence, we have:

Corollary 1. *If X has a density, then $P(X = c) = 0$ for all c.*

In pictures, the probability that X lies between a and b is the area under the graph of $f(x)$ between the limits a and b.

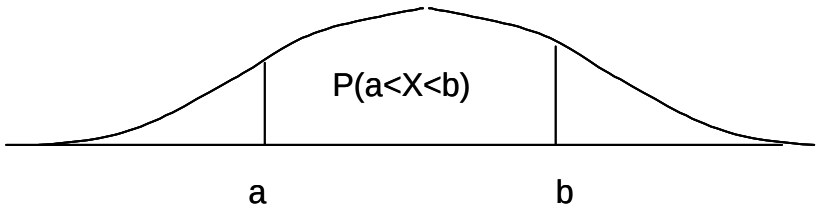

Figure 2.1. *A Density Function*.

The following are typical examples of densities.

Example 1. (*A simple Beta density*) Consider the density

$$\begin{aligned} f(x) &= Cx\,(1-x) \qquad 0 \leq x \leq 1 \\ &= 0 \qquad\qquad\quad \text{otherwise.} \end{aligned}$$

The normalization constant C can be found by requiring that

$$\int_0^1 f(x)dx = C\int_0^1 x\,(1-x)\,dx = \frac{C}{6} = 1.$$

Taking $C = 6$ gives a density. We may then compute, for example,

$$P(\frac{1}{4} < X < \frac{3}{4}) = \int_{1/4}^{3/4} 6x\,(1-x)\,dx = \frac{11}{16}.$$

The *range* of X is the interval $(0, 1)$. ■

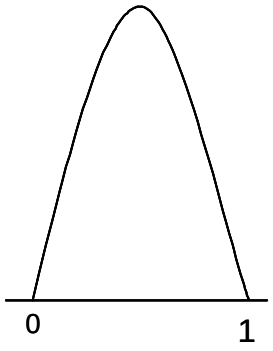

Figure 2.2. *The Beta density* $6x(1-x)$.

Example 2. (*Exponential Density*) For the *wait* T *for the first call*, we found that

$$P(t < T \leq s) = e^{-at} - e^{-as} = \int_t^s ae^{-ax}dx.$$

The random variable T therefore has the density

$$\begin{aligned} f(x) &= ae^{-ax} \qquad \text{for } x \geq 0 \\ &= 0 \qquad\quad \text{for } x < 0. \end{aligned}$$

This very important density is known as the *exponential density*.■

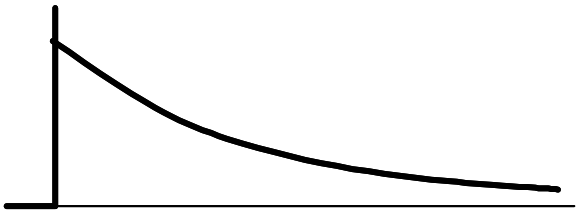

Figure 2.3 *The Exponential density.*

Example 3. (*Uniform Density*) Consider next the *position* X *of a random point* on the line segment $[a, b]$. We saw that the probability that X lies in some interval subinterval I of $[a, b]$ is proportional to the length of I. That is, if $I = [c, d] \subset [a, b]$, we have

$$\begin{aligned} P(X \in I) &= P(c < X < d) = \frac{d-c}{b-a} \\ &= \int_c^d \frac{1}{b-a}dx. \end{aligned}$$

where $a \leq c < d \leq b$. The random variable X therefore has the density

$$\begin{aligned} f(x) &= \frac{1}{b-a} \qquad a \leq x \leq b \\ &= 0 \qquad\qquad\quad \text{otherwise.} \end{aligned}$$

This density is known as the *uniform density*. It is constant on the interval $[a.b]$, and zero outside of it.■

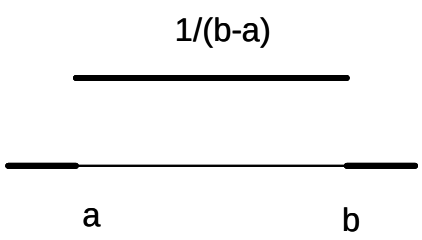

Figure 2.4. *The uniform density.*

Example 4. (*Standard Normal Density.*) The Standard Normal or *Gaussian* density is defined as

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

As shown in Appendix A, the constant $1/\sqrt{2\pi}$ in front makes the integral equal to 1. As we shall see, this is probably the most important density function in Probability and certainly the most important in Statistics. The normal density curve is sometimes referred to as the "*Bell Curve*" because of its shape.

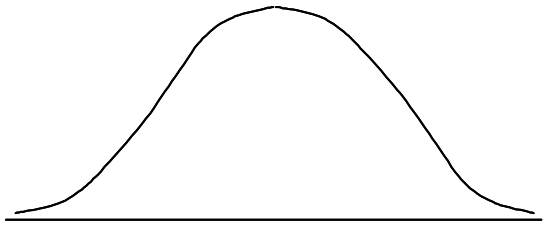

Figure 2.5. *The normal density.*

Remark: The density of a r.v. X is always *zero off the range of* X. Thus, the exponential density is *not just* ae^{-ax}; it is ae^{-ax} for $x \geq 0$ and zero for $x < 0$. Similarly, the Beta density of Example 1 is zero off the range $[0, 1]$ of X, and the uniform density of Example 3 is zero off the range $[a, b]$. This needs to be borne in mind when computing with densities.

Thus, whenever confronted with a r.v. X, the *first question one should always ask is: "What values can this r.v. take on ?"*; that is, what is the range of X ?

Gamma and Beta densities.

The Gamma density $\Gamma(p, a)$ is a generalization of the exponential density defined by

$$f(x) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax} \qquad x > 0.$$

where $a > 0, p > 0$ for convergence. The factor $\Gamma(p)$ is defined to make the integral equal to 1.

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx.$$

Its properties are discussed in Appendix A 2. In particular, $\Gamma(n+1) = n!$ if n is a positive integer.

The Beta density $Beta(a, b)$ is

$$f(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} \qquad 0 < x < 1$$

where

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

is the Beta function.

7.3 Expectation and Variance .

Expectation.

The expectation of a discrete random variable was defined by

$$E(X) = \sum_i x_i p_i.$$

For a r.v. X with density $f(x)$, the *expectation* is obtained simply by *replacing the discrete sum by an integral*:

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

Example 1. The expectation of the Beta density

$$f(x) = 6x(1-x) \qquad 0 \le x \le 1$$

is

$$E(X) = 6 \int_0^1 x \cdot x (1-x) dx = 6 \int_0^1 \left(x^2 - x^3\right) dx = \frac{1}{2}. \blacksquare$$

Example 2. For the *exponential density*

$$f(x) = ae^{-ax} \qquad 0 \le x < \infty$$

we have, with $s = ax$,

$$E(X) = \int_0^\infty xae^{-ax}dx = \frac{1}{a}\int_0^\infty se^{-s}ds = \frac{1}{a}\Gamma(2) = \frac{1}{a}.\blacksquare$$

Example 3. For the *Gamma density* $\Gamma(p, a)$

$$f(x) = \frac{a^p}{\Gamma(p)}x^{p-1}e^{-ax} \qquad x > 0.$$

we have, with $s = ax$,

$$E(X) = \frac{a^p}{\Gamma(p)}\int_0^\infty xx^{p-1}e^{-ax}dx = \frac{1}{a}\int_0^\infty s^p e^{-s}ds = \frac{\Gamma(p+1)}{\Gamma(p)}\frac{1}{a} = \frac{p}{a^2}.\blacksquare$$

Example 4. The expectation of the *uniform density* on $[a, b]$, is

$$E(X) = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{a+b}{2}.\blacksquare$$

Example 5. For the *Standard Normal density*

$$E(X) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\, e^{-x^2/2}dx = 0$$

since this is the integral of an odd function over symmetric limits.$\blacksquare$

The expectation has the same basic properties as before.

Theorem 1. *(Properties of Expectation.)*

(a.) $E(1) = 1$.

(b.) If $X \ge 0$, *then* $E(X) \ge 0$.

(c.) $E(cX) = cE(X)$.

(d.) $E(X + Y) = E(X) + E(Y)$.

(e.) If X *and* Y *are independent, then* $E(XY) = E(X)E(Y)$.

The proofs of (a.) - (c.) are the same as for discrete r.v., since they involve only the basic properties of expectation, and a little algebra. Parts (d.) and (e.) are more difficult, and will be omitted.

Variance.

The *variance* is defined as before:

$$\sigma^2(X) = var(X) = E(X-\mu)^2$$

where $\mu = E(X)$. It has the same properties as before.

Theorem 2. *(Properties of Variance.)*

(a.) $var(X) \geq 0$.

(b.) $var(cX) = c^2 var(X)$.

(c.) $\sigma^2 = \mu_2 - \mu^2$, *where* $\mu_2 = E(X^2)$ *is the second moment.*

(d.) If X *and* Y *are independent, then* $var(X+Y) = var(X) + var(Y)$.

As in the discrete case, the *Law of the Unconscious Statistician* is useful for computing variances. For random variables with densities, it takes the following form.

Theorem 3. (*Law of the Unconscious Statistician.*) *If* X *has density* $f(x)$*, then*

$$E(\phi(X)) = \int_{-\infty}^{\infty} \phi(x) f(x) dx.$$

Example 6. For the simple Beta density

$$f(x) = 6x\,(1-x) \qquad 0 \leq x \leq 1$$

the second moment is

$$\mu_2 = E(X^2) = 6\int_0^1 x^2 \cdot x\,(1-x)\,dx = 6\int_0^1 \left(x^3 - x^4\right) dx = \frac{3}{10}.$$

and so the variance is

$$\sigma^2 = \mu_2 - \mu^2 = \frac{3}{10} - \left(\frac{1}{2}\right)^2 = 0.05.\blacksquare$$

Example 7. For the exponential density, we have, with $s = ax$,

$$\mu_2 = \int_0^\infty x^2 a e^{-ax} dx = \frac{1}{a^2}\int_0^\infty s^2 e^{-s} ds = \frac{1}{a^2}\Gamma(3) = \frac{2!}{a^2}.$$

and

$$\sigma^2 = \mu_2 - \mu^2 = \frac{2}{a^2} - \frac{1}{a^2} = \frac{1}{a^2}.\blacksquare$$

Example 8. For the *Gamma density,* we have, with $s = ax$,

$$E(X^2) = \frac{a^p}{\Gamma(p)} \int_0^\infty x^2 x^{p-1} e^{-ax} dx = \frac{1}{\Gamma(p)a^2} \int_0^\infty s^{p+1} e^{-s} ds = \frac{\Gamma(p+2)}{\Gamma(p)} \frac{1}{a^2} = \frac{(p+1)p}{a^2}.\blacksquare$$

and

$$\sigma^2 = \mu_2 - \mu^2 = \frac{p(p+1)}{a^2} - \frac{p^2}{a^2} = \frac{p}{a^2}.\blacksquare$$

Example 9. For the Standard Normal density,

$$\mu_2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \, e^{-x^2/2} dx = 1$$

according to Corollary 1 of Apendix A, and so

$$\sigma^2 = \mu_2 - \mu^2 = 1 - 0^2 = 1.\blacksquare$$

Remarks. (1.) Since the expectation and variance for general r.v. have the same basic properties as for discrete r.v., many of the results we proved for discrete r.v. also hold in general. In particular, the Theorems of Chapter 5 on the sample mean and variance, the formulas for the variance of sums and properties of covariance, the Markov and Chebyshev inequalities, the Inclusion-Exclusion proof and the Weak Law of Large Numbers all hold as stated.

*(2.) As a matter of fact, the notion of expectation can be extended to (at least) *all* bounded r.v., whether they have densities or not. The method is simply to approximate a continuous random variable by discrete random variables and take the limit. However, to prove the basic properties of expectation rigorously is rather technical, and a fully satisfactory treatment requires an analytical sophistication that we have not assumed. Analytical complexities, important though they are, are not our focus here, and so we will simply use without proof for continuous r.v., the properties of the expectation that we proved for discrete r.v.

7.4 The Cumulative Distribution Function.

Although many r.v. do not have densities, the distribution of a perfectly general r.v. X can be described by a single function, known as the *cumulative distribution function*. We define the *cumulative distribution function* (*c.d.f.*) of X to be

$$F_X(t) = P(X \leq t).$$

The probability

$$P(a < X \leq b)$$

can then be computed as

$$P(a < X \leq b) = P(X \leq b) - P(X < a) = F_X(b) - F_X(a).$$

Thus, to describe the distribution of X, it is sufficient so to give its c.d.f.

Theorem 4. *The c.d.f has the following properties.*

(a.) $F(t)$ *is increasing: if* $t \leq s$, *then* $F(t) \leq F(s)$.

(b.) $\lim_{t \to -\infty} F(t) = 0$.

(c.) $\lim_{t \to +\infty} F(t) = 1$.

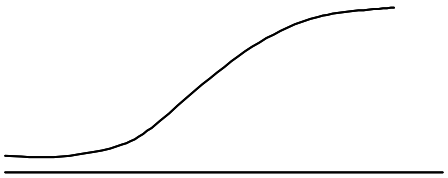

Figure 4.1. *A Cumulative Distribution Function.*

The *density is the derivative of the c.d.f.*

Theorem 5. *If* X *has a density* $f(x)$ *and c.d.f.* $F(x)$, *then*

$$f(x) = F'(x).$$

Proof. We have

$$F(t) = P(X \leq t) = \int_{-\infty}^{t} f(x)dx$$

so by the Fundamental Theorem of Calculus

$$F'(t) = \frac{d}{dt} \int_{-\infty}^{t} f(x)dx = f(t). \square$$

Example 1. For the simple Beta density

$$f(x) = 6x\,(1-x) \qquad 0 \leq x \leq 1$$

the c.d.f. is

$$\begin{array}{rcll} F(t) & = & 0 & x \leq 0 \\ & = & \displaystyle\int_0^t 6x\,(1-x)\,dx = 3t^2 - 2t^3 & 0 \leq x \leq 1 \\ & = & 1 & x \geq 1. \blacksquare \end{array}$$

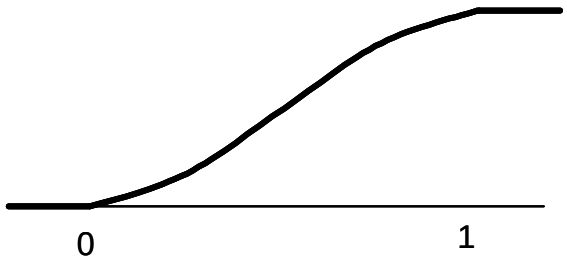

Figure 4.2. *c. d. f. of the Beta density* $6x(1-x)$.

Example 2. (*Exponential Density.*) For the exponential density, the c.d.f. is

$$\begin{aligned} P(T \leq t) = 1 - P(T > t) = 1 - e^{-at} \quad & \text{for } t > 0. \\ = 0 \qquad\qquad\qquad\qquad & \text{for } t > 0. \blacksquare \end{aligned}$$

Figure 4.3. *c.d.f of the Exponential density.*

Example 3. (*Uniform Density.*) For the uniform density on $[0, 1]$, the c.d.f. is

$$\begin{aligned} F(t) &= 0 \qquad \text{for } t \leq 0. \\ &= x \qquad \text{for } 0 \leq t \leq 1. \\ &= 1 \qquad \text{for } t \geq 1. \blacksquare \end{aligned}$$

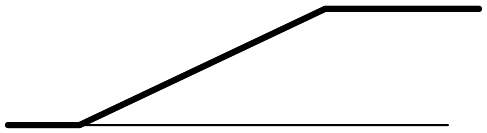

Figure 4.4.. *c.d.f of the Uniform Density.*

Discrete r.v. have c.d.f.'s, too.

Example 4. Consider the Binomial r.v. X with $n = 2$ and $p = 1/2$. The distribution is

$\mathbf{x}_i$	0	1	2
$\mathbf{p}_i$	$1/4$	$1/2$	$1/4$

The c.d.f.

$$F(t) = P(X \leq t)$$

is therefore given by

$$\begin{aligned} F(t) &= 0 && \text{for } t < 0. \\ &= \frac{1}{4} && \text{for } 0 \leq t < 1. \\ &= \frac{3}{4} && \text{for } 1 \leq t < 2. \\ &= 1 && \text{for } t \geq 2. \blacksquare \end{aligned}$$

The distribution and c.d.f are shown in Figure 4.5.

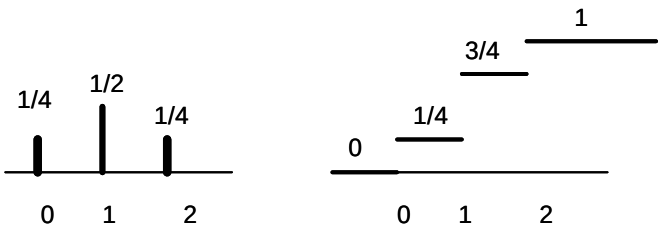

Figure 4.5 *Distribution and c.d.f. of* $Bin(2, \frac{1}{2})$.

The c.d.f. has jumps at the points where the possible values - the range - of the discrete r.v. occur. The jumps - not the *values* of $F(t)$, but its *jumps* - are equal to the probabilities that these values occur.■

Warning! *A point of frequent confusion is the difference between densities and cumulative distribution functions. Now is the time to get this straight !* Things will go much more smoothly from here on if this is clear in your mind.

The facts are as follows:

1. *Every r.v. has a c.d.f., but not every r.v. has a density*.

2. *If X has a density, the density is the derivative of the c.d.f. and the c.d.f. is the integral of the density.*

7.5 Translation and Scaling of Densities.

The c.d.f. is useful for finding and manipulating densities. A typical use is illustrated by the following result. The idea is to *find the c.d.f. first and then differentiate it to get the density*.

Theorem 6. *Let X have density $f(x)$. Then*

(a.) The density of $X + a$ is

$$f(x - a)$$

(b.) If $c \neq 0$, the density of cX is

$$\frac{1}{|c|} f(cx).$$

Proof. (a.) We want to compute the c.d.f $F_Y(t)$ of $Y = X + a$. It is

$$F_Y(t) = P(Y < t) = P(X + a < t) = P(X < t - a) = \int_{-\infty}^{t-a} f(x)dx.$$

By the Fundamental Theorem of Calculus,

$$f_Y(t) = F_Y'(t) = f(t - a)\frac{d}{dt}(t - a) = f(t - a).$$

(b.) Problem.□

Example 1. Let X be exponential with mean $1/a$. Find the distribution of $Y = aX$.

Solution. X has density $f(x) = ae^{-ax}$. By part (a.), Y has density

$$\frac{1}{a} f(\frac{x}{a}) = \frac{1}{a} \cdot ae^{-a(x/a)} = e^{-x}.$$

This is the unit exponential distribution; the parameter a has disappeared. Thus the parameter a in the exponential distribution is just a *scale parameter*.■

Similarly, the parameter a in the distribution $\Gamma(p, a)$ is a scale factor.

The General Normal Density.

The density

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

is called the general Normal density. It is obtained from the Standard Normal by scaling and translation.

Example 2. Let Z have the Standard Normal distribution. Find the density of $X = \sigma Z + \mu$.

Solution. Z has density

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

By (a.), the density of σZ is

$$\frac{1}{\sigma}\varphi(\frac{x}{\sigma}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$$

Now adding μ to σZ by part (b.) gives the density of $X = \sigma Z + \mu$ as

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

This is the normal density $N(\mu, \sigma)$. It has mean mean μ and variance σ. This is clear from the definition, since

$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \mu$$

and

$$var(\sigma Z + \mu) = var(\sigma Z) = \sigma^2 var(Z) = \sigma^2. \blacksquare$$

7.6 Densities of Functions of a Continuous Random Variable.

The random variables $X + a$ and cX are very simple functions of X. However, we can use the same method to find the densities of more general functions of X.

The method is the following:
Given the density $f_X(x)$, to find the density $f_Y(t)$ of $Y = \phi(X)$,

(1.) *compute the c.d.f. of $\phi(X)$, and*

(2.) *differentiate to find the density $f_Y(t) = F'_Y(t)$.*

Explicitly, the method consists of the following steps.

(1.) *Identify the ranges of X and $\phi(X)$; the densities are zero off these sets.*

(2.) *Write*

$$F_Y(t) = P(\phi(X) \leq t)$$

of $\phi(X)$.

(3.) *Express $F_Y(t)$ as*

$$F_Y(t) = P(\phi(X) \leq t) = \int_{S(t)} f(x)dx$$

where $S(t) = \{x : \phi(x) \leq t\}$.

(4.) *Differentiate to find the density:*

$$f_Y(t) = F'_Y(t).$$

Only the second step presents any difficulty.

Let us consider a few examples.

Example 1. *Find the distribution of $X = Z^2$ if Z is a standard normal random variable.*

Solution. We have

$$\begin{aligned} F_X(t) &= P(X \le t) = P(Z^2 \le t) = P(-\sqrt{t} \le Z \le \sqrt{t}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{t}}^{\sqrt{t}} e^{-z^2/2} dz = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{t}} e^{-z^2/2} dz. \end{aligned}$$

Hence, the density of X is

$$f_X(x) = F'_X(t) = \frac{2}{\sqrt{2\pi}} \frac{d}{dt} \int_0^{\sqrt{t}} e^{-z^2/2} dz = \frac{2}{\sqrt{2\pi}} e^{-\left(\sqrt{t}\right)^2/2} \frac{d\sqrt{t}}{dt} = \frac{e^{-t/2}}{\sqrt{2\pi t}}.$$

This density actually has a name - the χ^2 *distribution with one degree of freedom,* denoted by $\chi^2(1)$. It is a special case of the Gamma density;in fact, $\chi^2(1) = \Gamma(\frac{1}{2}, \frac{1}{2})$ ■

Example 2. (*The Cauchy density.*) Particles are emitted randomly in all directions by a point source at a distance a from a screen. (Figure 6.1.) Find the distribution of the position at which an emitted particle strikes the screen.

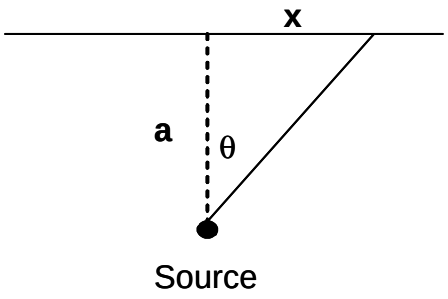

Figure 6.1

Solution. Let O be the point on the screen closest to the source. We want the distribution of and X, the distance from O to the point where the particle strikes the screen. The assumption is that the angle θ of emission is uniformly distributed on $-\pi/2 < \theta < \pi/2$. From Figure 6.1,

$$\tan\theta = \frac{X}{a}.$$

Hence, the c.d.f. is

$$F(t) = P(X \le t) = P(a\tan\theta \le t) = P(\theta \le \arctan(t/a)) = \frac{1}{\pi}\left(\arctan(t/a) + \pi/2\right)$$

Differentiating, we find the density

$$F'(t) = \frac{d}{dt}\left(\arctan(t/a) + \pi/2\right) = \frac{a}{\pi}\frac{1}{x^2+a^2}.$$

This density is called the Cauchy density. *It has no expectation*, since the integral

$$\frac{a}{\pi}\int_{-\infty}^{\infty}\frac{x}{x^2+a^2}dx.$$

does not converge. Thus, if we wanted to estimate the location of the source by looking at the pattern of spots where particles have hit the screen, computing the sample mean would not be useful.■

Example 3. *Let the interval* $[0,1]$ *be divided randomly divided into two pieces. Find the distribution of the product* Y *of the two pieces.*

Solution. We assume that the coordinate X of the division point is uniformly distributed on $0 \le x \le 1$. Then $Y = X(1-X)$. has the *range* $0 \le Y \le 1/4$. We have

$$F(t) = P(Y \le t) = P(X(1-X) \le t).$$

But $x(1-x) \le t$ holds iff either

$$x \le \frac{1-\sqrt{1-4t}}{2} \text{ or } x \ge \frac{1+\sqrt{1-4t}}{2}.$$

Thus, for $0 < t < 1/4$,

$$\begin{aligned} F(t) &= P(X(1-X) \le t) = \int_0^{\left(1-\sqrt{1-4t}\right)/2} dx + \int_{\left(1+\sqrt{1-4t}\right)/2}^1 dx \\ &= 1-\sqrt{1-4t}. \end{aligned}$$

and of course, $F(t) = 0$ for $t \le 0$, and $F(t) = 1$ for $t \ge 1/4$.

(At this point, it is good to check the formula by noting that $F(0) = 0$ and $F(\frac{1}{4}) = 1$.)

We can now differentiate to get

$$f(t) = F'(t) = \frac{2}{\sqrt{1-4t}} \qquad 0 \le t \le 1/4$$

and $f(t) = 0$ otherwise.■

Important Remark. *It is quite possible that the density of $\phi(X)$ may be different on different intervals.* The following example illustrates this problem.

Example 4. *Find the distribution of X^2 if X is uniform on $-1 < x < 2$.*

Solution. This is a little tricky. *We must be careful*.

First, check the *range* of Y. Since X lies between -1 and 2, its square lies between 0 and 4. The range is therefore $0 \leq Y \leq 4$.

The c.d.f. is therefore

$$F(t) = P(X^2 \leq t) = \int_{-\sqrt{t}}^{\sqrt{t}} f(x)\, dx$$

for $0 \leq t \leq 4$. Now the density is $f(x) = 1/3$ on $-1 < x < 2$, and zero elsewhere, so there are *two cases* here. *We cannot just stuff $f(x) = 1/3$ into the integral, because $f(x)$ is equal to $1/3$ only for $-1 \leq x \leq 2$, and is zero otherwise*. So the integral is different for $0 < t < 1$ than it is for $1 < t < 4$. We have

$$\begin{aligned} F(t) &= P(X^2 \leq t) = \int_{-\sqrt{t}}^{\sqrt{t}} f(x)\, dx \\ &= \int_{-\sqrt{t}}^{\sqrt{t}} \frac{1}{3}\, dx = \frac{2}{3}\sqrt{t} \qquad \text{if } 0 \leq t \leq 1 \\ &= \int_{0}^{\sqrt{t}} \frac{1}{3}\, dx + \int_{-1}^{0} \frac{1}{3}\, dx = \frac{1}{3}\sqrt{t} + \frac{1}{3} \qquad \text{if } 1 \leq t \leq 4. \end{aligned}$$

Let's check this formula before proceeding. At the ends of the range of X, we have $F(0) = 0$ and $F(4) = 1$, as we should. Both formulas match at $t = 1$, as they should . We also have

$$F(1) = P(X^2 \leq 1) = P(-1 < X < 1) = 2/3$$

which is also correct.

Things look encouraging, so we proceed to differentiate $F(t)$ to get $f(t)$. The density is

$$f(t) = F'(t) = \begin{cases} \frac{1}{3\sqrt{t}} & \text{if } 0 \leq t \leq 1 \\ \frac{1}{12\sqrt{t}} & \text{if } 1 \leq t \leq 4 \\ 0 & \text{otherwise.} \end{cases}$$

Be sure that you understand every step of this example. Nothing here is hard to figure out if you ask the right questions. But you must *think* at all stages.■

Remark. It is always good to check to see if your answer is reasonable. Always ask:

(1.) Does your c.d.f. have the properties of Theorem 4 ?

(2.) Does the final density integrate to 1?

(3.) Is the density zero off the range?

(4.) Is the density positive?

(5.) Can you think of anything else in your particular problem that you might check?

7.7 *A Censored Random Variable.

Suppose that we wish to measure the lifetime T of some device, say a light bulb. To measure T, we simply turn on the bulb and wait for it to burn out. There is, however, clearly a limit to our patience. We may not wish to wait for *every* bulb to fail. So, if a bulb is still burning after some cutoff time c, we will just record that it had a lifetime longer than c.

What we are actually measuring, then, is not T itself but *the minimum X of T and c.* This is called a *censored* r.v.

Suppose, for example, that T is exponential with parameter a, and let $X = \min(T, c)$, where $c > 0$. What is the distribution of X?

Let's compute the c.d.f. $F(t)$ of X. Note that for $t < c$, we have $X \leq t$ iff $T \leq t$, and that we always have $X \leq c$. Therefore,

$$F(t) = P(X \leq t) = \begin{cases} P(T \leq t) = 1 - e^{at} & \text{if } 0 \leq t < c \\ 1 & \text{if } t \geq c \end{cases}$$

The graph of $F(t)$ is shown in Figure 7.1.

Figure 7.1 *c.d.f. of* $\min(T, c)$.

$F(t)$ is smoothly varying up to the point c, at which it has a jump of size

$$P(X = c) = e^{-ac}.$$

Thus X is *not discrete*, since it may take on any value in the interval $[0, c]$, *nor does it have a density*, since it is equal to c with positive probability. The r.v. X thus has a continuous part $0 < x < c$, with density

$$f(x) = ae^{-ax} \qquad 0 < x < c$$

and a discrete part with a positive probability e^{-ac} to be equal to c.

The expectation of X may be easily computed by combining the discrete and continuous methods, averaging with the density over $[0, c]$ and treating the point c as for a discrete r.v. Explicitly,

$$\begin{aligned} E(X) &= \int_0^c x\, ae^{-ax} dx + cP(X = c) \\ &= \frac{1 - e^{-ac}}{a}. \end{aligned}$$

As a check, note that as $c \to \infty$, this tends to $1/a$, which is the expectation of T.

7.8 The Jacobian Method.

There is another method for computing the density of a function of a r.v. X. *When it applies*, it is much easier to use because it involves computation with a minimum of thought. The trouble is that it *does not always apply*, so thought is needed to determine when it may be used. It works for Example 2 of the previous section, but not for any of the other three Examples.

To be precise, the Jacobian method for the distribution of $Y = \phi(X)$ works whenever $\phi(x)$ is *one-to-one on the range of* X, that is to say, when $\phi(x)$ is either strictly increasing or strictly decreasing on the range of X. *Otherwise, it gives the wrong answer*.

The method is the following.

(1.) If $y = \phi(x)$ is *strictly increasing* on the range of X, we can *solve for* x *as a function of* y.

$$x = u(y).$$

(2.) Then

$$\begin{aligned} F_Y(t) &= P(\phi(X) \le t) = P(X \le u(t)) \\ &= \int_{-\infty}^{u(t)} f(x)\, dx \\ f(t) &= F'(t) = f(u(t))u'(t). \end{aligned}$$

If $y = \phi(x)$ is *strictly decreasing* on the range of X, we can still *solve for* x *as a function of* y, but

$$\begin{aligned} F_Y(t) &= P(\phi(X) \le t) = P(X \ge u(t)) \\ &= \int_{u(t)}^{\infty} f(x)\, dx \\ f(t) &= F'(t) = -f(u(t))u'(t). \end{aligned}$$

This is good, because for a decreasing function, with $u'(t) < 0$. The two formulas can be combined as

$$f(t) = F'(t) = f(u(t))\,|u'(t)|\,.$$

Example 1. *Let X be exponential with $a = 1$. Find the density of $Y = \sqrt{X}$.*

Solution. The range of X is $0 \le x < \infty$, and the function $\phi(x) = \sqrt{x}$ is one-to-one on $0 \le x < \infty$, so the method is applicable. Solving for Y gives $X = Y^2$. The density of X is

$$f(x) = e^{-x} \qquad x \ge 0$$

so we compute

$$f_Y(t) = f(t^2)u'(t) = 2t\,e^{-t^2} \qquad t \ge 0.\blacksquare$$

Example 2. Let X have the unit exponential density

$$f(x) = e^{-x} \qquad x > 0.$$

Find the density of $Y = e^{-X}$.

Solution. Note that the range of Y is $[0, 1]$. Solve $y = e^{-x}$ for y to get

$$x = u(y) = -\log y$$

The density is

$$f_Y(y) = f(u(y))\,|u'(y)| = e^{-(-\log y)}\left|\frac{-1}{y}\right| = y \cdot \frac{1}{y} = 1$$

on the range $[0, 1]$ of Y. ■

It is thus important to check the *ranges* of X and especially of Y. The formal calculation gives the formula for $f_Y(y)$, but does not tell *where it applies*. Remember that the range of Y need not be the range of X.

A *brief statement of the Jacobian Method*, which is easy to remember is to write

$$f_X(x)dx = f_Y(y)dy$$

Thus, in Example 1,. with $x = y^2$, we get

$$e^{-x}dx = e^{-y^2}2ydy$$

so that

$$f_Y(y) = 2y\,e^{-y^2}.$$

But remember to take the absolute value if $u(y)$ is decreasing.

Example 3. *Let X be uniform on $[0, 1]$. Find the density of $Y = e^X$.*

Solution. We have $X = \log Y$, and so

$$1 \cdot dx = \frac{1}{y} dy$$

so that

$$f_Y(t) = \frac{1}{y}.$$

But the integral of $1/y$ diverges on $[0, 1]$! How can that be?

Stop and think. The range of X is $[0, 1]$. Therefore, the range of Y is $\left[e^0, e^1\right] = [1, e]$. *The formula holds only on the range.* Thus more precisely, we have

$$\begin{aligned} f_Y(y) &= \frac{1}{y} \qquad 1 \le y \le e \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

The normalization integral is then

$$\int_1^e \frac{1}{y} dy = 1$$

as it should be. None of this is hard *if you ask the right questions*. Always ask first about ranges.■

Example 4. (*An Incorrect Calculation.*) What happens if you apply the Jacobian method when $\phi(x)$ is not one-to-one? Consider the problem of Example 1 of the preceding section: *Find the distribution of $Y = Z^2$ if Z is a standard normal random variable.*

If we simply used the Jacobian method without thought, we would set $Z = \sqrt{Y}$, and write

$$f(y)\, dy = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} dx$$

and hence

$$f_Y(y) = \frac{1}{2\sqrt{2\pi}} e^{-y/2} \qquad \text{for } y \ge 0.$$

This is *half the right answer*:

$$\int_0^\infty f_Y(y) dy = \int_0^\infty \frac{1}{2\sqrt{2\pi}} e^{-y/2} dy = \frac{1}{2}.$$

The problem is that $y = z^2$ is *not one-to-one on the range* $-\infty < z < \infty$, so the *Jacobian method does not work.* ■

Final Remarks.

(1.) We have discussed the c.d.f. method first because

(a.) the c.d.f. method is more general, and also because

(b.) students may tend to rely on the Jacobian method in all cases because it is computational and can be applied without much thought.

(2.) Why is this called the *Jacobian* method? It is because in the version for joint densities of several r.v., which we will encounter in the next chapter, the derivative $u'(y)$ is replaced by a Jacobian determinant.

7.9 Problems.

(1.) Determine the normalization constant C to make each of the following functions a density on the indicated range. Find the mean and variance. Sketch the graph of $f(x)$.

(a.) Triangular density.

$$f(x) = C(a - |x|), \quad -a \le x \le a.$$

(b.) Bilateral exponential (Laplace).

$$f(x) = Ce^{-a|x|}, \quad -\infty < x < \infty.$$

(c.) The density.

$$f(x) = C(1 - x^2), \quad -1 \le x \le 1.$$

(2.) Find and sketch the cumulative distribution function $F(x)$ for the *Cauchy density*,

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

where $a > 0$.

(3.) The probability density function of X, the lifetime of a certain type of electronic device (measured in hours), is given by

$$f(x) = \begin{cases} 10/x^2, & x > 10; \\ \\ 0, & x \le 10. \end{cases}$$

(a.) Find $P\{X > 20\}$.

(b.) What is the cumulative distribution function of X?

(c.) What is the probability that of 6 such devices at least 3 will function for at least 15 hours? What assumptions are you making?

(4.) A filling station is supplied with gasoline once a week. If its weekly volume of sales in thousands of gallons is a random variable with probability density function

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

what should the capacity of the tank be so that the probability of the supply's being exhausted in a given week is 0.01 ?

(5.) Prove that if the c.d.f. $F(x)$ of X is continuous, then the r.v. $F(X)$ is uniform on $[0,1]$.

(6.) Find the c.d.f. of the *Pareto density*

$$\frac{a}{(x+1)^{a+1}}, \qquad x > 0$$

When is the mean finite? When is the variance finite? Find mean and variance when they exist.

(7.) Let X have density

$$f(x) = \frac{3}{4}\left(1 - x^2\right) \qquad 0 \le x \le 1.$$

Find the following:

(a.) The mean of X.

(b.) The probability that $X^2 \ge \frac{1}{2}$.

(8.) Let X have density

$$f(x) = Cx^2\,(a - x) \qquad 0 \le x \le a.$$

Find the following:

(a.) The normalization constant C.

(b.) The mean of X.

(c.) The variance of X.

(d.) The cumulative distribution function.

(9.) Let X be a random variable such that for $t \ge 0$,

$$P(X \ge t) = e^{-t^2/2}.$$

(a.) Find the density of X.

(b.) Find the mean of X.

(10.) If the c.d.f. of the r.v. X is

$$F(t) = \begin{cases} 0 & \text{if } t \le 0. \\ 3x^2 - 2x^3 & \text{if } 0 < t < 1. \\ 1 & \text{if } t \ge 1. \end{cases}$$

find the density of X .

(11.) If the density of the r.v. X is

$$f(x) = \begin{cases} 2/x^3 & \text{if } x > 1. \\ 0 & \text{if } x < 1. \end{cases}$$

find the c.d.f. of X

(12.) Suppose that the r.v. X has the density

$$f(x) = \begin{cases} 30\ x^2\ (1-x)^2 & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Compute:

(a.) $P(X < \frac{1}{3})$.

(b.) $E(X)$.

(c.) The cumulative distribution function of X.

(13.) Let X have density

$$f(x) = \frac{2}{(1+x)^3} \qquad 0 \leq x < \infty.$$

Find the following:

(a.) The probability that $X > 2$.

(b.) The mean of X.

(c.) The variance of X.

(d.) The c.d.f. of X.

(e.) The range and density of $Y = 1/X$.

(14) (*Laplace's error distribution*.) Find the mean and variance of the density.

$$f(x) = \frac{1}{2} \log\left(1/\left|x\right|\right) \qquad -1 < x < 1.$$

This density was proposed by Laplace as a model for the distribution of errors in measurements, but was not a convenient model.

(15.) Trains headed for destination A arrive at the train station at 15-minute intervals starting at *7:00* a.m., whereas trains headed for destination B arrive at 15-minute intervals starting at *7:05* a.m.. If a passenger arrives at the station at a time uniformly distributed between 7 and 8 a.m. and gets on the first train that arrives, what proportion of time does he take the A train?

(16.) A bus travels between the two cities A and B, which are 100 miles apart. If the bus has a breakdown, the distance from the breakdown to city A has a uniform distribution over $(0, 100)$. There is a bus service station in city A, in city B, and in the center of the route

between A and B. It is suggested that it would be more efficient to have the three stations located $25, 50,$ and 75 miles, respectively, from A. Do you agree? Why?

(17.) In the preceding problem, find the optimal location of the three stations that minimizes the expected distance from breakdown to station.

(18.) You arrive at a bus stop at 10 o'clock, knowing that the bus will arrive at some time uniformly distributed between 10:00 and 10:30. What is the probability that you will have to wait longer than 10 minutes? If at 10:15 the bus has not yet arrived, what is the probability that you will have to wait at least an additional 10 minutes?

(19.) Let $X_1, X_2, \ldots, X_n$ be independent r.v. Define

$$Y = \min(X_1, \ldots, X_n),$$

and

$$Z = \max(X_1, \ldots, X_n).$$

Prove that

$$P(Y \geq t) = P(X_1 \geq t) \cdots P(X_n \geq t),$$

and

$$P(Z < t) = P(X_1 < t) \cdots P(X_n < t).$$

(20.) Calls arrive independently on each of n telephones. The waiting time T_k for the k^{th} telephone to ring is exponential, with mean λ. Let T be the waiting time until one of the phones rings. Prove that T is exponential and find its mean.

(21.) Suppose that 30 electronic devices, say $D_1, \ldots, D_{30}$, are used in the following manner: As soon as D_1 fails D_2 becomes operative. When D_2 fails D_3 becomes operative, etc. Assume that the time to failure of D_1 is an exponentially distributed random variable with parameter $= 0.1$ hour^{-1}. Let T be the total time of operation of the 30 devices. What is the probability that T exceeds 350 hours?

x (22.) Prove that if the c.d.f. $F(x)$ of X is continuous, then the r.v. $F(X)$ is uniform on $[0, 1]$.

(22.) Find the density of X if the r.v. $Y = X^\beta$ is a unit exponential r.v. (*i.e.*, with $a = 1$).

(23.) A point is chosen at random on the circumference of a circle of radius R with center at the origin (in other words, the polar angle of the point chosen is uniformly distributed in the interval $(-\pi, \pi)$). Find the density function for

(a.) the abscissa of the point selected;

(b.) the length of the chord joining this point to the point $(-R, 0)$.

(24.) Suppose that the continuous random variable X has density

$$f(x) = e^{-x}, \qquad x > 0$$

Find the density of the following random variables:

(a.) $Y = X^3$;

(b.) $Z = (X-1)^2$

(25.) If X is an exponential random variable with parameter $\lambda = 1$, compute the probability density function of the random variable Y defined by $Y = \log X$.

(26.) If X is uniformly distributed over $(0,1)$, find the density function of $Y = e^X$.

(27.) A point is chosen at random from the segment $[0,a]$, dividing it into two segments.

(a.) Find the probability that the longer segment L is at least 4 times as long as the shorter segment.

(b.) Find the expectations of L and S.

(c.) Find the mean and distribution of $Y = L/S$.

(28.) Find the distribution of the length of the chord in *Bertrand's Paradox* in each of the three cases.

(29.) Let X be exponentially distributed with $a = 1$. Find the density of the random variable $Y = e^{-X}$.

(30.) Find the density of $Y = X^2$ if X is uniformly distributed on $[-1,2]$.

(31.) A r.v Y has the *Weibull distribution* on $y \geq v$ iff

$$P(Y \geq t) = \exp\left\{-\left(\frac{t-v}{\alpha}\right)^{\beta}\right\}$$

Prove that Y is Weibull iff

$$X = \left(\frac{Y-v}{\alpha}\right)^{\beta}$$

is a unit exponential.

(32.) A r.v. W is *lognormal* iff $\log W$ has a normal distribution $N(\mu, \sigma^2)$. Find the density of W.

(33.) Let the r.v. X have the uniform distribution on $[0,1]$ and let $Y = \log\left(\frac{1}{X}\right)$.

(a.) Find the range of Y.

(b.) Find the density of Y.

(34.) Let T have the exponential distribution with $a = 1$. Find the density and mean of $X = e^{-T}$.

(35.) Let X have the exponential distribution with $a = 1$. Find the density for $Y = \sqrt{X}$.

Section 7.9 Problems.

(36.) Let X have the exponential distribution with $a = 1$. Find the *range* and *density* of $Y = e^X$.

(37.) Let X uniformly distributed on the interval $[0, 1]$. Find the *range* and *density* of $Y = 1/\sqrt{X}$.

(38.) Let X have the density

$$f(x) = 3x^2 \qquad 0 \leq x \leq 1.$$

(a.) Find the probability that $X > \frac{1}{2}$.

(b.) Find the expectation of $\sqrt{X}$.

(39.) Let the r.v. X have the uniform distribution on $[0, 1]$ and let $Y = (1 - 3X)^2$.

(a.) Find the range of Y.

(b.) Find the density of Y.

(40.) Solve the problem of Example 2 of section 6 if the screen is two-dimensional.

(41.) Show that the density

$$f(x) = \frac{3}{(x+1)^4} \qquad x \geq 0.$$

has finite mean and variance but no moments higher than the second.

(42.) Prove that the mean is the number that minimizes the function $f(a) = E(X - a)^2$.

(43.) (a.) Prove in general that the mean is the number that minimizes the function $f(a) = E(X - a)^2$.

(b.) Let X have positive density $f(x)$. For what value m is $f(m) = E\,|X - m|$ a minimum? (*Hint:* See section 4.6.)

(44.) Prove (b.) of Theorem 6.

Chapter 8
Several Continuous Random Variables

8.1 The Joint Density.

The cumulative distribution function of a single r.v. X tells us all that can be known about X alone from the view of probability theory. However, as we saw in the case of discrete r.v., given two r.v. X and Y, there are not only questions about the distributions of X and Y alone, but also about their relationship - about how the values of X correspond with those of Y.

The *joint distribution function* (j.d.f.) of the pair of r.v. (X, Y) is the function

$$F(t, s) = P(X \leq t \ \& \ Y \leq s).$$

From it, one can compute the probability

$$P(a < X \leq b \ \& \ c < Y \leq d)$$

for any intervals $(a, b]$ and $(c, d]$. It therefore describes completely the relation between X and Y. This will work for any pair of r.v., even discrete ones.

The j.d.f. is analogous to the c.d.f. for a single r.v. As with single continuous r.v., the j.d.f. *may* sometimes be expressed in terms of a density function.

Definition 1. We say that X and Y have *joint density* $f(x, y)$ if

$$F(t, s) = \int_{-\infty}^{s} \int_{-\infty}^{t} f(x, y) \, dxdy$$

If follows that for any (reasonable) set S

$$P(X \in S) = \int \int_{S} f(x, y) \, dxdy. \tag{8.1}$$

As with densities, in order for a function $f(x, y)$ to be used as a joint density, it is necessary that two things be true. First, the formula (8.1) cannot give negative numbers for any probability. Therefore, we need

$$f(x, y) \geq 0$$

for all x. Secondly, we need

$$P(-\infty < X < \infty \ \& \ -\infty < Y < \infty) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dxdy = 1.$$

Any function satisfying these two conditions is a possible joint density.

As with discrete r.v., we may compute the densities $f_X(x)$ of X and $f_X(x)$ of Y from $f(x,y)$ as the *marginal densities*

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\ f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx \end{aligned}$$

Example 1. Consider the joint density function

$$\begin{aligned} f(x,y) &= C\left(3x^2+4xy\right) \qquad 0\le x\le 1,\ 0\le y\le 1 \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned}$$

(a.) Determine C so that $f(x,y)$ is a joint density.

(b.) Compute $P(Y<X^2)$.

(c.) Find the densities of X and Y.

Solution. (a.) We first check that the function $f(x,y)$ is positive on the unit square, which it clearly is. The normalization constant C is determined by requiring that

$$1=\int_0^1 f(x,y)\,dxdy=\int_0^1\int_0^1 C\left(3x^2+4xy\right)\,dxdy=2C$$

so that $C=1/2$, and

$$\begin{aligned} f(x,y) &= \frac{3}{2}x^2+2xy \qquad 0\le x\le 1,\ 0\le y\le 1 \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned}$$

(b.) We have

$$\begin{aligned} P(Y &< X^2)=\int\int_{\{x,y);0<y<x^2\}} f(x,y)\,dxdy=\int_0^1\int_0^{x^2}\left(\frac{3}{2}x^2+2xy\right)\,dydx \\ &= \int_0^1\left[\frac{3}{2}x^2y+xy^2\right]_{y=0}^{y=x^2}dx=\int_0^1\left(\frac{3}{2}x^4+x^5\right)dx=\frac{7}{15}. \end{aligned}$$

(c.) We have

$$\begin{aligned} f_X(x) &= \int_0^1\left(\frac{3}{2}x^2+2xy\right)\,dy=\left[\frac{3}{2}x^2y+xy^2\right]_{y=0}^{y=1}=\frac{3}{2}x^2+x \qquad 0\le x\le 1. \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \\ f_Y(y) &= \int_0^1\left(\frac{3}{2}x^2+2xy\right)dx=\left[\frac{1}{2}x^3+x^2y\right]_{x=0}^{x=1}=y+\frac{1}{2} \qquad 0\le y\le 1 \\ & 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned}$$

The expectation of a function $\phi(X,Y)$ of X and Y can be computed from a version of the *Law of the Unconscious Statistician.*

Theorem 1. (Law of the Unconscious Statistician.) *If X and Y have joint density $f(x,y)$, then*

$$E(\phi(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi(x,y)f(x,y)dxdy.$$

Example 2. For the joint density of Example 1 above, compute the expectation of $\phi(X,Y) = XY$.

Solution. By Theorem 1

$$\begin{aligned} E(XY) &= \int_0^1\int_0^1 xy\,\left(\frac{3}{2}x^2+2xy\right)dxdy \\ &= \int_0^1\int_0^1 \frac{3}{2}x^3y+2x^2y^2\,dxdy = \frac{59}{144}.\blacksquare \end{aligned}$$

Joint Density of Independent r.v.

There is one case in which a joint density is easily found. If X and Y are *independent* r.v. with densities $f(x)$ and $g(y)$ respectively, then (X,Y) has the joint density

$$f(x,y) = f(x)g(y).$$

Conversely, if $f(x,y)$ has this form, then X and Y are independent. In other words, in analogy to the discrete case, *X and Y are independent if the joint density is the product of the marginal densities.*

Proof. By the definition of independence

$$\begin{aligned} P(a &< X \le b \,\&\, c < Y \le d) = P(a<X\le b)P(c<Y\le d) \\ &= \left(\int_a^b f(x)\,dx\right)\left(\int_c^d g(y)\,dy\right) = \int_a^b\int_c^d f(x)g(y)\,dxdy.\square \end{aligned}$$

Example 4. Let X and Y be independent r.v., uniformly distributed on $[0,a]$. The joint density is

$$f(x,y) = \frac{1}{a}\frac{1}{a} = \frac{1}{a^2} \qquad 0<x<a,\ 0<y<a.\blacksquare$$

Example 5. Let X and Y be independent unit normal r.v.. The joint density is

$$f(x,y) = \left(\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\right) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}.\blacksquare$$

Example 6. Let X and Y be independent exponential r.v., with parameter a. The joint density is

$$f(x,y) = \left(ae^{-x}\right)\left(ae^{-y}\right) = a^2e^{-(x+y)} \qquad 0 < x,\ 0 < y. \blacksquare$$

Example 7. Let X and Y have joint density

$$f(x,y) = \frac{1}{2}x^2e^{-(x+y)} \qquad 0 < x < \infty,\ 0 < y < \infty.$$

The joint density factors as

$$f(x,y) = \left(\frac{1}{2}x^2e^{-x}\right)\cdot\left(e^{-y}\right). \blacksquare$$

The r.v. X and Y are therefore independent.

Example 8. Remember, however, that *the domain of $f(x,y)$ must be taken into account*. Always ask where the density is zero; i.e. where the formula is valid. Consider, for example, two r.v. where the pair (X,Y) is uniformly distributed over the *unit disc*:

$$\begin{aligned} f(x,y) &= \frac{1}{\pi} \qquad x^2+y^2 < 1 \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Now the number $1/\pi$ is certainly the product of a (constant) function of x and a (constant) function of y, so a careless reader might think that X and Y are independent. Closer consideration shows that this cannot be so. For while the value of Y may be any number between -1 and $+1$, if X has the value x, then the value of Y must lie between $-\sqrt{1-x^2}$ and $\sqrt{1-x^2}$. So the value of X has a definite effect on the possible value of Y. *This is not independence*.

To pursue this further, let us compute the marginals. We have

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}dy = \frac{2}{\pi}\sqrt{1-x^2} \qquad -1 < x < 1.$$

By symmetry,

$$f_Y(y) = \frac{2}{\pi}\sqrt{1-y^2} \qquad -1 < y < 1.$$

The product of the marginals is

$$f_X(x)f_Y(y) = \left(\frac{2}{\pi}\right)^2\sqrt{1-x^2}\sqrt{1-y^2} \qquad -1 < x < 1, -1 < y < 1$$

which is definitely not $f(x,y)$.

Joint Densities of Several r.v.

One can also have a joint density $f(x_1, x_2, \ldots, x_n)$ of several r.v. $X_1, X_2, \ldots, X_n$, such that

$$P((X_1, X_2, \ldots, X_n) \in S) = \int \int \cdots \int_S f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n.$$

They occur widely in Statistics.

WARNING.

Two r.v. X and Y may not have a joint density, even if both X and Y are continuous r.v. with densities.

For example, if $Y = X^2$, the values of the pair (X, Y) all lie on the parabola $y = x^2$, so the joint density $f(x, y)$, if there were one, would have to be identically zero everywhere except on this parabola. But the integral over all values of such a function would be zero, so it could not be a joint density.

This situation occurs in Statistics where one sometimes deals with a set $X_1, X_2, ..., X_n$ of r.v. with

$$X_1 + X_2 + \cdots + X_n = 0.$$

8.2 The Distribution of Functions of X and Y.

The procedure for finding the distribution of a function $W = \phi(X, Y)$ of two variables is essentially the same as in the case of a function $\phi(X)$ of one variable:

(1.) Find the range of W.

(2.) Compute the c.d.f.

(3.) Differentiate.

This procedure may be, and usually is, more complicated with more than one variable, but the method is the same.

The chief complication is the following. The c.d.f. of W is

$$P(W \leq t) = \int \int_{S(t)} f(x, y)\, dxdy$$

where

$$S(t) = \{(x, y) : \phi(x, y) \leq t\}$$

is the set of points (x, y) where $\phi(x, y) \leq t$. We must therefore set the limits in a double integral over the set $S(t)$.

WARNING:

Do not try to set up these limits with out drawing a picture of the set $S(t)$ and the set where $f(x,y)$ is positive.

If you draw the picture correctly, you will probably get the correct limits *If you do not draw it, you will almost certainly get them wrong.* I guarantee it. Just trying to imitate an example, by sticking some formulas into what you hope is the right place *will not work*.

Don't even try it. *Draw the Picture*!

Example 1. *Find the distribution of $D = |X - Y|$ if X and Y are uniform on $[0, a]$.*

Solution. As above, the joint density is $f(x,y) = 1/a^2$ on the square. The c.d.f. of D is therefore, for $0 < t < a$,

$$F(t) = P(D \le t) = P(|X - Y| \le t) = \int\int_{S(t)} \frac{1}{a^2} dxdy = \frac{area\,(S(t))}{a^2}$$

where $S(t) = \{(x,y) : |x - y| \le t\}$.

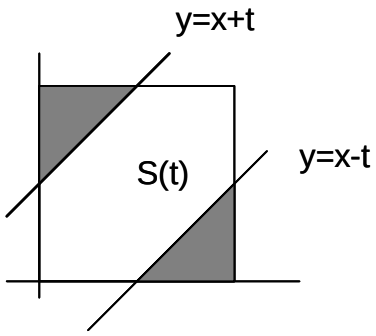

Figure 2.1

From Figure 2.1, we see that

$$area\,(S(t)) = a^2 - (a - t)^2 = 2at - t^2$$

Thus

$$F(t) = \frac{2at - t^2}{a^2}.$$

By differentiation we obtain the triangular density

$$f(t) = F'(t) = \frac{2}{a^2}(a - t) \qquad 0 \le t \le a.$$

which is shown in Figure 2.2. ■

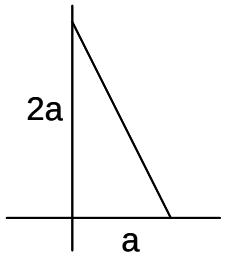

Figure 2.2.

Example 2. Find the distribution of $S = X + Y$ for the same distributions.

Solution. The c.d.f. of S is, for $0 < t < 2a$,

$$F(t) = P(D \leq t) = P(X + Y \leq t) = \int \int_{S(t)} \frac{1}{a^2} dxdy$$

where

$$S(t) = \{(x, y) : x + y \leq t,\ 0 < x < a,\ 0 < y < a\}$$

From Figure 2.3., we see that there are two cases, according as $0 < t < a$, of $a < t < 2a$.

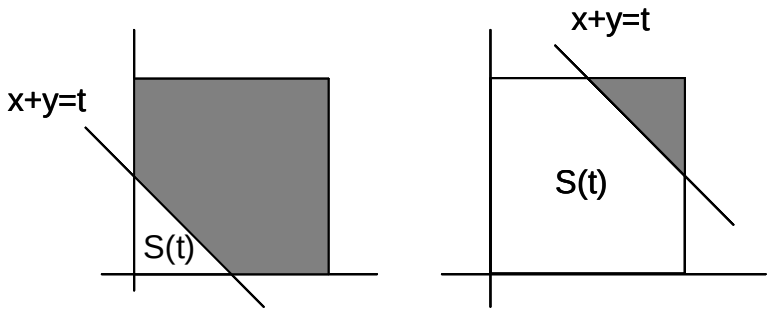

Figure 2.3a **Figure 2.3b**

In the case $0 < t < a$, $F(t)$ is $1/a^2$ times the area of the triangle in Figure2.3a:

$$F(t) = \frac{1}{a^2} \cdot \frac{1}{2} t^2 \qquad 0 < t < a$$

In the case $a < t < 2a$, $F(t)$ is 1 minus $1/a^2$ times the area of the triangle in Figure2.3b.

$$F(t) = 1 - \frac{1}{a^2} \cdot \frac{1}{2} (2a - t)^2 \qquad a < t < 2a$$

Differentiating gives the triangular density

$$\begin{aligned} f(x,y) &= t/a^2 && 0 \le t \le a \\ &= (2a-t)/a^2 && a \le t \le 2a \\ &= 0 && \text{otherwise} \end{aligned}$$

which is shown in Figure 2.4.■

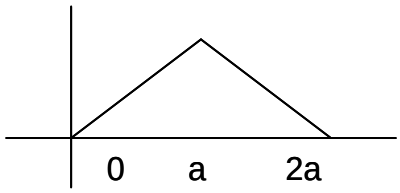

Figure 2.4. *Density of sum of independent uniform r.v.*

This problem shows the absolute necessity of

(1.) drawing a picture when setting up limits, and

(2.) paying attention to when the formulas apply.

Example 3. *Find the distribution of $Q = X/Y$ where X and Y are independent exponential r.v. with parameter a. as above, the joint density is*

$$f(x,y) = \left(ae^{-x}\right)\left(ae^{-y}\right) = a^2 e^{-(x+y)} \qquad 0 < x,\ 0 < y.$$

Solution. We write $F_Q(t) = 1 - P(Q > t)$, since it is easier to compute $P(Q > t)$. Let $S(t) = \{(x,y) : x/y > t\}$.

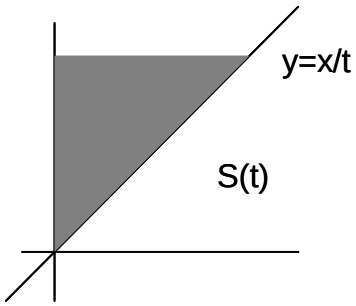

Figure 2. 5.

From Figure 2.5, we see that

$$\begin{aligned} P(Q &> t) = P(X > tY) = \int\int_{\{x>ty>0\}} a^2 e^{-(x+y)} dy dx = \int_0^\infty \int_{ty}^\infty a^2 e^{-(x+y)} dx dy \\ &= a^2 \int_0^\infty e^{-y} \left(\int_{ty}^\infty e^{-x} dx \right) dy = a \int_0^\infty e^{-y} e^{-aty} dy = \frac{1}{1+t}. \end{aligned}$$

The density is therefore

$$f_Q(t) = -\frac{d}{dt} P(Q > t) = \frac{1}{(1+t)^2}. \blacksquare$$

Remark: Note that *the density is independent of* a. This could have been seen from the first. Recall that if X is exponential with parameter a, then aX is exponential with parameter 1. But

$$Q = \frac{X}{Y} = \frac{aX}{aY}$$

so Q is actually the quotient of two unit exponentials. We could therefore have taken $a = 1$ from the start, which would have simplified our calculation somewhat. Noticing things like this *in advance* can often simplify calculations considerably.

Example 4. Find the distribution of $W = XY$ where X and Y have the joint density

$$f(x, y) = 4xy \qquad 0 < x, y < 1.$$

Solution. From Figure 2.6,

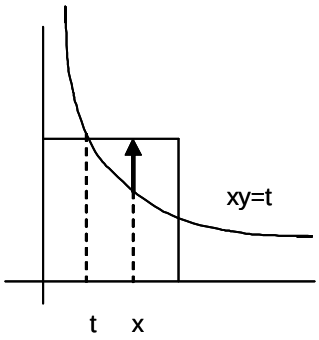

Figure 2.6.

we see that, for $0 < t < 1$, we have

$$\begin{aligned} F(t) &= P(XY \le t) = 1 - P(XY > t) \\ &= 1 - \int_t^1 \int_{t/x}^1 4xy \, dy \, dx = t^2 - 2t^2 \log t \end{aligned}$$

Differentiation gives the density

$$f(t) = 4t \log\left(\frac{1}{t}\right). \qquad 0 < t < 1. \blacksquare$$

8.3 *Buffon's Needle Problem.

Buffon in 1777 proposed the following problem.

Buffon's Needle Problem. *A needle of unit length is dropped at random on a grid of parallel lines of unit width. What is the probability that it touches a line?*

Let X be the position of the center of the needle - that is, its height above the line of the grid below it, and θ the angle of the needle with the horizontal.

We shall assume that X and θ are *independent and uniformly distributed* on $0 \le X < 1$ and $0 \le \theta < 2\pi$ respectively. *This is an assumption about how the needle is thrown onto the grid.* For example, if one holds the needle parallel to the lines of the grid and throws it with little or no spin, the angle θ will not be uniformly distributed.

Under this assumption, the joint density is the constant $1/\pi$.

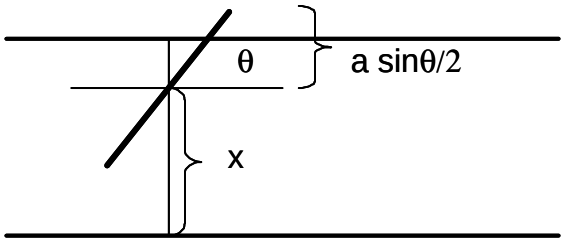

Figure 3.1. *Buffon's needle experiment.*

From Figure 3.1, we see that the needle touches the grid iff either

$$X - \frac{1}{2}\sin\theta < 0 \qquad \text{or} \qquad X + \frac{1}{2}\sin\theta$$

This region is shown in Figure 3.2.

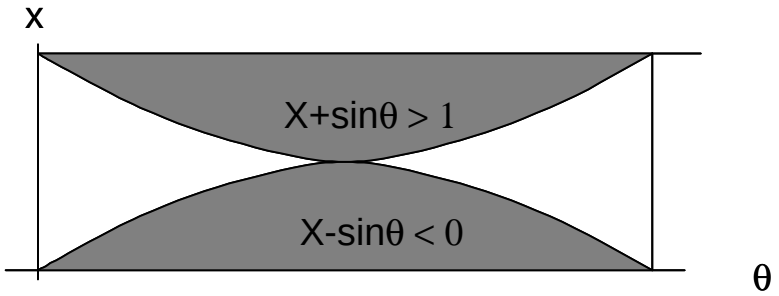

Figure 3.2.

The probability that the needle touches is therefore

$$2\int_0^\pi \frac{1}{2}\sin\theta\, d\theta = \frac{2}{\pi} = 0.6366.\blacksquare$$

It might occur to you that one could estimate the value of π experimentally in this manner.

This has indeed been tried, a sort of precursor of Monte Carlo calculations. Here are some results.

Investigator	**Date**	**Number of trials**	π **approximation**
Buffon	1777		
Wolf	1850	5000	3.1596
Smith	1855	3204	3.1553
Fox	1864	1100	3.1419
Lazzarini	1901	3408	3.15159292

The correct value of π is 3.141592654. If Lazzarini's result looks too good to be true, it is. See the article on Buffon's Needle on Wikipedia.

In the computer era, there have been numerous Monte Carlo simulations of Buffon's experiment, which in its way is a sort of precursor of Monte Carlo methods.

8.4 *Bertrand's Paradox.

The following question, discussed by Bertrand in 1889, illustrates an important fact about continuous r.v.

Bertrand's Problem. *A chord of a circle is selected at random. What is the probability that it is longer than the side of the inscribed equilateral triangle.*

Before proceeding, it may be good to remind the reader of a few facts about the inscribed equilateral triangle. By symmetry, the line from the center O of the circle to any vertex will bisect the angle of the triangle at that vertex. The angle $\angle OAP$ in Figure 4.1 is therefore 30^o, and so the segment $\overline{OP}$ has length

$$a \sin 30^o = a/2.$$

Armed with this fact, we proceed to the "solution" - or, more accurately, "solutions". Let $s = a\sqrt{3}$ be the length of the side.

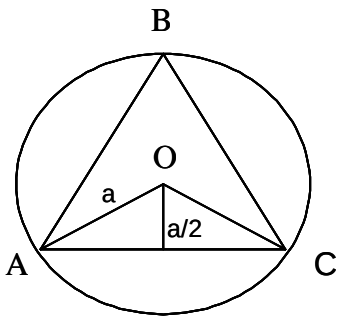

Figure 4.1. *Inscribed equilateral triangle.*

First solution. By symmetry, we may fix the direction of the chord in advance, so we may take the chord to be horizontal. Half the chords intersect the vertical within $a/2$ of the center, so $P = 1/2$.

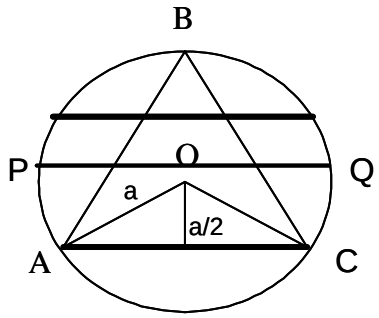

Figure 4.2. *Direction of chord fixed.*

Second Solution. By symmetry, we may fix one endpoint A of the chord in advance. Consider the angle θ at which the chord makes with the tangent at A. We have $0 \leq \theta \leq \pi$. Only those chords with $\pi/3 \leq \theta \leq 2\pi/3$ are longer than s, so $P = 1/3$.

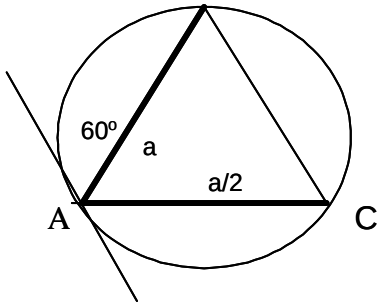

Figure 4.3. *Endpoint of chord fixed.*

Third Solution. Any point inside the circle can be the midpoint of a chord, and the midpoint determines the chord uniquely. The midpoints corresponding to chords longer than s are those within a distance $a/2$ of the center O. These midpoints occupy an area $\pi\,(a/2)^2$ equal to one quarter of the disc, so $P = 1/4$.

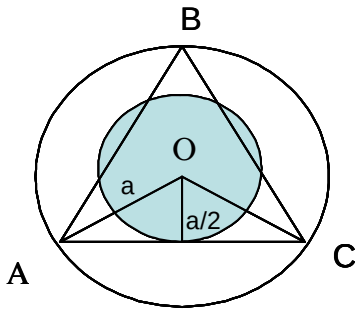

Figure 4.4. *Chord determined by midpoint.*

Three different answers. What is going on? The problem lies in the phrase "*a chord is selected at random*".

When we say that a ball is selected "*at random*" from an urn with a finite number n of balls, we mean that each ball has the same *positive* probability $1/n$ of being selected. But for continuous r.v. this doesn't work, because *all outcomes have probability zero*.

Thus, the statement "*a chord is selected at random*" has no precise meaning. The three solutions correspond to three different ways of selecting a chord *"at random*".

Physically, the three methods might be described as follows:

In case 1, a chord could be selected by rolling a bar down the vertical axis, and taking the chord across where it stops.

In case 2, a spinner could be placed on the point and spun, and the chord taken in the direction of the pointer when it stops.

In case 3, the circle could be placed on the wall and a dart cast randomly at it to determine the center.

In precise mathematical terms, we have assumed that:

In case 1, the y-component of the center is uniformly distributed on $[-a, a]$,

In case 2, the angle made with a tangent to the circle is uniformly distributed on $[-\pi, \pi]$, and

In case 3, the center is uniformly distributed on the disc.

8.5 *Sums of Independent r.v. and Convolutions.

Let X and Y be independent r.v. with densities $f(x)$ and $g(y)$ respectively, and let $S = X + Y$ be their sum. What is the density of S ?

The joint density of (X, Y) is

$$f(x, y) = f(x)g(y).$$

The c.d.f. of S is therefore

$$F(t) = P(X + Y \le t) = P(Y \le t - X) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-x} f(x)g(y)\, dy\, dx$$

Let $s = y + x$ in the inner integral to obtain

$$F(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t} f(x)g(s - x)\, ds\, dx = \int_{-\infty}^{t} \left(\int_{-\infty}^{\infty} f(x)g(s - x)\, dx \right) ds$$

Differentiating the c.d.f. gives the density

$$F'(t) = \int_{-\infty}^{\infty} f(x)g(t - x)\, dx.$$

Theorem 3. *If X and Y be independent r.v. with densities $f(x)$ and $g(y)$ respectively, then the sum $S = X + Y$ has the density*

$$f * g(t) = \int_{-\infty}^{\infty} f(x)g(t - x)\, dx. \tag{8.2}$$

The function is called the *convolution* of f and g.

Theorem 4. (**Properties of Convolution**.) *The convolution has the following properties.*

(a.) $(f + g) * h = (f * h) + (g * h)$.

(b.) If c is a constant, then $(cf) * g = c\,(f * g)$.

(c.) $f * g = g * f$.

(d.) $(f * g) * h = f * (g * h)$.

Proof. Parts (a.) and (b.) are clear. For part (c.), note that the density of $X + Y$ is the same as the density of $Y + X$. Similarly, for (d.), the density of $(X + Y) + Z$ is the same as the density of $X + (Y + Z)$. □

This proves parts (c.) and (d.) for densities. However, these properties hold for any functions, as long as the integrals exist.

Example 1. Let X and Y be independent $N(0, 1)$ r.v. Find the density of $X + Y$.

Solution. With $s = y - x/2$, we have

$$\begin{aligned} f * g(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-y)^2/2} e^{-y^2/2} dy = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left\{ -\left(y - \frac{x}{2}\right)^2 - \frac{x^2}{4} \right\} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left\{ -s^2 - \frac{x^2}{4} \right\} ds = \frac{1}{2\pi} e^{-x^2/4} \int_{-\infty}^{\infty} e^{-s^2} ds = \frac{1}{\sqrt{8\pi}} e^{-x^2/4}. \blacksquare \end{aligned}$$

The formula (8.2) takes a different form if X and Y are both positive. In this case, the densities $f(x)$ and $g(x)$ are zero for negative x. Hence, $f(x-y)g(y)$ is zero unless both $y > 0$ and $x - y > 0$, or in other words if $0 < y < x$. Thus (8.2) becomes

$$f * g(x) = \int_0^x f(x-y)g(y)dy.$$

This formula appears in the theory of Laplace Transforms.

Example 2. Let X and Y be independent exponential r.v., with mean 1. Find the density of $X + Y$.

Solution.

$$f * g(x) = \int_0^x e^{-(x-y)}e^{-y}dy = \int_0^x e^{-x}dy = xe^{-x}.\blacksquare$$

8.6 Transformation of Joint Densities. The Jacobian Method.

The Jacobian determinant arises in connection with change of variables in multiple integrals. Let

$$\begin{aligned} u &= u(x,y) \\ v &= v(x,y) \end{aligned}$$

be a continuously differentiable map from an open set S in the plane to another set G in the plane. The *Jacobian determinant* of the map is defined by

$$\frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} \partial u/\partial x & \partial v/\partial x \\ \partial u/\partial y & \partial v/\partial y \end{vmatrix} = \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}.$$

If the map

$$\begin{aligned} u &= u(x,y) \\ v &= v(x,y) \end{aligned}$$

is *one to one and onto* from S to G, then for any continuous function $f(u,v)$

$$\int\int_G f(u,v)dudv = \int\int_S f(u(x,y),v(x,y))\left|\frac{\partial(u,v)}{\partial(x,y)}\right|dxdy.$$

Hence, if (X, Y) have joint distribution $f(x, y)$ on S and $x = x(u, v)$, $y = y(u, v)$ is a one-one map of G onto S, then (U, V) have joint distribution

$$g(u, v) = f(x(u, v), y(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|.$$

on G.

An easy mnemonic is,

$$dudv = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| dxdy.$$

Important Remark. In using this method, you must be careful to understand the mapping completely; in particular, you must

(a.) know what S and G are, and

(b.) be sure you have a one-one map.

The tendency to just plug in and compute must be resisted.

Example 1. Let X and Y be independent $N(0, 1)$ r.v.

(a.) Find the joint distribution of the polar coordinates $R = \sqrt{x^2 + y^2}$, and $\Theta = \arctan(y/x)$.

(b.) Find the distribution of R.

Solution. (a.) The joint density of (X, Y) is

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$$

The mapping

$$\begin{aligned} r &= \sqrt{x^2 + y^2} \\ \theta &= \arctan\left(\frac{y}{x}\right). \end{aligned}$$

is a one-one mapping of $-\infty < x < \infty$, $-\infty < y < \infty$ (minus the negative x-axis) onto $0 < r < \infty$, $-\pi < \theta < \pi$. Solving for x and y in terms of r and θ , we find that

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta. \end{aligned}$$

Hence the Jacobian is

$$\frac{\partial(x, y)}{\partial(r, \theta)} = \begin{vmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{vmatrix} = r.$$

Thus the joint density of R and Θ is

$$g(r, \theta) = \frac{r}{2\pi} e^{-r^2/2}$$

on $0 < r < \infty$, $-\pi < \theta < \pi$.

More simply, we can just write

$$\frac{1}{2\pi} e^{-(x^2+y^2)/2}\, dxdy = \frac{1}{2\pi} e^{-r^2/2}\, rdrd\theta$$

(b.) The density of R is just the marginal distribution

$$\int_{-\pi/2}^{\pi/2} g(r,\theta)\, d\theta = re^{-r^2/2} \qquad 0 < r < \infty. \blacksquare$$

Example 2. *Find the density of the ratio of two independent exponentials X and Y, with $a = 1$.*

Solution. The joint density is

$$f(x,y) = e^{-(x+y)}$$

on $\mathbb{R}^2_+ = \{(x,y) : 0 < x < \infty, 0 < y < \infty\}$.

Let

$$\begin{aligned} u &= \frac{x}{y} \\ v &= x \end{aligned}$$

so that

$$\begin{aligned} y &= \frac{v}{u} \\ x &= v \end{aligned}$$

This mapping maps $\mathbb{R}^2_+$ one-one onto $\mathbb{R}^2_+$, with Jacobian

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} 0 & -v/u^2 \\ 1 & 1/u \end{vmatrix} = \frac{v}{u^2}.$$

Thus the joint density of (U,V) is

$$g(u,v) = \frac{v}{u^2} e^{-(v+\frac{v}{u})}.$$

The density of $U = X/Y$ is the marginal

$$\begin{aligned} f(u) &= \int_0^\infty \frac{v}{u^2} e^{-(v+\frac{v}{u})}\, dv = \frac{1}{u^2}\int_0^\infty ve^{-v(1+\frac{1}{u})}\, dv \\ &= \frac{1}{u^2}(1+\frac{1}{u})^{-2}\int_0^\infty se^{-s}ds = \Gamma(2)\frac{1}{(u+1)^2} \\ &= \frac{1}{(u+1)^2}. \end{aligned}$$

where $s = v(1 + \frac{1}{u})$.

This is a *Pareto density*. It has infinite expectation.■

Additional Remark. Suppose, more generally, that X is $\Gamma(1, \alpha)$ and Y is $\Gamma(1, \beta)$. Their joint density is

$$f(x, y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} y^{\beta-1} e^{-(x+y)} \qquad x.y > 0.$$

By the same transformation, $U = X/Y$ and $V = X$ have the joint density

$$g(u, v) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \frac{v^{\alpha+\beta-1}}{u^{\beta+1}} e^{-(v+\frac{v}{u})}.$$

The density of U is therefore

$$\begin{aligned} f(u) &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty \frac{v^{\alpha+\beta-1}}{u^{\beta+1}} e^{-(v+\frac{v}{u})} dv \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \frac{1}{u^{\beta+1}} (1 + \frac{1}{u})^{-(\alpha+\beta)} \int_0^\infty s^{\alpha+\beta-1} e^{-s} dv \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{u^{\alpha-1}}{(u+1)^{\alpha+\beta}} \qquad u > 0. \end{aligned}$$

where again $s = v(1 + \frac{1}{u})$. We will refer to this formula later in connection with R. A. Fisher's F distribution.■

Bonus Integral. Since this is a density, its integral must be one. This implies that

$$\int_0^\infty \frac{u^{\alpha-1}}{(u+1)^{\alpha+\beta}} du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = B(\alpha, \beta).$$

The next example shows the necessity of paying attention to domains.

Example 3. *As above, let X and Y be independent exponentials with $a = 1$. Find the density of $X + Y$.*

Solution.The joint density is.

$$f(x, y) = e^{-(x+y)}$$

on $\mathbb{R}^2_+$. Let

$$\begin{aligned} u &= x + y \\ v &= x \end{aligned}$$

so that

$$\begin{aligned} x &= v \\ y &= u - v. \end{aligned}$$

The Jacobian is

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

This is negative, so we must take its absolute value. The joint density of (u,v) is therefore

$$g(u,v) = e^{-u}$$

Without paying attention to domains, we get the "density"

$$f(u) = \int_0^\infty e^{-u} dv = e^{-u} \int_0^\infty dv = e^{-u} \cdot \infty.$$

The problem is that *the domain of* $g(u,v)$ - the set where it is not zero - *is not* $\mathbb{R}^2_+$. We have, rather $0 < v < \infty$, but $u = x + y > x = v > 0$. Thus the domain of $g(u,v)$ is

$$G : 0 < v < u < \infty.$$

which is shown in Figure 6.1

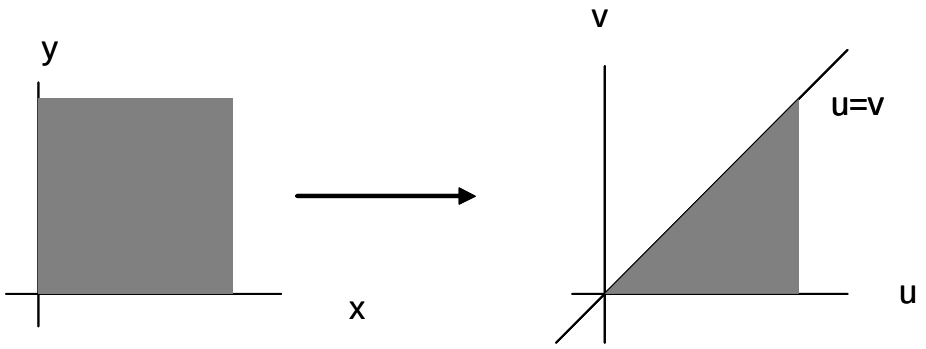

Figure 6.1

The integral for $f(u)$ should therefore be

$$f(u) = \int_0^u e^{-u} dv = u e^{-u}.$$

This is the distribution $\Gamma(1,2)$. ■

8.7 The Bivariate Normal Density.

One of the most important joint distributions, especially in Statistics is the Bivariate Normal distribution. It describes two normal r.v. which are, in general, not independent but are correlated in a special way.

The *Standard Bivariate Normal density*, in which both marginals are $N(0,1)$ is defined as follows For $-1 < r < 1$, define

$$f(x,y) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\left(x^2+y^2-2rxy\right)/2\left(1-r^2\right)\right]$$

For $r = 0$, the Bivariate Normal density reduces to the joint density of two *independent* normal r.v.

Theorem 5. *For the Bivariate Normal density,*

(a.) The marginal densities of X and Y are both $N(0,1)$.

(b..) $E(XY) = r$.

(c.) X and Y are independent iff $E(XY) = 0$.

Proof. By completing the square, we may write

$$x^2+y^2-2rxy = \left(1-r^2\right)x^2 + (y-rx)^2$$

so that

$$f(x,y) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi\left(1-r^2\right)}} e^{-(y-rx)^2/2(1-r^2)}$$

For (a.), the marginal density of X is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi\left(1-r^2\right)}} \int_{-\infty}^{\infty} e^{-(y-rx)^2/2(1-r^2)}dy \\ &= \frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi\left(1-r^2\right)}} \int_{-\infty}^{\infty} e^{-s^2/2(1-r^2)}ds = \frac{e^{-x^2/2}}{\sqrt{2\pi}}. \end{aligned}$$

where $s = y - rx$.

For (b.), we have

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)dy \\ &= \int_{-\infty}^{\infty} x\frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi\left(1-r^2\right)}} \int_{-\infty}^{\infty} \left[(y-rx)+rx\right] e^{-(y-rx)^2/2(1-r^2)}dy\, dx \\ &= \int_{-\infty}^{\infty} x\frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi\left(1-r^2\right)}} \int_{-\infty}^{\infty} \left[s+rx\right] e^{-s^2/2(1-r^2)}ds\, dx \\ &= \int_{-\infty}^{\infty} rx^2\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx = r\int_{-\infty}^{\infty} x^2\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx = r. \end{aligned}$$

For (c.), simply note that if $r = 0$, the joint density reduces to that of two independent normals. □

The General Bivariate Normal.

$$V = \frac{X - \mu_1}{\sigma_1} \quad \text{and} \quad W = \frac{Y - \mu_2}{\sigma_2}$$

In general, the joint distribution of two r.v. X and Y is Bivariate Normal iff the two normalized r.v.

$$V = \frac{X - \mu_1}{\sigma_1} \quad \text{and} \quad W = \frac{Y - \mu_2}{\sigma_2}$$

have the standard density

$$f(v,w) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\frac{1}{2\left(1-r^2\right)}\left(v^2 + w^2 - 2rvw\right)\right]$$

The Jacobian of the transformation

$$\begin{aligned} v &= \frac{x - \mu_1}{\sigma_1} \\ w &= \frac{y - \mu_2}{\sigma_2} \end{aligned}$$

is

$$\frac{\partial(v,w)}{\partial(x,y)} = \frac{1}{\sigma_1\sigma_2}$$

so that the general Bivariate Normal density is given by

$$\begin{aligned} & f(\frac{x-\mu_1}{\sigma_1}, \frac{y-\mu_2}{\sigma_2})\frac{\partial(v,w)}{\partial(x,y)} \\ = \quad & \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\frac{1}{2\left(1-r^2\right)}\left(\frac{\left(x-\mu_1\right)^2}{\sigma_1^2} + \frac{\left(y-\mu_2\right)^2}{\sigma_2^2} - \frac{2r\left(x-\mu_1\right)\left(y-\mu_2\right)}{\sigma_1\sigma_2}\right)\right] \end{aligned}$$

Corollary 1. *The covariance of X and Y is*

$$cov(X,Y) = r\,\sigma_1\sigma_2$$

and hence the coefficient of regression is r.

Proof.

$$\begin{aligned} cov(X,Y) &= E\left(\left(X-\mu_1\right)\left(Y-\mu_2\right)\right) = E((\sigma_1 V)(\sigma_2 W)) \\ &= \sigma_1\sigma_2 E(VW) = r\sigma_1\sigma_2.\square \end{aligned}$$

8.8 *Chi-square, t and F.

The Normal distribution gives rise to three standard distributions that are widely used in Statistics: the *Chi-square* distribution of Karl Pearson, *Student's* t distribution due to W. S. Gossett, and the F distribution of Snedecor and R. A. Fisher. We will not enter here into the various uses of these distributions, but will indicate what they are, and derive formulas for their densities.

Roughly speaking, these distributions may be described as follows:

(1.) The Chi-square is the distribution of the sum of the squares of independent standard normals.;

(2.) The F is (almost) the distribution of a quotient of Chi-squares, and

(3.) The t is (almost) the distribution of a normal divided by a chi-square.

The 'almost' means that the chi-squares are divided by their means, so that they are normalized to mean 1.

We will take these distributions up one by one.

Chi-square Distribution.

Definition 2. The Chi-square distribution with r degrees of freedom, $\chi^2(r)$, is the distribution of

$$X = Z_1^2 + Z_2^2 + \cdots + Z_r^2$$

where Z_1, Z_2, ...,Z_r are independent Standard normal r.v.

Since the square Z^2 of a standard normal Z has the Gamma distribution $\Gamma(\frac{1}{2}, \frac{1}{2})$, the distribution $\chi^2(r)$ is just the gamma distribution $\Gamma(\frac{r}{2}, \frac{1}{2})$, the density of which is

$$f(x) = \frac{1}{\Gamma(r/2)} \left(\frac{1}{2}\right)^{r/2} x^{r/2-1} e^{-x/2} \qquad x > 0.$$

Since the gamma $\Gamma(a, \nu)$has mean ν/a and variance ν/a^2, the mean and variance of $\chi^2(r)$ are $\mu = r$ and $\sigma^2 = 2r$.

F Distribution.

Note that if X_r is $\chi^2(r)$, then X_r/r has mean 1.

Definition 3. The F-distribution with degrees of freedom r and s is the distribution of

$$F_{r,s} = \frac{X_r/r}{X_s/s}$$

where X_r and X_s are independent r.v. with Chi-square distributions of r and s degrees of freedom respectively.

The factors of r and s appear to be more natural if we note that X_r/r *has mean one*.

In Example 2 of section 8.7, we found the density of the quotient of a $\Gamma(1,\alpha)$ r.v. by an independent $\Gamma(1,\beta)$ to be

$$f(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{u^{\alpha-1}}{(u+1)^{\alpha+\beta}} \qquad u > 0. \tag{8.3}$$

The $F-$density can be obtained from this by scaling. In the first place, since X_r is $\Gamma(\frac{1}{2}, \frac{r}{2})$, $X_r/2$ is $\Gamma(1, \frac{r}{2})$. Thus we can write

$$F = \frac{X_r/r}{X_s/s} = \frac{s}{r}\frac{X_r/2}{X_s/2} = \frac{s}{r}U$$

where the density of

$$U = \frac{X_r/2}{X_s/2}$$

is given by (8.3), with $\alpha = r/2$ and $\beta = s/2$ .By scaling, the density of F is

$$\frac{s}{r} f(\frac{r}{s}u)$$

with $\alpha = r/2$ and $\beta = s/2$. This works out to

$$f(u) = \frac{\Gamma(\frac{r+s}{2})}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \left(\frac{r}{s}\right)^{r/2} \frac{u^{(r-2)/2}}{(1+\frac{ru}{s})^{(r+s)/2}} \qquad u > 0.$$

Student's t Distribution.

Definition 4. *Student's t distribution* with r degrees of freedom is the distribution of

$$T = \frac{Z}{\sqrt{X_r/r}} = \sqrt{\frac{r}{2}}\frac{Z}{\sqrt{X_r/2}}$$

where Z has a Standard normal distribution, and X_r is an independent r.v. with $\chi^2(r)$ distribution.

To obtain the t density, we first prove:

Theorem 6. *Let X be $\Gamma(1,p)$, and Z be $N(0,1)$. Then the density of $Y = Z/\sqrt{X}$ is*

$$f(y) = \frac{\Gamma(p+\frac{1}{2})}{\Gamma(p)} \left(1 + y^2/2\right)^{-(p+1/2)}$$

Proof. The joint density of (Z, X) is

$$f(z,x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\Gamma(p)} x^{p-1} e^{-x} e^{-z^2/2} \qquad x > 0,\ -\infty < z < \infty.$$

Let

$$\begin{aligned} y &= z/\sqrt{x} \qquad -\infty < y < \infty \\ s &= x \qquad\qquad s > 0. \end{aligned}$$

Then

$$\begin{aligned} x &= s & x > 0 \\ z &= y\sqrt{s} & -\infty < z < \infty. \end{aligned}$$

The Jacobian is

$$\frac{\partial(z,x)}{\partial(y,s)} = \begin{vmatrix} 1 & y/2\sqrt{s} \\ 0 & \sqrt{s} \end{vmatrix} = \sqrt{s}$$

so that joint density of (y,s) is

$$f(y,s) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(p)} s^{p-1} e^{-s} e^{-sy^2/2} \sqrt{s}$$

and the $y-$marginal distribution of $Y = Z/\sqrt{X}$ is

$$\begin{aligned} f(y) &= \int_0^\infty f(y,s)ds = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(p)} \int_0^\infty s^{p-1/2} e^{-s\left(1+y^2/2\right)} ds \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(p)} \left(1+y^2/2\right)^{-(p+1/2)} \int_0^\infty v^{p-1/2} e^{-v} dv \\ &= \frac{1}{\sqrt{2\pi}} \frac{\Gamma(p+1/2)}{\Gamma(p)} \left(1+y^2/2\right)^{-(p+1/2)} \end{aligned}$$

where $v = s\left(1+y^2/2\right)$. □

Again a distribution must integrate to one, so we obtain, with a little algebra,

Corollary 2. (*A Bonus Integral*.)

$$\int_{-\infty}^{\infty} \frac{1}{(1+u^2)^p} du = \sqrt{\pi} \frac{\Gamma(p-1/2)}{\Gamma(p)}.$$

Corollary 3. *The t-distribution with r degrees of freedom has the density*

$$g(t) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma((r+1)/2)}{\Gamma(r/2)} \frac{1}{(1+t^2/r)^{(r+1)/2}}.$$

Proof. This follows by scaling by $c = \sqrt{r/2}$, and setting $a = r/2$:

$$g(t) = \sqrt{2/r} f(\sqrt{2/r}t) = \frac{1}{\sqrt{r\pi}} \frac{\Gamma((r+1)/2)}{\Gamma(r/2)} \frac{1}{(1+t^2/r)^{(r+1)/2}}. \square$$

The t distribution was first obtained in 1908 by the statistician W. S. Gossett. Gossett was employed by the Guinness brewery, who did not want it to get about that they were using statistical methods, so Gossett published his papers under the pseudonym "*A. Student*", whence the name "*Student's t*".

8.9 Problems.

(1.) Let the random point (X, Y) be uniformly distributed over the unit square. Are X and Y independent? What if (X, Y) is uniformly distributed over the unit disc?

(2.) (*Rayleigh distribution*.) Let X and Y be independent standard normal r.v., and let (R, θ) be polar coordinates in (X, Y)-space.

(a.) Prove that R and θ are independent.

(b.) Find the densities of R and θ. (The density of R is called the *Rayleigh distribution*.)

(.3.) At a random time X during a 24-hour day, a light is turned on. At a random time Y between X and midnight, it is turned off.

(a.) Find the joint density of X and Y.

(b.) Find the probability that the light was on more than half the day.

(4.) The joint probability density function of X and Y is given by

$$f(x, y) = c(y^2 - x^2)e^{-y}, \quad -y \leq x \leq y,\ 0 < y < \infty.$$

(a.) Find c.

(b.) Find the marginal densities of X and Y.

(5.) (*Yuletide Special*) If a spherical plum pudding of radius a contains n indefinitely small plums, compute the expected distance of the nearest one from the surface of the pudding.

(6.) Two points are chosen at random from the interval $[0, a]$ (i.e., their abscissas are independent and uniformly distributed on $[0, a]$). Find the distribution of the distance between the two points, its mean and variance.

(7.) Suppose that the joint density of the two-dimensional random variable (X, Y) is given by

$$\begin{aligned} f(x, y) &= x^2 + \frac{xy}{3}, && 0 < x < 1,\ 0 < y < 2, \\ &= 0 && \text{elsewhere.} \end{aligned}$$

Compute the following:

(a.) $P\left(X > \frac{1}{2}\right)$.

(b.) $P(Y < X)$.

(c.) Find the marginal densities.

(8.) Let the random variables X and Y have the joint density

$$f(x,y) = \frac{1}{4}\left(1 + \frac{xy}{2}\right)$$

for $-1 \leq x, y \leq 1$.

(a.) Find the marginal densities of X and Y.

(b.) Are X and Y independent?

(c.) Find $E(XY)$.

(d.) Find $P(XY > 0)$.

(9.) Solve Buffon's needle problem if $L > D$.

(10.) Let X and Y be independent and uniformly distributed on $[0,1]$. Find the mean and density of

(a.) XY.

(b.) $X + Y$.

(c.) $\log\left(\frac{1}{x}\right)$.

(d.) $\max(X,Y)$.

(e.) $\min(X,Y)$.

(f.) $|X - Y|$.

(11.) Let X and Y have the joint distribution

$$f(x,y) = \frac{1}{2}x^2 y e^{-(x+y)} \qquad x \geq 0, \quad y \geq 0.$$

(a.) What are the marginal distributions?

(b.) Are X and Y independent?

(c.) What is the distribution of $X + Y$?

(12.) Let X and Y be independent r.v., both uniformly distributed on the interval $[0,1]$. Find $P(Y \leq X^2)$.

(13.) Let the random variables X and Y have the joint density

$$\begin{aligned} f(x,y) &= 2 \qquad 0 \leq y \leq x \leq 1 \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

(Note that the density is non-zero on a *triangular set*, not a rectangle.)

(a.) Find $P(X > 2Y)$.

(b.) Find the marginal densities of X and Y.

(c.) Find $E(XY)$.

(d.) Find the density of $W = XY$.

(14.) (*Buffon's Checkerboard Problem.*) A coin of diameter d is thrown at random onto a checker board of squares of side $s > d$. What is the probability that the coin is completely contained in a single square?

(15.) (*A Honeycomb Problem.*) (a.) Solve Buffon's Checkerboard Problem for a hexagonal tiling of the plane.

(b.) Compare the results for square and hexagonal tilings of the same density; that is for which the squares and hexagons have the same area. For which is the probability of the coin lying within a single figure largest?

(16.) In Bertrand's Paradox, find the distribution of the length of the chord in each of the three cases.

(17.) An interval of is divided into two intervals of length a and b respectively, with $b < a$. Points x and y are selected at random in each, dividing the interval into three pieces. What is probability they form a triangle?

(18.) If X and Y are independent with same *even* density, prove that

$$P(X > Y \mid X > 0) = \frac{3}{4}.$$

(19.) Find the densities for $X + Y$ and $X - Y$ if X has density ae^{-ax} ($X > 0$) and the density of Y equals h^{-1} for $0 < x < h$.

(20.) Let X and Y be independent r.v., both having the exponential density ae^{-ax}, $x \geq 0$ Find the density for $X - Y$.

(21.) Let X and Y be independent exponential variables with mean $\frac{1}{a}$. Find the density of $W = Y/X$.

(22.) Let X and Y be independent exponential variables with mean $\frac{1}{a}$. Find the density of $W = |X - Y|$. Explain your answer.

(23.) In measuring T, the life length of an item, an error may be made which may be assumed to be uniformly distributed over $(-0.01, 0.01)$. Thus the recorded time (in hours) may be represented as $T + X$, where T has an exponential distribution wit parameter 0.2 and X has the above-described uniform distribution. If T and X are independent, find the p.d.f. of $T + X$.

(24.) Let X represent the life length of an electronic device and suppose that X is a continuous random variable with density

$$\begin{aligned} f(x) &= a/x^2, \quad x > a; \\ &= 0, \quad \text{otherwise} \end{aligned}$$

Let X_1 and X_2 be two independent determinations of the above random variable X.

(That is, suppose that we are testing the life length of two such devices.) Find the density of the random variable $Z = X_1/X_2$.

(25.) If X is uniformly distributed over $(0, 1)$ and Y is exponentially distributed with parameter $\lambda = 1$, find the distribution of (a) $Z = X + Y$, and (b) $Z = X/Y$. Assume independence of X and Y .

(26.) Let X and Y be positive, independent r.v. with densities $f(x)$ and $g(y)$. Find formulas for the densities of the r.v.

(a.) $Q = X/Y$.

(b.) $V = XY$.

(27.) Let $X_1, ..., X_n$ be independent exponential variables with mean $\frac{1}{a}$. Find the density of $W = \min(X_1, ..., X_n)$. Explain your answer.

(28.) Let X and Y be independent r.v., both uniformly distributed on the interval $[0, 1]$. Find the density of the r.v. $M = \max\{X, Y\}$, i.e. the maximum of X and Y.

(29.) Let X and Y be independent r.v., both uniformly distributed on the interval $[0, 1]$. Find the *range* and *density* of the random variable $Z = X/Y$. (*Hint*: Note that the formula for $t > 1$ differs from that for $t < 1$.)

(30.) Let Z_1 and Z_2 be independent normal random variables with mean 0, and variance 1. Find the distribution of $Y = Z_1^2 + Z_2^2$.

(31.) Let X and Y be independent normal random variables with mean 0, and variance 1. Find the distribution of the quotient Y/X. (*Hint* There are separate integrals for $X > 0$ and $X < 0$.)

(32) Let the random variable X and Y have the joint density

$$f(x, y) = \frac{1}{(1 + x)^2} \frac{1}{(1 + y)^2} \qquad x > 0, y > 0.$$

Find the joint density of the random variable .

$$U = \frac{Y}{X} \qquad \text{and} \qquad V = X.$$

(33.) If T has Student's t-distribution with r degrees of freedom, find the distribution of T^2.

Chapter 9
Moment Generating Functions.

9.1 Definition of the Moment Generating Function.

We shall now introduce a technical device known as the *moment generating function* that will be useful for manipulating distributions.

First, the definition.

Definition 1. Let X be a random variable. The *moment generating function* (m.g.f.) of X is the function of t, defined by

$$\phi(t) = E(e^{tX}).$$

As always, an expectation may or may not exist, so *some r.v. will not have m.g.f.'s*.

Example 1. Let X have the *Binomial distribution* $Bin(n,p)$. The moment generating function is

$$\phi(t) = E(e^{tX}) = \sum_{k=0}^{n} e^{tn}\binom{n}{k}p^k q^{n-k} = \sum_{k=0}^{n}\binom{n}{k}\left(e^t p\right)^k q^{n-k} = \left(e^t p + q\right)^n$$

by the Binomial Theorem.■

Example 2. Let Z have the *Standard Normal distribution* $N(0,1)$. The moment generating function is

$$\phi_Z(t) = E(e^{tZ}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-x^2/2}dx.$$

Completing the square in the exponent gives

$$\begin{aligned}\phi_Z(t) &= e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\left(x-2xt+t^2\right)/2}dx = e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x-t)^2/2}dx \\ &= e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-s^2/2}dx = e^{t^2/2}.\blacksquare\end{aligned}$$

Example 3. Let N have the *Poisson distribution with mean* λ. The m.g.f. is

$$\phi_N(t) = E(e^{tN}) = \sum_{n=0}^{\infty} e^{tn}\frac{\lambda^n}{n!}e^{-\lambda} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\left(e^t\lambda\right)^n}{n!} = e^{-\lambda}\exp(\lambda e^t).\blacksquare$$

Example 4. As an example of a distribution which does not have an m.g.f., let Y have the *Cauchy density*

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

The moment generating function is defined as the integral

$$\phi(t) = E(e^{tY}) = \frac{1}{\pi}\int_{-\infty}^{\infty} e^{tx}\frac{1}{x^2+1}dx,$$

which fails to exist for $t \neq 0$.■

9.2 Moments.

There are *Five Facts* to remember about m.g.f's.

Fact # 1. Recall that the n^{th} *moment* μ_n *of* X is the expectation of X^n:

$$\mu_n = E(X^n).$$

The First Fact is that *the moment generating function of* X *generates the moments of* X, in the sense that the n^{th} derivative of the m.g.f. at $t = 0$ is the n^{th} moment of X.

Theorem 1 (*Fact # 1.*). *Let* $\phi(t)$ *be the m.g.f. of the r.v.* X. *Then*

$$\phi^{(n)}(0) = \mu_n.$$

Proof. Expand the exponential in series:

$$e^{tX} = \sum_{n=0}^{\infty}\frac{1}{n!}(tX)^n = \sum_{n=0}^{\infty}\frac{X^n}{n!}t^n$$

Taking the expectation gives

$$\phi(t) = E(e^{tX}) = \sum_{n=0}^{\infty}\frac{E(X^n)}{n!}t^n = \sum_{n=0}^{\infty}\frac{\mu_n}{n!}t^n.$$

Now recall *Taylor's formula* from Calculus

$$\phi(t) = \sum_{n=0}^{\infty}\frac{\phi^{(n)}(0)}{n!}t^n.$$

Comparing coefficients, we find that

$$\frac{\phi^{(n)}(0)}{n!} = \frac{\mu_n}{n!}$$

or

$$\phi^{(n)}(0) = \mu_n.\blacksquare$$

Fact # 1 lets us use the m.g.f. to compute the mean and variance.

Example 1. For the *Binomial distribution* $Bin(n,p)$, the moment generating function is

$$\phi(t) = \left(e^t p + q\right)^n$$

Its derivative is

$$\phi'(t) = n\left(e^t p + q\right)^{n-1} e^t p$$

and so we obtain for the mean

$$\mu = \mu_1 = \phi'(0) = np.$$

For the variance, we compute the second derivative to be

$$\phi''(t) = n(n-1)\left(e^t p + q\right)^{n-2}\left(e^t p\right)^2 + n\left(e^t p + q\right)^{n-1} e^t p$$

from which we obtain the second moment

$$\begin{aligned} \mu_2 &= \phi''(0) = n(n-1)p^2 + np = (np)^2 + np^2 - np \\ &= (np)^2 + npq \end{aligned}$$

and the variance

$$\sigma^2 = \mu_2 - \mu^2 = (np)^2 + npq - (np)^2 = npq.\blacksquare$$

Example 2. If N has the *Poisson distribution* with mean λ, the m.g.f. is

$$\phi_N(t) = e^{-\lambda} e^{\lambda e^t}.$$

The derivative is

$$\phi'(t) = e^{-\lambda} e^{\lambda e^t} \lambda e^t$$

and so we obtain for the mean

$$\mu = \mu_1 = \phi'(0) = \lambda.$$

For the variance, we compute the second derivative to be

$$\phi''(t) = e^{-\lambda} e^{\lambda e^t}\left(\lambda e^t\right)^2 + e^{-\lambda} e^{\lambda e^t} \lambda e^t$$

from which we obtain the second moment

$$\mu_2 = \phi''(0) = \lambda^2 + \lambda.$$

and the variance

$$\sigma^2 = \mu_2 - \mu^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.\blacksquare$$

Example 3. Let Z have the *Standard Normal distribution* $N(0,1)$. The moment generating function is

$$\phi(t) = e^{t^2/2}$$

We compute that

$$\begin{aligned} \phi'(t) &= te^{t^2/2} \\ \phi''(t) &= \left(t^2+1\right)e^{t^2/2} \end{aligned}$$

from which we obtain

$$\mu = \mu_1 = \phi'(0) = 0.$$

and

$$\begin{aligned} \mu_2 &= \phi''(0) = 1 \\ \sigma^2 &= \mu_2 - \mu^2 = 1. \end{aligned}$$

In fact, *we can in fact easily find all the moments of* Z by comparing coefficients in two series. We have

$$\phi_Z(t) = \sum_{n=0}^{\infty} \frac{\mu_n}{n!} t^n = e^{t^2/2} = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{t^{2k}}{2^k}.$$

It follows that

$$\mu_{2k+1} = 0.$$

and

$$\frac{\mu_{2k}}{(2k)!} = \frac{1}{k!2^k}$$

so that

$$\mu_{2k} = \frac{(2k)!}{k!2^k}$$

If we note that the product of the *even* integers up to $2k$ is

$$2 \cdot 4 \cdot \cdots \cdot (2k) = k!2^k$$

it follows that

$$\mu_{2k} = \frac{(2k)!}{k!2^k} = 1 \cdot 3 \cdot \cdots \cdot (2k-1) = (2k-1)!!.$$

Thus, for example,

$$\mu_8 = 1 \cdot 3 \cdot 5 \cdot 7 = 105.\blacksquare$$

9.3 Translation and Scaling.

Fact # 2. The second fact tells how the m.g.f. behaves under translation and scaling.

Theorem 2. (*Fact # 2.*) *If the r.v. X has the m.g.f. $\phi(t)$, then*

(a.) $X + a$ has m.g.f. $\phi(t)e^{ta}$ and

(b.) cX has m.g.f. $\phi(ct)$.

Proof. For the proof , we have

$$\phi_{X+a}(t) = E(e^{t(X+a)}) = E(e^{tX}e^{ta}) = e^{ta}E(e^{tX}) = e^{ta}\phi_X(t)$$

and

$$\phi_{cX}(t) = E(e^{t(cX)}) = E(e^{(ct)X}) = \phi_X(ct).\square$$

Example 1. We have seen that the general Normal distribution $N(\mu, , \sigma^2)$ is the distribution of $X = \sigma Z + \mu$ where Z has the distribution $N(0, 1)$. As computed above, Z has the m.g.f,

$$\phi_Z(t) = e^{t^2/2}$$

so, by *(b.)*, σZ has the m.g.f.

$$\phi_{\sigma Z}(t) = e^{\sigma^2 t^2/2}$$

and hence, by *(a.)*, $X = \sigma Z + a$ has the m.g.f

$$\phi_X(t) = \phi_{\sigma Z+\mu}(t) = e^{\sigma^2 t^2/2}e^{\mu t}.\blacksquare$$

9.4 Independence.

Fact #3. The third fact gives the m.g.f. of the *sum of independent r.v.*

Theorem 3. (*Fact #3.*) *If X and Y are independent r.v. with m.g.f.'s $\phi_X(t)$ and $\phi_Y(t)$ respectively, then $X + Y$ has m.g.f.*

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$

Proof. If X and Y are independent, then so are the r.v. e^{tX} and e^{tY}, so

$$\phi_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY}) = \phi_X(t)\phi_Y(t).\square$$

Example 1. (*m.g.f. of the Binomial distribution.*) Let S be an event with probability p. The m.g.f. of 1_S is

$$E(e^{t1_S}) = e^{t\cdot 0}q + e^{t\cdot 1}p = q + e^t p.$$

As we noted before, the number X of Successes in n independent trials, which has the Binomial distribution $Bin(n,p)$, may be written as

$$X = 1_{S_1} + 1_{S_2} + \cdots + 1_{S_n}$$

where S_k is Success on the k^{th} trial. All the 1_{S_k} have the same m.g.f. $(q + e^t p)$, so by Fact # 3, the m.g.f. if X is

$$\left(q + e^t p\right)^n .\blacksquare$$

Fact # 4. The fourth fact says that the *m.g.f. determines the distribution*.

Theorem 4. (*Fact # 4.*) *If two r.v. have the same m.g.f., then they have the same distribution.*

The proof somewhat difficult, and will not be given.

Example 2. Let X and Y be two independent normal r.v., where X is $N(\mu_1, \sigma_1^2)$ and Y is $N(\mu_2, \sigma_2^2)$. If $S = X + Y$, then S has m.g.f.

$$\phi_S(t) = \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = e^{\sigma_1^2 t^2/2}e^{\mu_1 t}\cdot e^{\sigma_2^2 t^2/2}e^{\mu_2 t} = e^{\left(\sigma_1^2+\sigma_2^2\right)t^2/2}e^{(\mu_1+\mu_2)\mu t}.$$

This is the m.g.f. of the normal distribution $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. By Fact # 4, S must actually have this distribution. We obtain, therefore,

Theorem 5. *If X and Y are independent, and X is $N(\mu_1, \sigma_1^2)$ and Y is $N(\mu_2, \sigma_2^2)$, then $S = X + Y$ is $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Remark. Note that we already know from the elementary properties of expectation that $X + Y$ has mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Thus the essential content of Theorem 5 is that the *sum of independent normal r.v. is normal*.

9.5 The Continuity Theorem.

Fact # 5. We are now ready to discuss the *Fifth Fact* about the m.g.f. - the *Continuity Theorem*.

Theorem 6. (*Fact # 5. The Continuity Theorem.*). *Let X be a r.v. with m.g.f.'s $\phi(t)$, and c.d.f. $F(t)$. Let X_n be a sequence of r.v. with m.g.f.'s $\phi_n(t)$, and c.d.f. $F_n(t)$.*
If $\phi_n(t) \to \phi(t)$, then c.d.f. $F_n(t) \to F(t)$.

The proof is beyond the scope of this course. We shall proceed with a few applications.

Example 1.(*Poisson's approximation to the Binomial*.) We have shown directly that the Binomial distribution $Bin(n,p)$ with $p=\lambda/n$ is approximated for large n by the Poisson distribution with mean λ. Let's prove that using the Continuity Theorem.

The distribution $Bin(n,p)$ has the m.g.f.

$$\left(q+pe^t\right)^n=\left(1+p\left(e^t-1\right)\right)^n.$$

Setting $p=\lambda/n$ gives

$$\phi_n(t)=\left(1+\frac{\lambda\left(e^t-1\right)}{n}\right)^n\to e^{-\lambda}e^{\lambda e^t}.$$

But

$$\phi(t)=e^{-\lambda}e^{\lambda e^t}$$

is the m.g.f. of the Poisson distribution with mean λ.■

For the next two examples, we will need the limit

$$\lim_{n\to\infty}n\left(e^{t/n}-1\right)=\lim_{n\to\infty}t\frac{e^{t/n}-1}{t/n}=t\lim_{x\to0}\frac{e^x-1}{x}=t. \tag{9.1}$$

Example 2. (*The exponential distribution as limit of geometric distributions*.) Let T be geometric with probability p of Success. The m.g.f. of T is

$$\phi_T(t)=\frac{pe^t}{(1-qe^t)}.$$

Let $p=a/n$, so that the expected wait is

$$E(T_n)=\frac{n}{a}.$$

Let $Y_n=T_n/n$. The r.v. Y_n is a discrete waiting time with time step $1/n$, and fixed mean equal to $1/a$. The m.g.f. of Y_n is

$$\phi_{Y_n}(t)=\frac{\frac{a}{n}e^{t/n}}{\left(1-e^{t/n}+\frac{a}{n}e^{t/n}\right)}.$$

Using (9.1), we have

$$\lim_{n\to\infty}\phi_{Y_n}(t)=\lim_{n\to\infty}\frac{ae^{t/n}}{\left(ae^{t/n}-n\left(e^{t/n}-1\right)\right)}=\frac{a}{a-t}$$

which is the m.g.f. of the exponential distribution $\Gamma(a,1)$.■

Example 3. (*The uniform distribution as the limit of discrete uniform distributions*.)

If X is uniform on $[0,1]$, it has m.g.f.

$$\phi(t)=\int_0^1e^{tx}dx=\frac{e^t-1}{t}.$$

Let X_n take on the values $\frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1$, all with equal probabilities $\frac{1}{n}$. The m.g.f. of X_n is

$$\phi_n(t) = \sum_{k=1}^{n} e^{tk/n} \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} \left(e^{t/n}\right)^k = \frac{e^{t/n}}{n} \frac{(1-e^t)}{1-e^{t/n}} \to \frac{(1-e^t)}{t}$$

where we have used the formula

$$1 + x + x^2 + \cdots + x^{n-1} = \frac{1-x^n}{1-x}. \blacksquare$$

Thus, by (9.1),

$$\lim_{n\to\infty} \phi_n(t) = \lim_{n\to\infty} e^{t/n} \frac{(e^t-1)}{n\left(e^{t/n}-1\right)} = \frac{(e^t-1)}{t}$$

which is the m.g.f. of the uniform distribution on $[0,1]$.$\blacksquare$

Example 4. *The Poisson distribution for large mean is approximately Gaussian.*

We will show that for large λ, the distribution $Poi(\lambda)$ is approximately $N(\lambda, \sqrt{\lambda})$. The m.g.f. of $Poi(\lambda)$ is

$$\phi(t) = e^{-\lambda} e^{\lambda e^t}$$

By translation, $N - \lambda$ has m.g.f.

$$e^{-\lambda t}\phi(t) = e^{-\lambda} e^{-\lambda t} e^{\lambda e^t}$$

and by scaling,

$$Z_\lambda = \frac{N-\lambda}{\sqrt{\lambda}}$$

has m.g.f.

$$\phi_\lambda(t) = e^{-\lambda} e^{-\lambda t/\sqrt{\lambda}} e^{\lambda e^{t/\sqrt{\lambda}}}$$

We want the limit as $\lambda \to \infty$. Take the logarithm:

$$\begin{aligned}
\log \phi_\lambda(t) &= -\lambda - \sqrt{\lambda}t + \lambda e^{t/\sqrt{\lambda}} \\
&= -\lambda - \sqrt{\lambda}t + \lambda \left[1 + \frac{t}{\sqrt{\lambda}} + \frac{1}{2}\frac{t^2}{\lambda} + O(\frac{1}{\lambda^{3/2}})\right] \\
&= -\lambda - \sqrt{\lambda}t + \lambda + \sqrt{\lambda}t + \frac{1}{2}t^2 + O(\frac{1}{\sqrt{\lambda}}) \\
&= \frac{1}{2}t^2 + O(\frac{1}{\sqrt{\lambda}}) \to \frac{1}{2}t^2
\end{aligned}$$

as $\lambda \to \infty$. Exponentiating gives $\phi_\lambda(t) \to e^{t^2/2}$.

This approximation is good for $\lambda \geq 10$, within the range $n = \lambda \pm c\sqrt{\lambda}$, provided that a continuity correction is used.$\blacksquare$

Example 5. (*The Weak Law of Large Numbers.*)
If the m.g.f. of a r.v. X is

$$\phi(t) = 1 + \mu t + \frac{\sigma^2}{2}t^2 + \cdots$$

then the m.g.f. of the sample mean

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is

$$[\phi(t)]^n = \left[1 + \frac{\mu t}{n} + \frac{\sigma^2}{2n^2}t^2 + \cdots\right]^n \to e^{\mu t}.$$

But $e^{t\mu}$ is the m.g.f. of the distribution of the constant r.v. $M = \mu$.■

We have used the following Lemma.

Lemma. *If* $\epsilon_n \to 0$, *then*

$$\lim_{n\to\infty}\left(1 + \frac{x + \epsilon_n}{n}\right)^n = e^x$$

Proof. Since

$$\lim_{x\to 0}\frac{\log(1+x)}{x} = 1$$

we have

$$\begin{aligned}\log\left(1 + \frac{x+\epsilon_n}{n}\right)^n &= n\log\left(1 + \frac{x+\epsilon_n}{n}\right)\\ &= (x+\epsilon_n)\left(\frac{n}{x+\epsilon_n}\right)\log\left(1 + \frac{x+\epsilon_n}{n}\right) \to x.\square\end{aligned}$$

9.6 *Characteristic Functions.

We have not emphasized technical matters in this text, but a few remarks about the existence of m.g.f's may not be out of place.

As we have remarked, by no means all random variables will have m.g.f.'s. In fact, since the coefficients of the m.g.f. are determined by the moments of X, it follows that all moments of X must exist. For example, the Cauchy distribution of Example 4 of section 1 has no finite moments, not even a mean.

Existence of all moments is a very strong condition, since it means that the probabilities of large values of X must be very small. If X has density $f(x)$, the integral

$$E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

must converge for at least for some $t > 0$. This means roughly that $f(x)$ needs to go to zero exponentially at infinity. To be precise, it if $\phi(a)$ is finite for some a, then $P(X > t)$ must

go to zero exponentially, because

$$\begin{aligned} P(X &> t) = \int_t^\infty f(x)dx = \int_t^\infty e^{-at}e^{at}f(x)dx \\ &\leq e^{-at}\int_t^\infty e^{ax}f(x)dx \leq e^{-at}\int_{-\infty}^\infty e^{ax}f(x)dx = e^{-at}\phi(a). \end{aligned}$$

Fortunately, many useful distributions have this property. In particular, all *bounded* r.v. will have m.g.f.'s.

It is actually quite possible to define a substitute for the m.g.f. which works for *every* r.v. It is called the *characteristic function* $\psi(t)$ of X, and is obtained by replacing t in the m.g.f. by it where $i^2 = -1$. Thus, the characteristic function of X is

$$\psi(t) = E(e^{itX}).$$

Since $\left|e^{itX}\right| = 1$, this is the expectation of a bounded r.v., and so will always exist.

If the m.g.f. exists. we have

$$\psi(t) = \phi(it)$$

The analogues of the five facts hold.

Theorem 7. (Properties of the Characteristic Function.)

(a.) $\psi^{(n)}(0) = i^n\mu_n$, *provided that the moment exists.*

(b.) $\psi_{cX}(t) = \psi(ct)$

(c.) $\psi_{X+a}(t) = e^{iat}\psi(t)$

(d.) If X *and* Y *are independent, then* $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t)$

(e.) If $\psi_X(t) = \psi_Y(t)$, *then* X *and* Y *have the same distribution.*

(f.) If $\psi_n(t) \to \psi(t)$, *then* $F_n(x) \to F(x)$

As an example, the characteristic function of the Cauchy density

$$f_a(x) = \frac{1}{\pi}\frac{a}{x^2 + a^2}$$

is

$$\psi_a(t) = e^{-a|x|}.$$

Thus since

$$\psi_a(t)\psi_b(t) = e^{-a|x|}e^{-b|x|} = e^{-(a+b)|x|}$$

it follows that the sum of two independent Cauchy r.v. is also Cauchy.

9.7 Problems.

(1.) Find the m.g.f. of the Geometric distribution, and use it to find the mean and variance.

(2.) Find the m.g.f. of the Pascal distribution. (*Hint*: Use problem 1.)

(3.) Let X have the Poisson distribution with mean λ.

(a.) Compute the m.g.f. of X.

(b.) Use the m.g.f. to find the mean and variance of X.

(c.) Prove that the sum of two independent Poisson r.v. of means λ_1 and λ_2 is Poisson with mean $\lambda_1 + \lambda_2$.

(4.) For the moments of the Poisson distribution, prove that

$$\mu_{n+1}(\lambda) = \lambda \left[\mu_n(\lambda) + \mu_n'(\lambda)\right].$$

(*Hint* Prove by induction that

$$\phi^{(n)}(t) = e^{-\lambda} e^{\lambda e^t} P_n(\lambda e^{-\lambda t})$$

where $P_n(x)$ is a polynomial satisfying the recursion

$$P_{n+1}(x) = x \left[P_n(x) + P_n'(x)\right].$$

Then set $t = 0$.)

(5.) Compute the m.g.f. of the censored r.v. $X = \min[T, c]$ where T is exponential with $a = 1$. Use it to find the mean and variance. Use scaling to get result for general a.

(6.) A r.v. W is *lognormal* iff $\log W$ has a normal distribution $N(\mu, \sigma^2)$. Prove that

$$\log E(W) = E(\log W) + \frac{\sigma^2}{2}.$$

(7.) Prove that if $f(x)$ is a density with m.g.f. $\phi(t)$, then for a positive integer r,

$$g(x) = \frac{1}{\mu_r} x^r f(x)$$

is a density with m.g.f $\phi^{(r)}(t)/\mu_r$, provided that either r is even or $f(x) = 0$ for $x < 0$.

(8.) Find the m.g.f. of the distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} x^2 e^{-x^2/2}$$

and use it to compute the mean and variance.

(9.) Prove that

$$f(x) = \frac{x^{2n}e^{-x^2/2}}{\sqrt{2\pi}\,(2n-1)!!}$$

is a density and compute its m.g.f. and find its moments.

(10.) Prove that if p_n is a discrete probability distribution on $n \geq 1$, with m.g.f. $\phi(t)$, then for any positive even integer n,

$$P_n = \frac{1}{\mu_r} n^r p_n$$

is a distribution with m.g.f $\phi^{(p)}(t)/\mu_p$.

(11.) Find the m.g.f. of the distribution

$$P_n = np^2q^{n-1} \qquad n \geq 1.$$

and use it to compute its mean and variance.

(12.) Let X_n have the Binomial distribution $Bin(n,p)$ with $p = \lambda/n$. Use the Continuity Theorem to prove that the distribution of X_n converges to the Poisson distribution with mean λ.

(13.) Let W_n have the Geometric distribution with $p = a/n$. Use the Continuity Theorem to prove that the distribution of

$$T_n = \frac{W_n}{n}$$

converges to the exponential distribution $\Gamma(a,1)$ with mean $1/a$.

(14.) Let X_n have Binomial distribution $Bin(n,p)$ with $p = 1/2$. Use the Continuity Theorem to prove that the distribution of

$$Z_n = \frac{X_n - \frac{n}{2}}{2\sqrt{n}}$$

converges to the standard Normal distribution $N(0,1)$.

(15.) (a.) Find the m.g.f $\phi(t)$ of the uniform distribution on $[0,1]$.

(b.) Let the r.v. X_n take on the values $k = 1, 2, ..., n$ with equal probabilities:

$$P\left(X_n = k\right) = \frac{1}{n} \qquad k = 1, 2, ..., n$$

for all $k = 1, ..., n$. Find the m.g.f. of $\phi_n(t)$ of X_n.

Show that the distribution of

$$Y_n = \frac{1}{n} X_n$$

converges to the uniform distribution.

Cumulants. The logarithm

$$g(t) = g_X(t) = \log \phi(t) = \log E(e^{tX})$$

of the m.g.f. of a distribution is called its *cumulant generating function (c.g.f.)* and its derivatives at the origin are the *cumulants*

$$\kappa_n = \kappa_n(X) = g^{(n)}(0)$$

of the distribution. They were first introduced in 1889 by Torvald Thiele.

(16.) Prove that $\kappa_1 = \mu$ and $\kappa_2 = \sigma^2$.

(17.) Find $g(t)$ for the unit normal distribution $N(0, 1)$.

(18.) Find $g(t)$ and κ_n for the Poisson distribution with mean λ.

(19.) Show that

(a.) $\kappa_1(X + c) = \kappa_1(X) + c$

(b.) $\kappa_n(X + c) = \kappa_n(X)$ for $n \geq 2$.

(c.) $\kappa_n(cX) = c^n \kappa_n(X)$.

(20.) Show that if X and Y are independent, then

$$g_{X+Y}(t) = g_X(t) + g_Y(t)$$

(21.) Find the c.g.f. of the Bernoulli distribution $Bin(1, p)$. Use problem (20.) to find the c.g.f. the distribution $Bin(n, p)$.

(22.) Find the c.g.f. of the (a.) geometric and (b.) Pascal distributions. (*Hint*: Use problem 5 for (b.))

Chapter 10
The Gamma Distribution and the Poisson Process.

10.1The Gamma Distribution.

The Gamma distribution is a generalization of the exponential distribution. It is called the Gamma distribution because of the Gamma function appearing in its definition.

We shall discuss its properties first, and see afterwards how it arises in interesting situations.

Definition 1. *For $a > 0$ and $\nu > 0$, the Gamma distribution $\Gamma(a, \nu)$ is defined by the density*.

$$f_{a,\nu}(x) = \frac{a^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-ax} \qquad x > 0.$$

We need $a > 0$ in order for the integral of $f_{a,\nu}(x)$ to converge at ∞, and $\nu > 0$ in order for it to converge at $x = 0$.

We have already encountered two special cases of the Gamma distribution

Example 1. If $\nu = 1$, $\Gamma(a, 1)$ has the exponential density

$$ae^{-ax} \qquad x \geq 0.\blacksquare$$

Example 2. If Z is Standard normal $N(0, 1)$, then Z^2 has the density

$$\frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}.$$

This is the Gamma density $\Gamma(\frac{1}{2}, \frac{1}{2})$.(See Example 1 of section 6 of Chapter 7). ■

Moment Generating Function. The m.g.f of the Gamma density is easily computed.

Theorem 1. *The m.g.f. of $\Gamma(a, \nu)$ is*

$$\phi(t) = \frac{a^\nu}{(a-t)^\nu}.$$

Proof. We have

$$\phi(t) = \frac{a^\nu}{\Gamma(\nu)} \int_0^\infty e^{tx} x^{\nu-1} e^{-ax} dx = \frac{a^\nu}{\Gamma(\nu)} \int_0^\infty x^{\nu-1} e^{-(a-t)x} dx.$$

Let $s = (a-t)x$ to obtain

$$\phi(t) = \frac{a^\nu}{(a-t)^\nu}\frac{1}{\Gamma(\nu)}\int_0^\infty s^{\nu-1}e^{-s}ds = \frac{a^\nu}{(a-t)^\nu}.\square$$

Just as for the exponential, the parameter a is a *scale factor*.

Corollary 1. *If X is $\Gamma(a,\nu)$, then aX is $\Gamma(1,\nu)$*

Proof. The m.g.f. of aX is

$$\phi(at) = \frac{1}{(1-t)^\nu}$$

which is the m.g.f. of $\Gamma(1,\nu)$.$\square$

Corollary 2. *The distribution $\Gamma(a,\nu)$ has mean ν/a and variance ν/a^2.*

Proof. See problem 1. $\square$.

The sum of independent Gammas with the *same scale factor* a is also a Gamma.

Theorem 2. *If X and Y are independent, X is $\Gamma(a,\nu_1)$ and Y is $\Gamma(a,\nu_2)$, then $X+Y$ is $\Gamma(a,\nu_1+\nu_2)$.*

Proof. We have

$$\frac{a^{\nu_1}}{(a-t)^{\nu_1}}\frac{a^{\nu_2}}{(a-t)^{\nu_2}} = \frac{a^{\nu_1+\nu_2}}{(a-t)^{\nu+\nu_2}}.\square$$

Corollary 3. *If $X_1, X_2, ..., X_n$ are independent exponential r.v. with parameter a, then $S = X_1 + X_2 + \cdots + X_n$ has the distribution $\Gamma(a,n)$.*

Corollary 4. *If $Z_1, Z_2, ..., Z_r$ are independent unit normal r.v., then $X = Z_1^2 + Z_2^2 + \cdots + Z_r^2$ has the distribution $\Gamma(\frac{1}{2},\frac{r}{2})$.*

The distribution $\Gamma(\frac{1}{2},\frac{r}{2})$ is important in Statistics, and is referred to there as $\chi^2(r)$, the *"χ^2-distribution with r degrees of freedom"*.

We can write down the general moment of the Gamma.

Corollary 5. *The n^{th} moment of $\Gamma(1,\nu)$ is*

$$\mu_n = \frac{\Gamma(\nu+n)}{\Gamma(\nu)}.$$

Proof. See problem 2.c

10.2 The Exponential Distribution.

A special case of the Gamma distribution is the *exponential distribution* $\Gamma(1, a)$.

We have noted the Markov or memoryless property of the Geometric distribution on the positive integers:

$$P(T > n + m \mid T > m) = P(T > n).$$

This follows easily since for the Geometric,

$$P(T > n) = q^n$$

where $q = 1 - p$.

The Geometric distribution is the *only memoryless waiting time on the positive integers*; that is the only memoryless waiting time if time is measured by the number of trials.

When time is measured continuously, one has a waiting time T with values in the positive real numbers. In this case, it is the *exponential distribution*

$$P(T > t) = e^{-at}.$$

that is the only memoryless waiting time on the positive real numbers. This should not be too surprising, since we originally obtained the exponential as the limit of geometric distributions in our discussion of the phone call model. (See also Example 2 of section 9.5.)

For a continuous r.v., the Markov Property is

$$P(T > t) = P(T > t + s \mid T > s).$$

Since

$$P(T > t) = P(T > t + s \mid T > s) = \frac{P(T > t + s)}{P(T > s)}.$$

the Markov property is equivalent to

$$P(T > t + s) = P(T > t)P(T > s)$$

This is clearly true for the exponential distribution.

Conversely, suppose that

$$F(t) = P(T > t)$$

satisfies

$$F(t + s) = F(t)F(s)$$

We shall prove that T is exponential under the assumption that $F(t)$ is differentiable. (This assumption can be eliminated, but that is a technical matter that we do not wish to pursue.) Differentiating with respect to s gives

$$F'(t + s) = F(t)F'(s)$$

Set $s = 0$ to get

$$F'(t) = F(t)F'(0) = -aF(t)$$

where $F'(0) = -a$, which implies

$$\frac{d}{dt}\left[e^{at}F(t)\right] = ae^{at}F(t) + e^{at}F'(t) = 0$$

Hence,

$$e^{at}F(t) = C$$

for some constant C. But setting $t = 0$ gives

$$C = F(0) = P(T > 0) = 1$$

so that

$$P(T > t) = F(t) = e^{-at}.$$

Note that we must have $a > 0$, since $P(T > t)$ is decreasing. ■

Sums of Independent Exponentials.

Let $T_1, T_2, ..., T_n, ...$ be i.i.d. exponential random variables with parameter a. Let $Y_n = T_1 + T_2 + \cdots + T_n$ be the sum of the first n variables.

Theorem 3. *S_n has density*

$$g_n(x) = \frac{a^n x^{n-1}}{(n-1)!}e^{-ax} \qquad 0 \le x < \infty.$$

and c.d.f.

$$P(Y_n < x) = G_n(x) = 1 - e^{-ax}\left(1 + \frac{ax}{1!} + \frac{(ax)^2}{2!} + \cdots + \frac{(ax)^{n-1}}{(n-1)!}\right)$$

Proof. By Corollary 3, S_n has the Gamma distribution $\Gamma(a, n)$, which has just the density $g_n(x)$ above. Integrating by parts gives

$$\begin{aligned} G_n(x) &= \frac{a^{n-1}}{(n-1)!}\int_0^x s^{n-1}e^{-as}ads \\ &= \frac{a^{n-1}}{(n-1)!}\left[-s^{n-1}e^{-as}\right]_0^x + \frac{a^{n-1}}{(n-1)!}\int_0^x (n-1)\, s^{n-2}e^{-as}ds \\ &= e^{-ax}\frac{(ax)^{n-1}}{(n-1)!} + \frac{a^{n-1}}{(n-2)!}\int_0^x s^{n-2}e^{-as}ds \\ &= e^{-ax}\frac{(ax)^{n-1}}{(n-1)!} + G_{n-1}(x). \end{aligned}$$

But since $G_1(x) = 1 - e^{-ax}$, we have the result by induction.■

Clearly, we have also

Corollary 6.

$$P(Y_n > x) = e^{-ax}\left(1 + \frac{ax}{1!} + \frac{(ax)^2}{2!} + \cdots + \frac{(ax)^{n-1}}{(n-1)!}\right).$$

10.3 The Poisson Process.

We shall now derive the Poisson process from general principles.

Consider a process such as the phone model where events (e.g. calls) are occurring at random times. Let T_1 be the wait for the first event and T_n be the *wait between the* $(n-1)^{st}$ *and the* n^{th} *events*.

We shall assume:

(1.) *The times* $T_1, T_2, ..., T_n, ...$ *are independent*

(2.) *They are memoryless with the same distribution.*

Hence, these times have an exponential distribution with some parameter a. It follows that

(1.) The wait for the n^{th} event is

$$Y_n = T_1 + T_2 + \cdots + T_n.$$

is therefore $\Gamma(a, n)$, and has the c.d.f. $G_n(x)$ given above.

(2.) If $N(t)$ is the number of events in the interval $[0, t]$, then

$$\begin{aligned} P(N(t) &= n) = P(Y_n \leq t \ \& \ S_{n+1} > t) \\ &= P(Y_n \leq t) - P(\ Y_{n+1} \leq t) = G_n(x) - G_{n+1}(x) \\ &= e^{-ax} \frac{(at)^n}{n!}. \end{aligned}$$

Therefore, $N(t)$ *is Poisson with mean* at. The parameter a is therefore the *mean number of events (calls) per unit time.*

The family $N(t)$ of r.v. is called the *Poisson process*.

10.4 Problems.

(1.) Prove that the distribution $\Gamma(a, \nu)$ has mean ν/a and variance ν/a^2.

(2.) Prove that the n^{th} moment of the distribution $\Gamma(1, \nu)$ is

$$\mu_n = \frac{\Gamma(\nu + n)}{\Gamma(\nu)}.$$

(3.) Use Newton's Binomial series

$$(1 + x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k$$

where

$$\binom{\alpha}{k} = \frac{\alpha\,(\alpha-1)\,(\alpha-2)\cdots(\alpha-k+1)}{1\cdot 2\cdot 3\cdot\cdots\cdot k}$$

for any real number, to show that the moments of $\Gamma\,(\nu, a)$ are

$$\mu_{n+1} = \frac{\nu\,(\nu+1)\cdots(\nu+n)}{a^{n+1}}.$$

(4.) A forest ranger mounts his tower to look for forest fires at random times, with independent exponential waits between, but an average of twice an hour. In order to prevent crying "wolf," he adopts the policy of not reporting a fire until he sees it for the second time. Lightning strikes a tree at noon, starting a fire.

(a.) What is the probability that he looks at least twice between noon and 2 p.m.?

(b.) According to Smokey's Third Law, the area of a forest fire is LT^2, where T is the length of time the first has been burning and L is a constant. Find the expected area of the fire at the time it is reported.

(5.) Cars on a lonely road pass randomly at an average rate of one car every 5 minutes.

(a.) What is the probability that no car passes in the first 15 minutes?

(b.) If no car has passed for 10 minutes, what is the probability that no car will pass in the next 5 minutes?

(6.) A store opens at 9 a.m. Customers then enter at an average rate of one every 5 minutes. Assume that the waiting times between customers are independent and have the same exponential distribution.

(a.) What is the probability that no customer enters in the first 15 minutes ?

(b.) What is the probability that exactly 5 customers enter in the first half-hour?

(c.) What is the probability that the third customer enters after $9:20$ a.m.

(7.) At beginning of each month, a manager orders toasters. If the average demand per month is 4, how many should he order to ensure an 80% chance of not running out.?

(8.) Motorists arrive at a gas station at Poisson rate of 20 per hour. What is probability that no one arrives in a given 5 minute period?

Chapter 11
The Normal Distribution and the Central Limit Theorem.

11.1 The Normal Distribution.

Let us now summarize some properties of the Normal distribution that we have already proved.

The *Standard Normal density*

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

has mean zero and variance one. Its m.g.f. is

$$\phi(t) = e^{t^2/2}.$$

If Z is a Standard Normal r.v., then $X = \sigma Z + \mu$ has the general Normal distribution $N(\mu, \sigma^2)$ with mean μ, variance σ^2 and density

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

An important property of the Normal distribution is that the sum of independent normal r.v. is again a normal r.v.

Theorem 1. *If X and Y are independent, and X is $N(\mu_1, \sigma_1^2)$ and Y is $N(\mu_2, \sigma_2^2)$, then $S = X + Y$ is $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

The Normal is perhaps the most important distribution in Probability, and certainly in Statistics. It was first obtained by DeMoivre in 1733 as an approximation to the Binomial distribution, an early version of the Central Limit Theorem. It is also known as the called the Gaussian distribution, since it was used by Gauss in his Theory of Errors in 1809,, and is referred to as the "*Bell Curve*" because its shape resembles a bell. No one, however, seems to refer to it as DeMoivre's distribution, after its discoverer.

11.2 Computing Normal Probabilities.

Section 11.2 Computing Normal Probabilities.

Probabilities for a r.v. X with the Normal distribution $N(\mu, \sigma^2)$ can be reduced to probabilities for the Standard Normal by normalizing to mean zero and variance one, that is, by noting that

$$Z = \frac{X - \mu}{\sigma}$$

has the Standard Normal distribution.

It is not possible to express the integral

$$P(a < Z < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt$$

in terms of the elementary functions of Calculus. However, probabilities for the Standard Normal are tabulated in widely available tables.

What is actually tabulated there is the c.d.f

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

so that

$$P(a < Z < b) = \Phi(b) - \Phi(a).$$

The table gives $\Phi(x)$ only for $x \geq 0$. However, since the integrand $e^{-t^2/2}$ is an even function, the normal curve is symmetric in the origin, and we have

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} e^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-t^2/2} dt$$

or

$$\Phi(-x) = 1 - \Phi(x).$$

From this, all probabilities can be found. One can use the formulas above, but it is often useful to draw a picture to keep things straight.

Example 1. Let X be $N(3, 4)$. Find $P(4 < X < 5)$.

Solution. We first normalize X to mean zero and variance one; that is, we note that the r.v.

$$Z = \frac{X - 3}{2}$$

has a Standard Normal distribution Thus,

$$\begin{aligned} P(4 &< X < 5) = P(\frac{4-3}{2} < Z = \frac{X-3}{2} < \frac{5-3}{2}) \\ &= P(0.5 < Z = \frac{X-3}{2} < 1.0) = \Phi(0.5) - \Phi(1.0) \\ &= 0.8413 - 0.6915 = 0.1798 = 0.18. \blacksquare \end{aligned}$$

Example 2. Let X be $N(3, 4)$. Find $P(1 < X < 4)$.

Solution. As above,

$$\begin{aligned} P(1 &< X < 4) = P(-1.0 < Z < 0.5) = \Phi(0.5) - \Phi(-1.0) \\ &= 0.8413 - (1 - 0.6915) = 0.5328.\blacksquare \end{aligned}$$

$$\begin{aligned} P(1 &< X < 4) = P(-1.0 < Z < 0.5) = \Phi(0.5) - \Phi(-1.0) \\ &= 0.8413 - (1 - 0.6915) = 0.5328.\blacksquare \end{aligned}$$

Example 3. Let X be $N(3, 4)$. Find $P(X < 0)$

Solution. As above,

$$P(X < 0) = P(Z < -1.5) = 1 - P(Z < 1.5) = 1.0 - 0.9332 = 0.0668.\blacksquare$$

Back of the Envelope Gaussian Probabilities.

For applications, especially to Statistics, it is good to have a rough idea of Normal probabilities. To begin with, note that if X is Normal with mean μ and variance σ^2, then since

$$Z = \frac{X - \mu}{\sigma}$$

is a Standard Normal $N(0, 1)$, the probability that X falls within say r *standard deviations* σ of its mean μ is just $P(|Z| \leq r)$. i.e.

$$P(|X - \mu| \leq r\sigma) = P(|Z| \leq r).$$

Thus, it is good to know $P(|Z| \leq r)$ approximately for various simple values of r. The most useful such fact is that

$$P(|Z| \leq 1.96) = 0.05.$$

Now $1.96 \simeq 2$, so we can say that *roughly* 95% *of the time a normal r.v. will fall within two standard deviations of its mean.*

The second most useful fact is the 99% *of the area is within about* 2.6 *standard deviations of the mean.*

For *one* Standard deviation,

$$P(|Z| \leq 1) = 0.68.$$

so *roughly* 68% *, or about* $2/3$ *of the time, a normal r.v. will fall within one standard deviation of its mean.*

Some exact, and approximate, figures are these:

$\mathbf{P(Z \leq r)}$	0.50	0.68	0.90	0.95	0.99
$\mathbf{r}$	0.674	1.0	1.645	1.96	2.576
$\mathbf{r} \simeq$	$2/3$	1	$5/3$	2	2.6

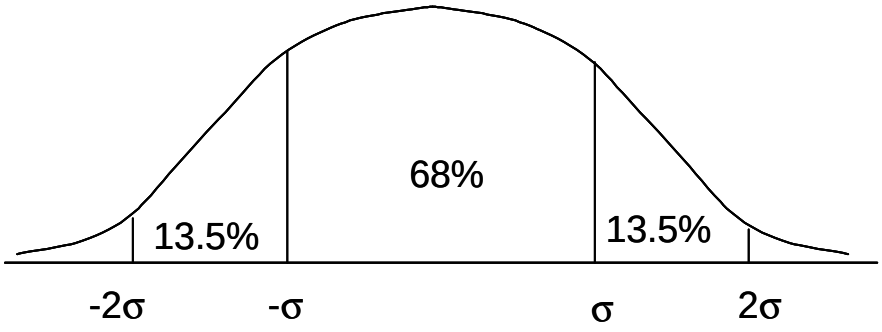

Figure 2.1. *Approximate Normal Probabilities.*

11.3 The Central Limit Theorem.

The Central Limit Theorem is one of the most important theorems of probability theory. First discovered in a special case by DeMoivre in 1733 and subsequently extended by Laplace, it states that the sum of a large number of independent identically distributed r.v. with finite mean and variance will have a normal distribution *regardless of the details of the distribution of the summands.*

Let $X_1, X_2, ..., X_n, ...$be a sequence of i.i.d. r.v. with mean μ and variance σ^2. We know from section 5.8 that the sum

$$S_n = X_1 + X_2 + \cdots + X_n$$

has mean $n\mu$ and, by independence, variance $n\sigma^2$.

The Central Limit Theorem states essentially that the distribution of S_n is approximately normal, with the known mean $n\mu$ and variance $n\sigma^2$, that is, S_n is approximately $N(n\mu, n\sigma^2)$. To be more accurate, what it states is that the r.v.

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

obtained by normalizing S_n to mean zero and variance one, has approximately the Standard Normal distribution $N(0, 1)$. This is fine for us, because, of course, we only find normal probabilities by reducing them to the Standard Normal, since that is what is tabulated.

Theorem 2. (*Central Limit Theorem.*) *Let $X_1, X_2, ..., X_n, ...$ be a sequence of i.i.d. r.v. with mean μ and variance σ^2, and*

$$S_n = X_1 + X_2 + \cdots + X_n$$

Then the distribution of

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

converges to the standard normal distribution.

To be specific,

$$\lim_{n\to\infty} P(a < Z_n < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

Proof. We will prove the result in the case that the distribution of the X_n's has a m.g.f., by showing that the m.g.f. of S_n converges to the m.g.f. $e^{t^2/2}$ of the Standard Normal. To simplify, we may assume that the X_k's have mean zero and variance one. Explicitly, we can write

$$Z_n = \frac{Y_1 + Y_2 + \cdots + Y_n}{\sqrt{n}}$$

where

$$Y_k = \frac{X_k - \mu}{\sigma}.$$

Now, Y_k has mean zero and variance one, so its m.g.f. $\phi(t)$ has as it first few terms

$$\phi(t) = 1 + \frac{1}{2}t^2 + \cdots$$

By independence, the m.g.f. of $Y_1 + Y_2 + \cdots + Y_n$ is $\phi(t)^n$, and by scaling, the m.g.f. of Z_n is

$$\begin{aligned} \phi_n(t) &= \phi(\frac{t}{\sqrt{n}})^n = \left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + \cdots\right]^n \\ &= \left(1 + \frac{t^2}{2n} + \cdots\right)^n \to e^{t^2/2}. \blacksquare \end{aligned}$$

This does not prove all that is stated, since of course, many r.v. have finite variance without having any moments higher that the second, much less an m.g.f.. Nevertheless, the theorem holds with under the stated conditions. Complete proofs may be found in many places.

Example 1 The weight in ounces of a pretzel is a random variable with the triangular density

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

Estimate the probability that a bag of 102 pretzels weighs at least 200 ounces.

Solution. Let X be the weight of a single pretzel. Then, by symmetry

$$E(X) = \mu = 2$$

and

$$\sigma^2 = var(X) = E(X-\mu)^2 = 2\int_2^3 (x-2)^2\,(3-x)\,dx = \frac{1}{6}.$$

The weight of the bag is

$$W = X_1 + X_2 + \cdots + X_{102}$$

where X_n is the weight on the n^{th} pretzel. By the Central Limit Theorem, W is approximately $N(n\mu, n\sigma^2) = N(204, 17)$. Hence,

$$P(W \geq 200) = P(Z = \frac{W-204}{\sqrt{17}} \geq -\frac{4}{\sqrt{17}} = -0.970) = 0.8340.\blacksquare$$

Example 2. According to problem 3 (c.) of chapter 9, a r.v. with distribution $Poi(n)$ for some integer n is the sum of n independent r.v. of distribution $Poi(1)$. By the Central Limit Theorem, we should expect it to be approximately normal for large n. This makes clearer the result of Example 4 of section 9.6,.according to which the Poisson distribution $Poi(\lambda)$ is approximately normal for large λ.

11.4 The DeMoivre-Laplace Approximation.

As we have already noted, the Binomial distribution $Bin(n,p)$ is the distribution of

$$X_n = 1_{S_1} + 1_{S_2} + \cdots + 1_{S_n}$$

where S_k is the event "Success on the k^{th} trial". Thus X_n is the sum of the n i.i.d. r.v. with finite mean p and variance pq. Thus for large n, the Binomial distribution is approximately $N(np, npq)$.

Theorem 3. (DeMoivre-Laplace Approximation.) *If X_n is Binomial $Bin(n,p)$, then the distribution of*

$$Z_n = \frac{X_n - np}{\sqrt{npq}}$$

converges to the Standard Normal distribution as $n \to \infty$.

Example 1. A fair coin is tossed 100 times. Estimate the probability the number N of Heads differs from the mean by no more than 5.

Solution. The mean of N is $np = 50$ and the variance is $npq = 25$. We want

$$P(45 \leq N \leq 55) = P(-1 \leq Z = \frac{N-50}{5} \leq 1) \approx 0.68.\blacksquare$$

The Continuity Correction.

Actually, this estimate is not as accurate as it could be. We are approximating a discrete distribution supported on the integers by a continuous one, so we need to be careful where we put the limits on the Gaussian density function. A little thought suggests that for each integer value, we should integrate over the interval of length one centered at that integer. This means that it is more accurate to approximate $P(44.5 \leq N \leq 55.5)$ rather than $P(45 \leq N \leq 55)$ by the Gaussian.

Thus, we write

$$\begin{aligned} P(45 &\leq N \leq 55) = P(44.5 \leq N \leq 55.5) = P(-1.1 \leq Z = \frac{N-50}{5} \leq 1.1) \\ &\approx 2\,(0.8643) - 1.0000 = 0.7286 \approx 0.73. \end{aligned}$$

This is called the *Continuity Correction* to the DeMoivre-Laplace Approximation.

Accuracy of approximations to distributions have been extensively studied in the literature, but this is beyond the scope of the present work.

11.5 Asymptotic Formula for the Normal Distribution.

Most tables of values of $\Phi(x)$ extend only up to around 3.0 or 3.5. For values larger than this, $\Phi(x)$ is quite close to one, but we may wish to know how close. What do we do if the numbers we need are not in the table?

The reason that the table stops where it does is that there is an asymptotic formula that gives a good approximation to $\Phi(x)$ for large x.

Theorem 4. *If $\Phi(x)$ is the c.d.f. of the standard normal distribution, then*

$$1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt \sim \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x}. \tag{11.1}$$

Example 1. As an example, suppose a fair coin is tossed 900 times. Estimate the probability of at least 510 heads.

Solution. The mean of N is $np = 450$ and the variance is $\sigma^2 = npq = 225$, and the standard deviation is $\sigma = 15$. We want

$$P(N \geq 510) = P(Z = \frac{N - 450}{15} \geq -\frac{60}{15} = 4.0)$$

The table contains no entry for 4.0, but by the approximation (11.1)

$$\begin{aligned} P(Z &> 4.0) = 1 - \Phi(4.0) \simeq \frac{1}{\sqrt{2\pi}} \frac{e^{-4^2/2}}{4} = \frac{1}{\sqrt{2\pi}} \frac{e^{-8}}{4} \\ &\simeq 3.3 \times 10^{-5}. \blacksquare \end{aligned}$$

*Derivation. The formula is obtained by integration by parts.

Lemma. *Define, for* $p \geq 0$,

$$I_p(x) = \int_x^\infty \frac{1}{t^p} e^{-t^2/2} dt.$$

Then $I_p(x) > 0$, *and,*

$$I_p(x) = \frac{e^{-x^2/2}}{x^{p+1}} - (p+1)\, I_{p+2}(x)..$$

Proof of the Lemma.

$$\begin{aligned} I_p(x) &= \int_x^\infty \frac{1}{t^p} e^{-t^2/2} dt = \int_x^\infty \frac{1}{t^{p+1}} e^{-t^2/2} t dt = \int_x^\infty \frac{1}{t^{p+1}} d\left[-e^{-t^2/2}\right] \\ &= \left[-\frac{e^{-t^2/2}}{t^{p+1}}\right]_x^\infty - (p+1)\int_x^\infty \frac{e^{-t^2/2}}{t^{p+2}} dt \\ &= \frac{e^{-x^2/2}}{x^{p+1}} - (p+1)\int_x^\infty \frac{e^{-t^2/2}}{t^{p+2}} dt \\ &= \frac{e^{-x^2/2}}{x^{p+1}} - (p+1)\, I_{p+2}(x). \blacksquare \end{aligned}$$

Proof of Theorem 4. We have

$$I_0(x) = \frac{e^{-x^2/2}}{x} - I_2(x) = \frac{e^{-x^2/2}}{x} - \frac{e^{-x^2/2}}{x^3} + I_4(x)$$

Hence

$$\frac{e^{-x^2/2}}{x} - \frac{e^{-x^2/2}}{x^3} < I_0(x) < \frac{e^{-x^2/2}}{x}$$

which implies that

$$1 - \frac{1}{x^2} < x e^{x^2/2} I_0(x) < 1$$

and hence

$$I_0(x) \sim \frac{e^{-x^2/2}}{x}.$$

But then

$$1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} I_0(x) \sim \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x}. \blacksquare$$

The derivation gives more, an upper and lower bound for $1 - \Phi(x)$, which lets us assess the accuracy of the approximation:

Corollary 1. For $x > 0$,

$$\frac{1}{\sqrt{2\pi}} \left(\frac{e^{-x^2/2}}{x} - \frac{e^{-x^2/2}}{x^3} \right) < 1 - \Phi(x) < \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x}.$$

Note that $1/\sqrt{2\pi} = 0.3989$ is approximately 0.4. This corresponds to the approximation $\pi \simeq 3\frac{1}{8}$.

*Remarks on Chebychev estimates.

Chebychev's inequality has two advantages (a.) it is quite general, since it requires only finite mean and variance; and (b.) it is very easy to prove. Its generality makes useful for proving general theorems, Its defect, which comes with its great generality, is that in special cases it is not numerically a good estimate. Chebychev's inequality states that the tail of the distribution drops off like $1/t^2$. But distributions like the exponential and the Gaussian decay much faster.

For example, for the standard normal density,

$$P(|Z| \geq 2) \approx 0.05$$

while according to Chebyshev,

$$P(|Z| \geq 2) \leq \frac{E(Z)}{4^2} = 0.125$$

which is true, but but *off by a factor of* 25..

More generally, the Chebychev estimate is,

$$P(|Z| > t) = P(Z^2 > t^2) \leq \frac{E(Z^2)}{t^2} = \frac{1}{t^2}.$$

while actually,

$$P(|Z| > t) \sim \frac{2}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t}.$$

which drops off very much faster.

11.6 Statistical Applications of the Central Limit Theorem.

1. Test of a Hypothesis.

Suppose that a coin is tossed 900 times, and produces 480 Heads? Is this evidence that the coin is not fair. Of course, one would automatically say "Yes". The expected number of Heads is 450, and this is considerably more than that, so one might think that the coin favored heads. On the other hand, one does not expect to hit *exactly* 450 - the probability of that is only about 0.02 - so the question is *how convincing is this evidence* that Heads is favored?

This is an example of what is known in Statistics as a problem of *Hypothesis Testing*. The first known author to discuss such a problem was an Englishman named John Arbuthnot, physician to Queen Anne, who was the first to translate Huygens book on probability into English. In a paper of 1710, he examined births in London for the 82 years from 1629 -1710,

and observed that in *every year*, male births outnumbered female births. He argued that *this could not have been due to chance because the probability of such a chance occurrence is* $1/2^{82} \simeq 2 \times 10^{-25}$. He claimed that this was Divine Providence compensating for the increased risk of premature death of men in wars. This is considered the first instance of Statistical Inference.

Applying this idea to our problem, we are led to ask: *"How likely is it that* 900 *tosses of a fair coin would give as many as* 480 *Heads*"? This is simple enough. The number N of Heads is a Binomial r.v. with $p = 1/2$ and $n = 900$. We want $P(N \geq 480)$. We may use the Central Limit Theorem the mean of M is 450, the variance is $\sigma^2 = npq = 225$, and the standard deviation $\sigma = \sqrt{225} = 15$. Thus $Z = (N - 450)/15$ is approximately a standard normal and

$$P(N \geq 480) = P(Z = \frac{N - 450}{15} \geq \frac{30}{15} = 2) = 0.025.$$

This is fairly small, and we may reasonably conclude that the die is not fair.

Weldon's Dice Data.

In problems we have generally assumed that all faces of a die are equally likely to arise when the die is rolled. One may ask however, to what extent this holds in practice. This, of course, depends on the particular physical objects that are rolled. Dice may certainly be modified to favor certain outcomes, but what of standard unaltered dice?

In 1894, the zoologist W. F. R. Weldon (1860 - 1906) performed the experiment of rolling 12 dice at a time a total of $26,306$ times, the equivalent of $12 \times 26,306 = 315,672$ rolls of a single die. He recorded the number of times that either a five or a six appeared. His data are given in the following table, where n_k is the number of times that k five or six appeared.

$\mathbf{k}$	0	1	2	3	4	5	6	7	8	9	10	11	12
$\mathbf{n}_k$	185	1149	3265	5475	6114	5194	3067	1331	403	105	14	4	0

The total number of fives and sixes is

$$\sum_{k=0}^{12} k n_k = 106,602.$$

According to the usual assumption, the number of fives or sixes should have a Binomial distribution with $p = 1/3$ and $n = 315,672$. In fact, the estimate for p here is

$$\hat{p} = 106,602/315,672 = 0.33770.$$

If we assume that $p = 1/3$, how likely are these data? The variance and standard deviation of $\hat{p}$ are

$$\begin{aligned} \sigma^2 &= pq/n = \frac{1}{3}\frac{2}{3}\frac{1}{315,672} = 7.03 \times 10^{-7} \\ \sigma &= 0.0008390 \end{aligned}$$

so that the Z value is

$$Z = \frac{0.33770 - 0.33333}{0.0008390} = 5.244.$$

This is very small, so Weldon's dice cannot be "fair", and must have some small asymmetry. A similar experiment with physically different dice might well give a different result.

Weldon's data were analyzed by Karl Pearson in 1900 using the χ^2 test, and later by R. A. Fisher in 1925.

2. The Margin of Error.

Another Statistical problem is that of *Estimation* One has a r.v.X with a distribution containing a parameter- for example, its mean μ - whose value we don't know, but wish to determine by making repeated measurements of the r.v. In the case of the mean, suppose we have a sample of the distribution of X, that is, n independent measurements $X_1, X_2, \cdots, X_n$ of X. The *Law of Large Numbers* of Chapter 5 suggests that that a good way to estimate μ is by the sample mean

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Indeed, the quantity $\bar{X}_n$ has the following desirable properties:

(1.) it has mean equal to μ,

(2.) its variance σ^2/n is small when n is large, and, very importantly,

(3.) it is a *function of the sample values alone*, and does not involve any unknown parameters, like μ or σ^2.

A quantity with property (3.) is called a *statistic*. Finding a satisfactory statistic - an *estimator* - to estimate a parameter in an assumed distribution is called the problem of *Point Estimation*.

In the special case where $X = 1_S$ is the indicator of an event S, the mean of X is $p = P(S)$, the probability of S. An estimator $\bar{X}_n$ is the frequency ratio

$$\bar{X}_n = \hat{p}_n = \frac{N}{n}$$

where

$$N = 1_{S_1} + 1_{S_2} + \cdots + 1_{S_n}$$

is the number of occurrences of S in the n trials. Using $\hat{p}_n$ to estimate p is quite reasonable, since

(a.) $\hat{p}_n$ has mean equal to p and

(b.) the variance of $\hat{p}_n$ is pq/n,which is small if n is large.

Having found an estimator of a parameter, the question arises as to *how good the estimate is likely to be*, that is how close the estimate is to the true value. This is the Statistical problem of *Interval Estimation*.

An example is the familiar "*margin of error*" of the political pollsters. Consider a poll of the usual sort, where a (presumably) random sample of voters is asked which of two

candidates they favor. Assuming that the number of voters is far greater that the number sampled, so that the trials may be considered independent, the number N favoring candidate A has a Binomial distribution $Bin(n,p)$, where n is the number of voters in the sample and p is the true fraction of voters favoring A.

What we are interested, of course, is the number p, which we estimate by the frequency ratio.

$$\hat{p}_n = \frac{N}{n}.$$

The main point to realize is that *a Statistical Estimator is a random variable*. In particular, the number $\hat{p}_n$ is a *random variable*, because *if another poll were taken, we could - and likely would - get another value for* $\hat{p}_n$. The expectation of $\hat{p}_n$ is

$$E(\hat{p}_n) = E\left(\frac{N}{n}\right) = p$$

and we know from the Law of Large Numbers that $\hat{p}_n$ will be close to p if n is large enough, since the variance pq/n of $\hat{p}$ is small However, we have only a particular finite number n of voters in our poll. What we want to know is: *How reliable is our estimate of* p?

The answer is given by what is called a *Confidence Interval*. We proceed as follows. The r.v. N has mean np and variance $\sigma^2 = npq$. We do not know this variance because we don't know p. However, we can estimate it if we note that for $0 \leq p \leq 1$, $pq = p\,(1-p)$ has a maximum of $1/4$ on at $p = 1/2$, so that

$$\sigma^2 \leq \frac{n}{4}$$

Moreover, pq is very close to the value $1/4$ unless p is near 0 or 1, which is not usually the case in polls. So we may estimate σ^2 by $n/4$. with reasonable accuracy. Let z_α be the number such that for a standard normal r.v.

$$P(|Z| \geq z_\alpha) = \alpha.$$

For example, if $\alpha = 0.05$, then $z_\alpha = 1.96 \simeq 2$. Then

$$\begin{aligned} 1-\alpha &= P(|Z| \leq z_\alpha) \simeq P\left(\left|\frac{N-np}{\sqrt{npq}}\right| \leq z_\alpha\right) = P\left(\frac{|\hat{p}-p|}{\sqrt{pq/n}} \leq z_\alpha\right) \\ &= P\left(|\hat{p}_n - p| \leq z_\alpha\sqrt{pq/n}\right) = P\left(\hat{p}_n - z_\alpha\sqrt{pq/n} \leq p \leq \hat{p}_n + z_\alpha\sqrt{pq/n}\right) \\ &\simeq P\left(\hat{p}_n - \frac{z_\alpha}{2\sqrt{n}} \leq p \leq \hat{p}_n + \frac{z_\alpha}{2\sqrt{n}}\right). \end{aligned}$$

Taking $\alpha = 0.05$, and hence $z_\alpha = 2$, this says that *the probability that the true value of* p *lies in the interval*

$$\hat{I} = \left[\hat{p}_n - \frac{1}{\sqrt{n}}, \hat{p}_n + \frac{1}{\sqrt{n}}\right]$$

is approximately 0.95. In shorthand,.

$$p = \hat{p}_n \pm \frac{1}{\sqrt{n}}.$$

The interval $\hat{I}$ is called a 95% *confidence interval for* p. The number $1/\sqrt{n}$ is what is called the margin of error in polls. If you will check your favorite news source, you will find almost universally the the margin of error is just $1/\sqrt{n}$, where n is the number polled.

Let us note carefully, though, what this really means. The interval $\hat{I}$ is a *random interval*; its endpoints are random variables. *A repeat of the survey, with the same number of voters, will likely give a different interval.* The probability that the true value of p lies in $\hat{I}$ is 0.95. That is, if surveys are repeated over and over, then 19 times out of 20, the interval so computed will contain p.

In particular, *it does not mean that the true value must lie within the margin of error* about $\hat{p}$,.only that it does so 95% of the time. Indeed, there is no way to obtain certainty, short of polling the entire population. This method is often used, and is referred to as an election.

Interval estimation requires a knowledge of the distribution of the estimator. The role of the Central Limit Theorem is that it provides an approximation to the distribution of $\hat{p}_n$ for large n.

11.7 *Gauss's Theory of Errors.

Maximum Likelihood Estimators.

One way of obtaining an estimator of a parameter μ is the *Method of Maximum Likelihood*. Essentially, one takes as the estimate of μ the *value* $\hat{\mu}$ *that makes the observed values* $x_1, x_2, \cdots, x_n$ *most probable*. To be precise, let X be a r.v. with density $f(x \mid \mu)$ depending on a parameter μ. If n independent measurements $X_1, X_2, \cdots, X_n$ of X are made the joint density of $X_1, X_2, \cdots, X_n$ is

$$f(x_1, x_2, \cdots, x_n \mid \mu) = f(x_1 \mid \mu) f(x_2 \mid \mu) \cdots f(x_n \mid \mu).$$

The probability of obtaining the values $x_1, x_2, \cdots, x_n$ is essentially $f(x_1, x_2, \cdots, x_n \mid \mu)$. One chooses $\hat{\mu}$ to be the value of μ that maximizes this number for fixed observed values $x_1, x_2, \cdots, x_n$.

Example 1. Suppose that we wish to estimate the parameter a in the exponential density

$$f(x \mid a) = ae^{-ax} \qquad x > 0.$$

We have

$$f(x_1, x_2, \cdots, x_n \mid a) = a^n e^{-a\Sigma_n}$$

where $\Sigma_n = x_1 + x_2 + \cdots + x_n$. The maximum is obtained when

$$\frac{\partial}{\partial a} f(x_1, x_2, \cdots, x_n \mid a) = \left(na^{n-1} - a^n \Sigma_n\right) e^{-a\Sigma_n} = 0$$

or

$$a = \frac{n}{\Sigma_n} = \frac{1}{\bar{X}_N}$$

The Maximum Likelihood Estimator of a is therefore the reciprocal of the sample mean, a reasonable result.■

The Theory of Errors.

An early use of the Normal distribution was in connection with errors in measurements, particularly in Astronomy. Different measurements of the same quantity typically give slightly different values. The problem was to combine several divergent values to obtain an estimate of the true value. The model was that when a quantity A was measured the result was equal to the true value a plus a random error ε. The question then arose as to what distribution the error ε might have.

If one assumes that ε is due the sum of a large number of more or less independent factors, a normal distribution for ε becomes a very reasonable conjecture. Gauss, however, gave a different derivation for the normal law. Gauss assumed that the error $\phi(x)$ distribution has the following properties.

(1.) $\phi(x)$ is an even function of x, and hence has mean zero.

(2.) $\phi(x)$ has a maximum at $x = 0$.

(3.) For any given sample values $x_1, x_2, \ldots, x_n$ the Maximum Likelihood Estimator of the mean of the quantity measured is

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right).$$

Theorem 5. *Let $\phi(x)$ be a continuously differentiable density satisfying the assumptions (1.) - (3.). Then for some $\sigma > 0$,*

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

Proof. We will only need to assume (3.) for $n = 3$. The distribution of the measured r.v. is $\phi(x - \mu)$, so that the Likelihood function is

$$f(x, y, z \mid \mu) = \phi(x - \mu)\phi(y - \mu)\phi(z - \mu).$$

We have for the maximum

$$\frac{\partial}{\partial \mu} f(x, y, z \mid \mu) = \ell(x - \mu) + \ell(y - \mu) + \ell(z - \mu) = 0$$

where

$$\ell(x) = \phi'(x)/\phi(x).$$

Note that $\phi(x)$ is an odd function

By (3.), this occurs if

$$\mu = \frac{1}{3}(x + y + z)$$

With a change of variables, therefore, we must have

$$\ell(x) + \ell(y) + \ell(z) = 0$$

whenever

$$x + y + z = 0$$

that is,

$$\ell(x) + \ell(y) + \ell(-(x + y)) = 0$$

or since $\ell(x)$ is odd,

$$\ell(x) + \ell(y) = -\ell(-(x + y)) = \ell(x + y)$$

Thus $\ell(x)$ is linear:

$$\frac{\phi'(x)}{\phi(x)} = -cx$$

Putting $c = -1/\sigma^2$ and integrating gives

$$\phi(x) = Ce^{-x^2/2\sigma^2}$$

By normalization,

$$C = 1/\sqrt{2\pi\sigma^2}.\blacksquare$$

As a justification for using the normal to model the error distribution, Gauss's argument seems less convincing than the appealing to the Central Limit.However, it throws additional light on the normal density.

11.8 Problems.

(1.) The diameter of an electric cable is normally distributed with mean 0.8 and variance 0.0004. What is the probability that the diameter will exceed 0.81 inch?

(2.) In a large, well-mixed bin or marbles, there are three sizes: 20% weigh 4 oz., 30% weigh 1 oz., and 50% weigh 2 oz. What is the probability that a bag of 200 marbles selected randomly from the bin weighs at least 400 oz.?

(3.) Find the probability that among $10,000$ random digits the digit 7 appears more than 968 times.

(4.) A model for long chain molecules is as follows: The molecule is assumed to lie in a plane. Each component is of length ℓ and is joined in the preceding component at an angle θ, uniformly distributed on $[0, 2\pi]$. If there are n components altogether, find the distribution of the difference X in the x-coordinates of the two end points, approximately for large n.

(5.) A die is rolled 1200 times, and it is noted that an ace is obtained 270 times. Is this good evidence that the die is not fair?

(6.) Forty-eight numbers are rounded off and then added together. Assuming that the roundoff errors are uniformly distributed over the interval $(-0.5, 0.5)$, what is the probability that the total roundoff error is no more than 2?

(7.) One thousand independent rolls of a fair die will be made. Compute an approximation to the probability that number 5 will appear less than 150 times.

(8.) A total of n votes are cast in an election, with a votes going to candidate A, and b to candidate B. Suppose that A wins by a margin of $d = a - b > 0$. However, the voting machines used have a small probability p of error, and will occasionally attribute an A vote to B and vice versa. Approximate the probability that machine error will change the result of the election.

For a numerical examples, take $n = 6,000,000$, $d = 500$ and $p = 0.01$ and $p = 0.0012$. How does the result differ if a machine error simply invalidates the vote?

(9.) A player plays a game which he has probability p to win and $q = 1 - p$ to lose for a stake of $\$1$ per play. His winnings after n games are therefore

$$S_n = X_1 + \cdots + X_n$$

where X_k is his gain on the k^{th} game. If $d = q - p > 0$, estimate the probability that he is ahead after n games if n is large.

(10.) If 1000 fair dice are rolled, what is the probability that the sum of the numbers on them lies between 3400 and 3600 ?

(11.) (*Twenty-six*) Thirteen dice are rolled 10 times! If a total of 26 or more sixes are rolled, the house pays 3:1 odds; for a total of 33 or more, it pays 7:1.

(a.) Compute the advantage of the house.

(b.) Compute the advantage if the odds are changed to 4:1 and 18:1.

(12.) In a large, well-mixed bin of marbles, there are three sizes: 20% weigh 4 oz., 30% weigh 1 oz., and 50% weigh 2 oz. What is the probability that a bag of 200 marbles selected randomly from the bin will weigh between 400 and 430 oz.?

(13.) One thousand independent rolls of a fair die will be made. Approximate the probability that number 5 will appear less than 150 times.

(14.) If 1000 fair dice are rolled, what is the probability that the sum of the numbers on them is between 3400 and 3600 ?

(15.) Bags on an airline average 40 pounds in weight with a standard deviation of 10 pounds. What is the probability that 100 bags exceed 4200 pounds in weight?

(16.) Suppose that candidate A actually has a 4-point lead over candidate B; that is, A has 52% of the vote and B has 48%. Estimate the probability that a random poll of 900 voters will show B leading.

(17.) A die is rolled 1200 times, and it is noted that an ace is obtained 270 times. Is this good evidence that the die is not fair?

(18.) A poll of 1600 students at Citsong College shows that 60% believe that extraterrestrials are responsible for global warming. Find 95% and 99% confidence intervals for the true fraction of those who so believe. (Use the maximal variance $pq = 1/4$.) What is the "margin of error"?

(19.) 100 snails are randomly sunning themselves on a 100 ft. walk.

(a.) What is the probability that the first 3 ft. of the walk contains no snails?

(b.) What is the probability that the first 50 ft. of the walk contains at least 60 snails?

(20.) In a certain year, 100 accidents occurred on a 200 mile stretch of road. Assume that the accidents are distributed uniformly and independently along the road.

(a.) What is the approximate probability that there were no accidents in the first 10 miles?

(b.) What is the approximate probability that there were at least 60 accidents in the last 50 miles?

(21.) The phone company estimates that an average of 14 lines between .cities A and B will be in use at the peak period. How many lines are required so that a caller is only 5% likely not to be able to get through?

(23.) A store opens at 9 a.m. Customers then enter at an average rate of one every 5 minutes. Assume that the waiting times between customers are independent and have the same exponential distribution. What is the probability that the 50^{th} customer enters between $12:30$ and 1 p.m.?

(24.) According to one national columnist, in one Ohio precinct, exit polling showed Kerry with 67% of the vote, while the official count gave him 38%. It was stated that "The probability of a sampling error this large is approximately $1/867,205,553$."
How many voters were in the survey?

Chapter 12
Conditional Expectation.

12.1 Conditional Probability.

Recall that the conditional probability $P(E \mid A)$ of E given A is given by

$$P(E \mid A) = \frac{P(EA)}{P(A)}$$

It is the probability for the experiment $(\mathcal{E} \mid A)$.

Example 1. Calls come in on two independent phones at a rate of λ_1 and λ_2 calls per hour respectively. After an hour, there have been total of n calls on both phones combined. What is the probability that k of them were on the first phone?

Solution. Let X be the number of calls on the first phone, and Y the number on the second. The numbers X and Y are independent Poisson r.v. with means λ_1 and λ_2. We are asking for the distribution of X given that $X + Y = n$.

As we have shown previously, $X + Y$ is Poisson with mean $\lambda_1 + \lambda_2$. The joint distribution of X and Y is

$$P(X = k \,\&\, Y = m) = e^{-(\lambda_1+\lambda_2)} \frac{1}{k!m!} \lambda_1^k \lambda_2^m.$$

Thus

$$\begin{aligned} P(X &= k \mid X + Y = n) = \frac{P(X = k \,\&\, X + Y = n)}{P(X + Y = n)} \\ &= \frac{P(X = k \,\&\, Y = n - k)}{P(X + Y = n)} = \frac{e^{-(\lambda_1+\lambda_2)} \lambda_1^k \lambda_2^{n-k} / k!\,(n-k)!}{e^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^n / n!} \\ &= \frac{n!}{k!\,(n-k)!} \frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n} = \binom{n}{k} p^k q^{n-k} \end{aligned}$$

where $p = \lambda_1 / (\lambda_1 + \lambda_2)$ and $q = 1 - p = \lambda_2 / (\lambda_1 + \lambda_2)$. Thus, conditioned on $X + Y = n$, X is $Bin(n, p)$ with $p = \lambda_1 / (\lambda_1 + \lambda_2)$.■

This result can be easily understood if we recall that the note that the respective probabilities of a call on the two phones in a short time Δt are$\lambda_1 \Delta t$ and $\lambda_2 \Delta t$. It is thus reasonable that the probability that the ***first call*** is on the first phone is $p = \lambda_1 / (\lambda_1 + \lambda_2)$. But the Poisson process is memoryless, so once a call has come in, every thing starts over, and the probability that the *next* call is on the first phone is again p, etc. So we have n independent trials (since $X + Y = n$) of an experiment with probability p.

The Law of Total Probability.

Recall next the *Law of Total Probability* ":(Theorem 1 of Chapter 3)

$$P(E) = \sum_k P(E \mid A_k)P(A_k).$$

where $A_1, ..., A_n$ is a set of mutually exclusive and exhaustive events. If X and Y are discrete r.v., then taking $E = \{X = x_i\}$ and $A_k = \{Y = y_k\}$, we obtain the marginal distribution of X

$$P(X = x_i) = \sum_k P(X = x_i \mid Y = y_k)P(Y = y_k).$$

where $P(X = x_i \mid Y = y_k)$ is the *conditional distribution*

$$P(X = x_i \mid Y = y_k) = \frac{P(X = x_i \text{ and } Y = y_k)}{P(Y = y_k)}.$$

If X and Y have joint density $f(x, y)$, we define the *conditional density of X given Y* to be

$$f(x \mid Y = y) = \frac{f(x, y)}{f_Y(y)}$$

where $f_Y(y)$ is the *marginal density* of Y. Then

$$f_X(x) = \int f(x, y)\, dy = \int \frac{f(x, y)}{f_Y(y)} f_Y(y)\, dy = \int f(x \mid Y = y)\, f_Y(y)\, dx$$

This is the *Law of Total Probability for continuous r.v.*

Example 2. Let X and Y have the joint density

$$f(x, y) = \frac{1}{y} e^{-(x/y)} e^{-y} \qquad \text{on } x, y \geq 0.$$

Then

$$f_Y(y) = \int_0^\infty \frac{1}{y} e^{-(x/y)} e^{-y} dy = e^{-y}.$$

and so

$$\begin{aligned} f(x \quad \mid \quad Y = y) &= \frac{f(x, y)}{f_Y(y)} = \frac{1}{y} e^{-(x/y)} e^{-y} \cdot \frac{1}{e^{-y}} \\ &= \frac{1}{y} e^{-(x/y)} \qquad \text{on } y \geq 0. . \blacksquare \end{aligned}$$

Example 3. A coin with unknown probability p of Heads is tossed n times and gives k Heads. Given this result, what is the distribution of p?

Solution. To solve this problem, we shall *assume* that since initially we know nothing about p, its distribution on uniform on $[0,1]$. Thus the number N of heads is $Bin(n,p)$ where p is a uniformly distributed r.v. Thus we have

$$P(N=k \mid p=s) = \binom{n}{k} s^k (1-s)^{n-k}$$

and

$$\begin{aligned} P(N &= k) = \int_0^1 \binom{n}{k} s^k (1-s)^{n-k}\, ds \\ &= \binom{n}{k} B(k+1, n-k+1) = \binom{n}{k} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} \\ &= \binom{n}{k} \frac{k!\,(n-k)!}{(n+1)!} = \frac{1}{n+1}. \end{aligned}$$

Thus, not surprisingly, under these hypotheses all values of N are equally probable.

We want $P(p=s \mid N=k)$. By Bayes formula,

$$\begin{aligned} P(p &= s \mid N=k) = P(N=k \mid p=s) \frac{P(p=s)}{P(N=k)} \\ &= \binom{n}{k} s^k (1-s)^{n-k} \frac{1}{1/(n+1)} \\ &= \frac{1}{B(k+1, n-k+1)} s^k (1-s)^{n-k}. \end{aligned}$$

This is a continuous Beta density which peaks out at $s = k/n$.■

12.2 Conditional Expectation, I.

Discrete Random Variables.

Considered as a function of the event E, the conditional probability $P(E \mid A)$ is a probability for the experiment $(\mathcal{E} \mid A)$. If X is a discrete r.v., the expectation of X for this experiment is

$$E(X \mid A) = \sum_i x_i P(X = x_i \mid A).$$

This is called the *conditional expectation of X given A*.

Just as. using the Law of Total Probability, we could often simply computations of *probabilities* by conditioning on various outcome of an experiment, we may compute *expectations* in the same way.

Theorem 1. (Law of Total Expectation.) *Let $A_1, ..., A_n$ be a partition of Ω; i.e. a set of mutually exclusive and exhaustive events. Then*

$$E(X) = \sum_k E(X \mid A_k)P(A_k).$$

Proof.

$$\begin{aligned} E(X) &= \sum_i x_i P(X = x_i) = \sum_i x_i \sum_k P(X = x_i \mid A_k)P(A_k) \\ &= \sum_k \left[\sum_i x_i P(X = x_i \mid A_k) \right] P(A_k) = \sum_k E(X \mid A_k)P(A_k). \square \end{aligned}$$

Example 1. To illustrate, let us compute the mean μ of a r.v. T with geometric distribution with probability.p. To do this, we *condition on the result of the first trial*. We have $T = 1$ or $T > 1$ according as there is Success or Failure on the first trial. Thus, by Theorem 1,

$$\mu = E(T) = E(X \mid T = 1)P(T = 1) + E(X \mid T > 1)P(T > 1). \qquad (12.1)$$

Clearly, *given a Success on the first trial*,. $E(X \mid T = 1) = 1$. On the other hand. if there is Failure on the first trial, then T is equal to $T = 1 + T_1$, where T_1 is the wait for the first Success *after* the first trial. But since the trials are independent, T_1 has the same distribution as T, so that

$$E(X \mid T > 1) = 1 + E(T_1) = 1 + \mu$$

Thus, by (12.1)

$$\begin{aligned} \mu &= 1 \cdot p + (1 + \mu)\, q \\ \mu\,(1 - q) &= \mu p = p + q = 1 \end{aligned}$$

Solving for μ, we obtain

$$\mu = \frac{1}{p}. \blacksquare$$

Example 2. Consider a radio fund drive. Suppose that calls pledging donations come in according to a Poisson process with an average rate of a calls per hour, and that the contribution on a single call is a r.v. X with mean μ. What is the average total contribution in an hour?

Solution. The if X_n is the amount of the n^{th} contribution, then $X_1, X_2, ..., X_n, ...$ is a sequence of i.i.d. r.v of mean μ The total amount is there fore

$$S = X_1 + X_2 + \cdots + X_N$$

where N is the number of calls in an hour. The catch is that *the number N of terms is a Poisson r.v, with mean a*. We can handle this by conditioning on the value of N:

$$\begin{aligned} E(S) &= \sum_{n=0}^{\infty} E(S \mid N=n)\, P(N=n) \\ &= \sum_{n=0}^{\infty} E(X_1+X_2+\cdots+X_n \mid N=n)\, P(N=n) \\ &= \sum_{n=0}^{\infty} n\mu\, P(N=n) = \mu \sum_{n=0}^{\infty} n\, P(N=n) = \mu\, a. \blacksquare \end{aligned}$$

This result is simply the common sense guess

$$\text{mean contribution} \times \text{ average number of calls} = a\mu,$$

Note that the nature of the distribution of the X_n's is not important.

Continuous r.v.

We may extend the *Law of Total Expectation* to continuous r.v. by using the conditional density defined above. If X and Y are continuous r.v. with joint distribution $f(x,y)$, we define the conditional expectation of X given $Y=y$ to be

$$E(X \mid Y=y) = \int x\, f(x \mid Y=y)\, dx$$

The *Law of Total Expectation* is then

Theorem 2.

$$E(X) = \int E(X \mid Y=y)\, f_Y(y)\, dy$$

Proof.

$$\begin{aligned} E(X) &= \int\int x f(x,y) dx dy = \int\int x \frac{f(x,y)}{f_Y(y)} f_Y(y) dx dy \\ &= \int \left[\int x f(x \mid Y=y) dx \right] f_Y(y) dy = \int E(X \mid Y=y)\, f_Y(y)\, dy. \square \end{aligned}$$

Example 3. For the joint density

$$\frac{1}{y} e^{-(x/y)} e^{-y} \quad \text{on } \; x, y \geq 0$$

we computed above that

$$E(X \mid Y=y) = y.$$

and

$$f_Y(y) = e^{-y}$$

Thus,

$$E(X) = \int_0^{\infty} E(X \mid Y=y) f_Y(y) dy = \int_0^{\infty} y e^{-y} dy = \Gamma(2) = 1. \blacksquare$$

12.3 Conditional Expectation, II.

We will now define an object which is important in Mathematical Statistics, particularly in the construction of estimators. Pay careful attention, because it is a bit tricky.

Definition 1. Let X and Y be r.v. and put $\varphi(y) = E(X \mid Y = y)$. We define the *conditional expectation of X given Y* to be

$$E(X \mid Y) = \varphi(Y).$$

It is most important to understand that $E(X \mid Y)$ is a *random variable; it is a function of the r.v. Y*. The prescription for finding it is:

(a.) Compute the numerical function $\varphi(y) = E(X \mid Y = y)$;

(b.) Plug the r.v. Y into $\varphi(y)$ to get the r.v. $\varphi(Y)$.

Example 1. For the joint density

$$\frac{1}{y}e^{-(x/y)}e^{-y} \quad \text{on} \quad x, y \geq 0$$

we computed above that

$$\varphi(y) = E(X \mid Y = y) = y.$$

Therefore, in this case,

$$E(X \mid Y) = \varphi(Y) = Y.\blacksquare$$

Example 2. For a discrete example, let X and Y be independent Poisson r.v. with means λ_1 and λ_2. Compute

$$E(X \mid S)$$

where $S = X + Y$.

Solution: We computed in section 12.1 that

$$P(X = k \mid S = n) = \binom{n}{k} p^k q^{n-k}$$

where $p = \lambda_1 / (\lambda_1 + \lambda_2)$. This is a Binomial distribution, so

$$\varphi(n) = E(X \mid S = n) = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = pn$$

and hence,

$$E(X \mid S) = \varphi(S) = pS.\blacksquare$$

Theorem 3. *The conditional expectation has the following properties*

(a.) $E\left[E(X \mid Y)\right] = E(X)$.

(b.) $E(X + Y \mid Z) = E(X \mid Z) + E(X \mid Z)$.

(c.) $E(g(Y)X \mid Y) = g(Y)E(X \mid Y)$.

(d.) If X *and* Y *are independent, then* $E(X \mid Y) = E(X)$.

Proof. We will give the proofs for the case in which X and Y have a joint distribution.

(a.) Let

$$\varphi(y) = E(X \mid Y = y) = \int xf(x \mid Y = y)dx$$

Then

$$\begin{aligned} E\left(E(X \mid Y)\right) &= E(\varphi(Y)) = \int f_Y(y)\varphi(y)dy \\ &= \int f_Y(y) \int xf(x \mid Y = y)dxdy = E(X) \end{aligned}$$

by the "Law of Total Expectation".

(b.) See problem 36.

(c.) We have

$$\begin{aligned} E(g(Y)X \mid Y = y) &= \int g(y)xf(x \mid Y = y)dx \\ &= g(y) \int xf(x \mid Y = y)dx \\ &= g(y)E(X \mid Y = y) = g(y)\varphi(y). \end{aligned}$$

Hence,

$$E(g(Y)X \mid Y) = g(Y)\varphi(Y) = g(Y)E(X \mid Y).$$

(d.) We have $f(x, y) = f_X(x)f_Y(y)$. Thus,

$$f(x \mid Y = y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(x)} = f_X(x)$$

and so

$$E(X \mid Y = y) = \int xf(x \mid Y = y)dx = \int xf_X(x)dx = E(X).\blacksquare$$

The Variance Formula.

Now define

$$W = X - E(X \mid Y)$$

Consider the decomposition of X given by

$$X = E(X \mid Y) + W.$$

The Variance Formula says the the variance of X is the sum of the variance of the conditional expectation plus the mean variance of W.

In order to prove this we will need the following Lemma.

Lemma 1. *Let* $W = X - E(X \mid Y)$ *Then*

(a.) $E(W \mid Y) = 0$.

(b.) $E(g(Y)W) = 0$ *for every function* $g(Y)$ *of* Y.

Proof. (a.) Since $E(X \mid Y)$ is a function of Y, we have

$$\begin{aligned} E(W \mid Y) &= E(X - E(X \mid Y) \mid Y) \\ &= E(X \mid Y) \mid Y) - E(E(X \mid Y) \mid Y) \\ &= E(X \mid Y) - E(X \mid Y) = 0. \end{aligned}$$

(b.) We have

$$E(g(Y)W) = E\left[E(g(Y)W \mid Y) \mid Y\right] = E\left[g(Y)E(W \mid Y) \mid Y\right] = 0.\blacksquare$$

By (a.) of Lemma 1, $E(W \mid Y = y) = 0$ for every y, and so the variance of W for fixed y is just

$$v(y) = E(W^2 \mid Y = y). = E\left[(X - E\left(X \mid Y\right))^2 \mid Y = y\right].$$

The variance of W given Y is therefore given by the following definition.

Definition 2 The *conditional variance* $var(X \mid Y)$ of X given Y is the function of Y defined by

$$var(X \mid Y) = v(Y).$$

Theorem 4. (The Variance Formula. Cournot. 1843.)

$$var(X) = E(var(X \mid Y)) + var(E\left(X \mid Y\right))$$

Proof. Let $\mu = E(X)$. Then

$$\begin{aligned} E(X - \mu)^2 &= E\left[W + \left(E\left(X \mid Y\right) - \mu\right)\right]^2 \\ &= E\left(W^2\right) + E\left(E\left(X \mid Y\right) - \mu\right)^2 + 2E\left[W\left(E\left(X \mid Y\right) - \mu\right)\right]. \end{aligned}$$

But $E(X \mid Y) - \mu$ is a function of Y, so the third term drops out by Lemma 1. Hence,

$$\begin{aligned} E(X-\mu)^2 &= E\left(W^2\right) + E\left(E(X \mid Y) - \mu\right)^2 \\ &= E(var(X \mid Y)) + var(E(X \mid Y)). \blacksquare \end{aligned}$$

Example 1. Let $X_1 + \cdots + X_N$ be i.i.d. r.v. with mean μ and variance σ^2 and N a positive integer valued r.v.

Find the variance of $S_N = X_1 + \cdots + X_N$.

Solution. We have

$$\begin{aligned} E(S \mid N = n) &= E(X_1 + \cdots + X_n) = n\mu \\ var(S \mid N = n) &= var(X_1 + \cdots + X_n) = n\sigma^2. \end{aligned}$$

so that

$$\begin{aligned} E(S \mid N) &= \mu N \\ var(S \mid N) &= \sigma^2 N. \end{aligned}$$

By the Variance Formula

$$\begin{aligned} var(S) &= E(var(S \mid N)) + var(E(S \mid N) \\ &= E(\sigma^2 N) + var(\mu N) = \sigma^2 E(N) + \mu var(N). \blacksquare \end{aligned}$$

The Prediction Inequality.

Theorem 5. *For every function* $g(Y)$,

$$E\left[(X - E(X \mid Y))^2\right] \leq E\left[(X - g(Y))^2\right]$$

Proof. We have

$$X - g(Y) = X - E(X \mid Y) + E(X \mid Y) - g(Y) = W + f(Y)$$

where $f(Y) = E(X \mid Y) - g(Y)$ is a function of Y. Then, since $E(Wf(Y)) = 0$ by Lemma 1

$$\begin{aligned} E[X - g(Y)]^2 &= E(W + f(Y))^2 = E\left(W^2\right) + E(f(Y))^2 + 2E(Wf(Y)) \\ &= E\left(W^2\right) + E(f(Y))^2 \geq E\left(W^2\right). \blacksquare \end{aligned}$$

12.4 The Bivariate Normal Density.

The Standard Bivariate Normal.

The *Standard Bivariate Normal density*, with coefficient of regression r was defined in section 8.7 as

$$f(x,y) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\left(x^2+y^2-2rxy\right)/2\left(1-r^2\right)\right]$$

where $-1 < r < 1$.

Theorem 6. *Let (X,Y) have the standard bivariate normal density. Then*

(a.) the marginal distributions of X and Y are $N(0,1)$.

(b.) X and Y are independent iff $r = 0$.

(c.) The conditional density is

$$f(y \mid x) = \frac{f(x,y)}{f(x)} = \frac{\exp\left[-\left(y-rx\right)^2/2\left(1-r^2\right)\right]}{\sqrt{2\pi\left(1-r^2\right)}}$$

Proof. Parts (a.) and (b.), were proved in Theorem 5 of section 8.8. Part (c.) is an easy exercise.□

This conditional density is clearly $N(rx, 1-r^2)$, so we have immediately:

Corollary 1. *For the Bivariate Normal*
(a.) the conditional expectations of are

$$E(Y \mid X = x) = rx \qquad \textit{and} \qquad E(X \mid Y = y) = ry.$$

and hence

$$E(Y \mid X) = rX \qquad \textit{and} \qquad E(X \mid Y) = rY.$$

(b.) *the conditional variance is*

$$var(Y \mid X = x) = 1 - r^2.$$

and hence

$$var\left(Y \mid X\right) = 1 - r^2.$$

Corollary 2. $E(XY) = r$.

Proof.

$$E(XY) = E\left[E(XY \mid X)\right] = E\left[XE(Y \mid X)\right] = E(rX^2) = r.\square$$

As a check, the Variance Formula asserts that

$$\begin{aligned} var(Y) &= E(var(Y \mid X)) + var(E(Y \mid X)) \\ &= \left(1 - r^2\right) + var(rX) = \left(1 - r^2\right) + r^2 = 1. \end{aligned}$$

which is correct.

The line

$$y = rx \tag{12.2}$$

which gives the mean of the conditioned r.v $(Y \mid X = x)$ as a function of x is called the *line of regression of Y on X*.

Similarly, the *line of regression of X on Y* is

$$y = \frac{1}{r}x \tag{12.3}$$

The General Bivariate Normal.

In general , two r.v. X and Y have a joint Bivariate Normal distribution iff

$$X = \sigma_1 V + \mu_1 \qquad \text{and} \qquad Y = \sigma_2 W + \mu_2$$

where (V, W) have the Standard Bivariate Normal density. In this case, we have

Theorem 7..*For the general bivariate normal density,*

(a.) The conditional expectation is

$$E(Y \mid X = x) = \frac{\sigma_2 r}{\sigma_1}(x - \mu_1) + \mu_2$$

(b.) The conditional variance is

$$var(Y \mid X = x) = \sigma_2^2\left(1 - r^2\right)$$

Proof. For (a.), we have

$$\begin{aligned} E(Y \mid X = x) &= E(\sigma_2 W + \mu_2 \mid \sigma_1 V + \mu_1 = x) \\ &= E(\sigma_2 W + \mu_2 \mid V = \frac{x - \mu_1}{\sigma_1}) \\ &= \sigma_2 E(W \mid V = \frac{x - \mu_1}{\sigma_1}) + \mu_2 = \sigma_2 r\left(\frac{x - \mu_1}{\sigma_1}\right) + \mu_2. \end{aligned}$$

For (b.),

$$\begin{aligned} var(Y \quad | \quad X = x) &= var(\sigma_2 W + \mu_2 \mid \sigma_1 V + \mu_1 \ = x) \\ &= var(\sigma_2 W + \mu_2 \mid V = \frac{(x - \mu_1)}{\sigma_1}) \\ &= \sigma_2^2 var(W \mid V = \frac{(x - \mu_1)}{\sigma_1}) = \sigma_2^2 \left(1 - r^2\right). \square \end{aligned}$$

For (c.), simply replace x by $(x - \mu_1)/\sigma_1$ x and y by $(y - \mu_2)/\sigma_2$ in $y = rx$. similarly for (d.).□

If we replace x by $(x - \mu_1)/\sigma_1$ and y by $(y - \mu_2)/\sigma_2$ in (12.2), we obtain the *line of regression of Y on X*.

$$y - \mu_2 = \left(\frac{\sigma_2}{\sigma_1} r\right)(x - \mu_1).$$

It passes through the point of the two means and has slope $\sigma_2 r/\sigma_1$.

Similarly from (12.3) we find that the *line of regression of X on Y is*

$$(x - \mu_1) = \left(\frac{\sigma_1}{\sigma_2} r\right)(y - \mu_2)$$

or

$$y - \mu_2 = \left(\frac{\sigma_2}{r\sigma_1}\right)(x - \mu_1).$$

Regression to the Mean.

In order to understand the term '*regression*' and the significance of the line of regression, let us consider the classic 1877 *Sweet Pea Experiment* of Francis Galton Galton measured the weights X of a group of 490 sweet pea seeds and compared the weights Y of the filial seeds grown from the original parent seeds. He observed that:

(1.) The parent and filial populations were both approximately normal with the same mean μ and variance σ^2.

(2.) The descendents of subgroups of parent seeds of the same weights x were also approximately normally distributed, with a variance independent of x.

(3.) The means μ_x of the filial populations lay approximately on a line

$$y - \mu = r(x - \mu)$$

where r was a positive number less than 1; in Galton's case $r \simeq 1/3$.

What this means is that the mean weight of seeds grown from a larger than average parent was also larger than average, *but only about* $1/3$ *as much* as the parent. It is clear that (1.) - (3.) describe a Bivariate Normal distribution.

In another experiment, Galton observed the same phenomenon with the heights of parents and children: the heights of children of tall parents were tall, but on average less tall than their parents. This phenomenon is called '*regression to the mean*'.

A Linear Model.

The Bivariate Normal can be viewed as what is called in Statistics a "*Linear Model*". For the Standard Bivariate Normal, write.

$$Y = rX + W$$

By Lemma 1, $E(WX) = 0$. In fact, however, W and X are independent. For if one makes the transformation

$$\begin{aligned} x &= u \\ y &= w - rx \end{aligned}$$

then

$$\partial(x,y)/\partial(u,w) = \begin{vmatrix} 1 & 0 \\ -r & 1 \end{vmatrix} = 1$$

and the density of (u,w) is

$$g(u,w) = f(u, w-ru)\frac{\partial(x,y)}{\partial(u,w)} = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\frac{1}{2}\left(u^2 + \frac{w^2}{(1-r^2)}\right)\right]$$

This factors, as

$$g(u,w) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \times \frac{1}{\sqrt{2\pi}} e^{-w^2/2(1-r^2)}.$$

It follows that X and W are independent, and that W is normal with variance $1 - r^2$.

If r is close to 1 the variance of W is a small quantity $\sigma^2 = 1 - r^2$.Thus,

$$Y = rX + W$$

where W is $N(0, \sigma^2)$ and independent of X.

For the general Bivariate Normal density, we have

$$Y = \beta_1 X + \beta_0 + \varepsilon$$

where ε is $N(0, \sigma^2)$ and independent of X. This is what is called in Statistics a simple *Linear Model*.

Ellipses of Constant Probability and the Lines of Regression.

There is a geometrical interpretation of the lines of regression in terms of the *ellipses of constant probability*, given in the case of the standard Bivariate normal by

$$f(x,y) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\frac{1}{(1-r^2)}\left(x^2 + y^2 - 2rxy\right)\right] = C$$

which is the same as

$$x^2 + y^2 - 2rxy = c.$$

Differentiating implicitly gives

$$2x + 2yy' - 2ry - 2rxy' = 0$$

The *vertical tangent* occurs when $y\prime = \infty$, or

$$\begin{aligned} 2y - 2rx &= 0 \\ y &= rx. \end{aligned}$$

This is the line of regression of Y on X. Thus the line of regression of Y on X is found graphically by drawing a line from the center of the ellipse to the point where the tangent to the ellipse is vertical.

Similarly, the *horizontal tangent* occurs when $y' = 0$ or

$$\begin{aligned} 2x - 2ry &= 0 \\ x &= ry \end{aligned}$$

This is the line of regression of Y on X.

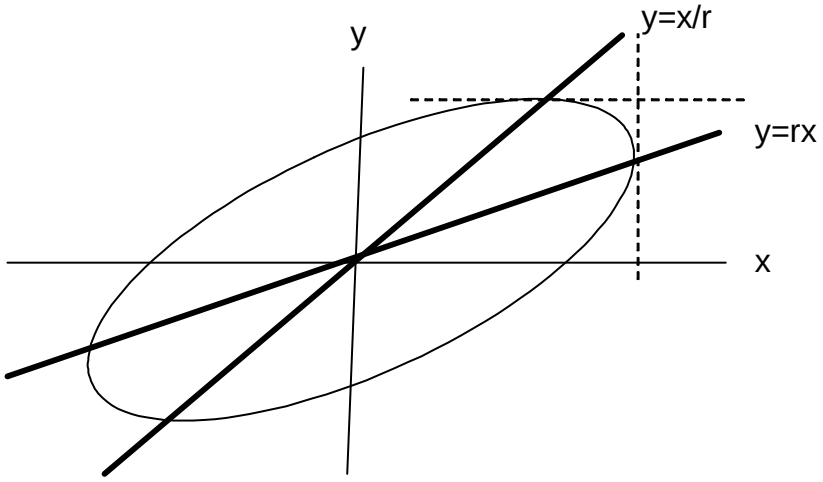

Figure 7.1. *Lines of regression and ellipses of constant probability.*

In general, the lines of regression of are determined by the vertical and horizontal tangents to the *ellipses of constant probability* defined by

$$(x-\mu_1)^2/\sigma_1^2 + (y-\mu_2)^2/\sigma_2^2 - 2r(x-\mu_1)(y-\mu_2)/\sigma_1\sigma_2 = c.$$

12.5 *Geometry of Random Variables.

There is a geometry of random variables. By this we mean that there is a way to picture r.v. as vectors in a Euclidean space, which, among other things, illuminates the notion of conditional expectation.

How are random variables like vectors? The properties of vectors that we have in mind are:

(1.) We can *add* vectors and *multiply them by scalars* (i.e. real numbers).

(2.) Vectors have a *dot product* .$a \cdot b = (a, b)$, which is

(a.) symmetric

$$(a, b) = (b, a)$$

(b.) linear

$$\begin{aligned} (a + c, b) &= (a, b) + (c, b) \\ (\alpha a, b) &= \alpha(a, b) \end{aligned}$$

(c.) and positive

$$(a, a) > 0 \qquad \text{if } a \neq 0.$$

From the dot product, we can get the *length* of a as

$$|a| = \sqrt{a \cdot a}$$

and the cosine of the *angle θ between a and b*:

$$\cos\theta = \frac{a \cdot b}{|a|\,|b|}$$

Two vectors a and b are therefore perpendicular or *orthogonal* iff

$$(a, b) = 0.$$

This is a lot of geometry, all based on addition and multiplication by numbers and the dot product. So, if we have a space of objects which we can add and multiply by numbers, and if we have something with the properties of a dot product, we can talk about things being orthogonal, and even about angles between them.

We can certainly add random variables and multiply them by numbers. Can we define a dot product - or as we say, an *inner product* (a, b) of random variables.? Yes; here's how it is done.

Definition 3. *Let X and Y be random variables with finite variance. The inner product of X and Y is defined to be*

$$(X, Y) = E(XY).$$

The *norm* or length of a r.v. is the square root of (X, X)

$$\|X\| = \sqrt{(X, X)} = \sqrt{E(X^2)}.$$

We use double lines for norm to distinguish the norm from the r.v. $|X|$ whose value is the absolute value of the number X.

Theorem 6. *Let X, Y and Z be r.v. with finite mean and variance. Then*

(a.) $(X + Y, Z) = (X, Z) + (Y, Z)$

(b.) $(cX, Y) = (X, cY) = c(X, Y)$

(c.) $(X, X) > 0$ if $X \neq 0$.

(d.) $|(X, Y)| \leq \|X\| \|Y\|$.

(e.) *If $|(X, Y)| = \|X\| \|Y\|$, then X and Y are linearly related, that is*

$$aX + bY = 0$$

for some a and b.

Proof. See Problem 37. Parts (d.) and (e.) are by Schwarz's inequality.□

The restriction to finite variance ensures by Schwarz's inequality that the r.v. XY actually has an expectation.

Several quantities of probability theory now have geometric interpretations.

1. First, the variance of X is just the squared length of the vector $X - \mu_x$

$$var(X) = E(X - \mu_x)^2 = \|X - \mu_x\|^2$$

By analogy with the dot product, we can speak of the "angle' θ between two r.v., defined by

2. Second, the cosine of the angle θ between X and Y is

$$\cos\theta = \frac{(X, Y)}{\|X\| \, \|Y\|}.$$

It follows that the cosine of the angle between $X - \mu_X$ and$Y - \mu_Y$ is the *correlation coefficient;*

$$\cos\theta = \frac{E(X - \mu_X, Y - \mu_Y)}{\sqrt{E(X - \mu_X)^2 \sqrt{E\,(Y - \mu_Y)^2}}} = \frac{cov(X, Y)}{\sqrt{var(X) var(Y)}} = \rho(X, Y);$$

In the extreme case that $\theta = 0$, and $\cos\theta = 1$, by Corollary 1 of section 5.4, X and Y are linearly related. Thus the correlation coefficient *measures the extent to which Y is a linear function of X*.

We must stress *linear function*, since, for example, if Z is $N(0,1)$, then

$$\rho(Z, Z^2) = 0$$

so that Z and Z^2 are uncorrelated, although Z^2 is clearly a function of Z.

Geometry of Conditional Expectation.

3. What is the geometrical picture of conditional expectation? According to Lemma 1, the r.v. $W = X - E(X \mid Y)$ is *orthogonal to every function of Y*, including, in particular, $E(X \mid Y)$. Thus, X is the sum of two perpendicular vectors:

$$X = E(X \mid Y) + W.$$

This is shown in Figure 5.1. The horizontal plane in the picture represents the space of functions of the r.v. Y, so $E(X \mid Y)$ is contained in this plane.

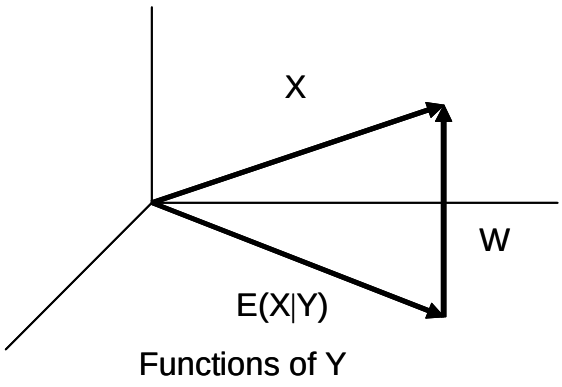

Figure 5.1. *The Variance Formula.*

We see that X, W and $E(X \mid Y)$ form the sides of a right triangle with X as hypotenuse, so Pythagoras Theorem states that

$$\|X\|^2 = \|E(X \mid Y)\|^2 + \|W\|^2$$

or

$$E(X^2) = E(W^2) + E\left[E(X \mid Y)^2\right]$$

This is the variance formula of Theorem 4. Thus, *in geometrical terms, $E(X \mid Y)$ is the projection of X onto the space of functions of Y*.

4. Going a bit farther, recall that the shortest distance from a point to a plane is the perpendicular distance. Thus, as shown in Figure 5.2, the distance from the point X to any point $g(Y)$ in the plane of functions of Y cannot be less that the distance from X to $E(X \mid Y)$; that is, for any function $g(Y)$,

$$E\left[(X - E(X \mid Y))^2\right] \leq E\left[(X - g(Y))^2\right]$$

This is the *Prediction Inequality* of Theorem 5.

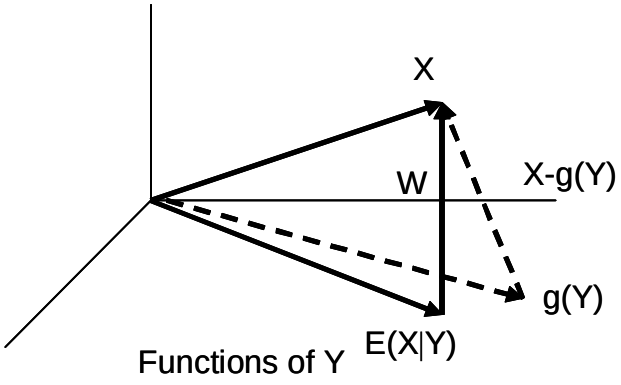

Figure 5.2. *The Prediction Inequality.*

12.6 Problems.

(1.) Let $X_1, ..., X_n$ be multinomial. Find $P(X_1, ..., X_r \mid X_{r+1}, ..., X_n)$.

(2.) Prove Theorem 4.

(3.) Let $N(t)$ be a Poisson process with parameter a and T an independent exponential waiting time with mean $1/b$. Find the distribution of $N(T)$.

(*Answer*. $N(T) + 1$ is geometric with $p = \frac{b}{a+b}$. This could by stated as a phone call problem: How many calls on phone #1 before phone #2 rings?

(4.) Same problem if means are different. Let X and Y be independent $\Gamma(a, 1)$ r.v. Find the distribution of $|X - Y|$. (Hint: Condition on $X > Y$ and $X < Y$.)

(5.) Let $f(x) = x + y$ on $0 \leq x \leq 1, 0 \leq y \leq 1$. Find $E(X \mid Y)$.

(6.) Let X and Y be independent $B(n, p)$ and $B(m, p)$ respectively, and $S = X + Y$.

Section 12.6 Problems.

(7.) Calls come in on two independent phones at a rate of λ_1 and λ_2 calls per hour respectively. After an hour, there have been total of n calls on both phones combined.

(a.) What is the probability that k of them were on the first phone?

(b.) If T_1 and T_2 are the waits for the first call on phones #1 and #2 respectively, show that

$$P(\text{first call is on phone } \#1) = P(T_2 > T_1) = \frac{\lambda_1}{\lambda_1 + \lambda_2} = p.$$

(8.) Let T be a r.v. with geometric distribution with probability.p. Compute the mean and variance of by conditioning on the result of the first trial.

(9.) Consider a radio fund drive. Suppose that calls pledging donations come in according to a Poisson process with an average rate of a calls per hour, and that the contribution on a single call is a r.v. with the mean μ. What is the average total contribution in an hour?

(10.) Let $N(t)$ be a Poisson process with parameter a and T an independent exponential waiting time with mean $1/b$. Find the distribution of $N(T)$.

(11.) Phone A rings at a rate a, phone B at rate b. What is the distribution of the number of times A rings before B rings?

(12.) Same problem if means are different. Let X and Y be independent $\Gamma(a, 1)$ r.v. Find the distribution of $|X - Y|$.

(*Hint*: Condition on $X > Y$ and $X < Y$.)

(13.) Same problem with different rates for X and Y.

(14.) Let X and Y be independent $B(n, p)$ and $B(m, p)$ respectively, and $S = X + Y$. Find $E(X \mid S)$.

(15.) Let X and Y be independent $B(n, p)$ and $B(m, p)$ respectively, and $S = X + Y$. Compute $E(X \mid S)$.

(16.) let X and Y be independent Poisson r.v. with means λ_1 and λ_2. Compute $E(X \mid S)$, where $S = X + Y$.

(17.) Let X and Y be independent Poisson r.v. with means λ_1 and λ_2 respectively. Let $S = X + Y$, Find $P(X = k \mid S = n)$. Interpret your answer.

(18.) Let X and Y be independent Binomial r.v. with $X \ \ Bin\ (n, p)$ and $Y \ \ Bin\ (m, p)$. Let $S = X + Y$. Find $P(X = k \mid S = r)$. Interpret your answer.

(19.) Let X be sampled from a Poisson distribution whose mean λ is an exponential r.v. with mean $\frac{1}{a}$. Find the distribution of X.

(20.) Let the random variables X and Y have the joint density

$$\begin{aligned} f(x, y) &= x^2 + \frac{xy}{3}, \qquad 0 < x < 1,\ 0 < y < 2, \\ &= 0 \qquad\qquad\quad \text{elsewhere.} \end{aligned}$$

Compute $E(X \mid Y = y)$ and $E(X \mid Y)$

(21.) Let the random variables X and Y have the joint density

$$f(x,y) = \frac{1}{4}\left(1 + \frac{xy}{2}\right)$$

for $-1 \leq x, y \leq 1$.

Compute $E(X \mid Y = y)$ and $E(X \mid Y)$

(22.) Let $N(I)$ be a Poisson point process on the real line, with density d. Let I be a finite interval and J an interval contained in I. Find the distribution of $N(J)$, given that there are n points in I. Interpret your answer.

(23.) Let the random variables X and Y have the joint density

$$\begin{aligned} f(x,y) &= \frac{1}{y}e^{-\left(\frac{x}{y}\right)}e^{-y}, \qquad 0 < x,\ 0 < y; \\ &= 0 \qquad\qquad\qquad \text{elsewhere.} \end{aligned}$$

(a.) Compute $f(x, Y = y)$.

(b.) Compute $E(X \mid Y = y)$.

(c.) Compute $E(X \mid Y)$.

(24.) Let $N(t)$ be a Poisson process with density a. If $0 < t < s$, find $P(N(t) = k \mid N(s) = n)$.

(25.) Let S_k be Success on the k^{th} Bernoulli trial, and $X_n = 1_{S_1} + \cdots + 1_{S_n}$ the number of Successes in n trials. Find

$$E(X_n \mid X_{n+m} = k)$$

(26.) (*Sum of a random number of r.v.*) Let $X_1, ..., X_n$ all have mean μ. Let $S_n = X_1 + \cdots + X_n$ and let

$$S = S_N = X_1 + \cdots + X_N$$

where N is a random positive integer valued r.v. with mean λ.

(a.) Find the mean of S.

(b.) Assume, in addition, that $X_1, ..., X_n$ are independent and have the same mean μ and variance σ^2, and that N has variance τ^2. Find $var\,(S)$.

(27.) Find the mean and variance of a geometric r.v. by conditioning on the outcome of the first trial.

(28.) Suppose that the number of eggs laid by an insect is Poisson with mean λ, and that the the probability of an egg developing is p. Assuming independence of the eggs, find the distribution of the number of eggs hatched.

Section 12.6 Problems.

(29.) Calls come in to a phone line at a rate of a calls per hour. When a call comes in, a biased coin with probability p of Heads is tossed. Find the mean of the number X of Heads thrown in the first three hours.

(30.) Consider a sequence of n Bernoulli trials. Given that there were m Successes in the n trials, find the probabilities of the various orders in which the sequence of m Succsses and $n-m$ Failures can occur.

(31.) Prove for the general Bivariate Normal density that the lines from the point $\left(\mu_x.\mu_y\right)$ to the points of vertical and horizontal tangency of the ellipse of constant probability are respectively the lines of regression of Y on X and of X on Y.

(32.) Prove that the conditional density of the standard bivariate normal density is

$$f(y \mid x) = \frac{\exp\left[-(y-rx)^2 / \left(1-r^2\right)\right]}{\sqrt{2\pi\left(1-r^2\right)}}$$

(33.) (a.) Prove that if X is normal and $(Y \mid X = x)$ is normal for each x, with mean linear in x and variance independent of x, then (X, Y) is bivariate normal. (b.) If in addition, X and Y are identically distributed, prove that

$$(Y - \mu \mid X = x) = r\,(x - \mu)$$

with $-1 < r < 1$.

(34.) Assuming that X, Y, W have joint density, prove Theorem 3 (b.).

Chapter 13
Random Walks.

13.1 Random Walk and Gambler's Ruin.

Here are two classic problems.

The Gambler's Ruin.

A gambler A, with initial capital k, plays repeatedly a game against an adversary B. A wins one dollar with probability p, and loses a dollar with probability $q = 1 - p$. Let B's initial capital be $a - k$, so that a is the total amount of money in the game. Each play is assumed to be independent of the others.

Suppose they agree to play until one player has won all the money. Several questions arise.

(a.) What is the probability that A is ruined?

(b.) What is the probability that the game continues forever?

(c.) What is the expected duration of the game?

A Random Walk.

Consider a set of points indexed by the first $n + 1$ nonnegative integers. $S = \{0, 1, 2, ..., a\}$.Starting at some point k in S, a Walker takes one step either to the right with probability p, or to the left with probability $q = 1 - p$, except that if he is at one of the endpoints $k = 0$ or $k = a$, he remains permanently at that endpoint. Each step is assumed to be independent of the others.

(a.) What is the probability that he eventually ends up at $k = 0$?

(b.) What is the probability that he continues walking forever, never reaching either endpoint?

(c.) What is the expected number of steps that he will take before reaching one of the endpoints?

Mathematically, these are the same problem. For suppose that the Walker starts at the point z corresponding to the players initial stake, steps right when player A wins and left when he loses. The answers to (a.), (b.) and (c.) will be the same. On the other hand, from the perspective of the Gambler's Ruin, suppose that A is betting that the Walker will step to the right.

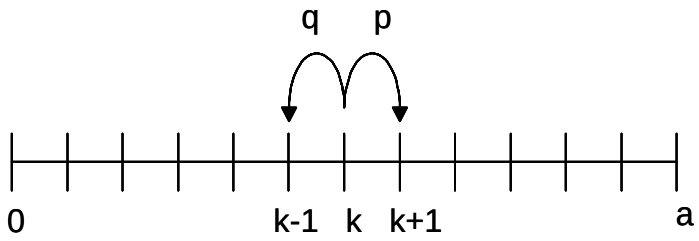

Figure 1.1. *A random walk on the integers* 0 *through* a.

Ruin Probabilities.

We begin by solving problem (a.), using conditional probabilities.

Let q_k be the probability that A is ruined starting from the point k. Conditioning on the result of the next game, we have

$$\begin{aligned} & P(\text{Ruin from } k) \\ = \;& P(\text{Win on first game})P(\text{Ruin from } k+1) + P(\text{Loss on first game})P(\text{Ruin from } k-1) \end{aligned}$$

That is,

$$q_k = pq_{k+1} + qq_{k-1.} \tag{13.1}$$

In addition,

$$q_0 = 1 \quad \text{and} \quad q_a = 0; \tag{13.2}$$

for if $k = 0$, A has lost already lost all his money, while if $k = a$, he has already won all of his opponents.

Now, equations (13.1) and (13.2) determine q_k uniquely. Equation (13.1) is what is called a *linear difference equation*. It can be solved in a manner similar to that used to solve linear differential equations.

Let us try for a solution of the form

$$q_k = r^k$$

for some number r, and try to determine r. Plugging in, we have

$$qr^k = pr^{k+1} + qr^{k-1}$$

Dividing by r^{k-1} gives a quadratic equation for r

$$\begin{aligned} qr &= pr^2 + q \\ pr^2 - qr + q &= (pr - q)(r - 1) = 0 \end{aligned}$$

with solutions

$$r = \frac{q}{p} \quad \text{and} \quad r = 1.$$

We therefore obtain two solutions

$$q_k = \left(\frac{q}{p}\right)^k \quad \text{and} \quad q_k = 1.$$

Note now that (because the equation is linear) any linear combination of two solutions of (13.1) is also a solution. Therefore,

$$q_k = C\left(\frac{q}{p}\right)^k + D$$

is a solution of (13.1) for any constants C and D. These two constants can now be determined to satisfy the two boundary conditions (13.2): we need

$$\begin{aligned} q_0 &= C + D = 1 \\ q_1 &= C\left(\frac{q}{p}\right)^a + D = 0 \end{aligned}$$

A little algebra now gives

$$q_k = \frac{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^k}{\left(\frac{q}{p}\right)^a - 1}. \tag{13.3}$$

Problem (b.) is now very simple if we note that B starts with $a-k$ dollars, and so the probability of B's ruin is simple q_{a-k}. But a computation shows that

$$q_k + q_{a-k} = 1. \tag{13.4}$$

Thus,one of A or B is ruined with probability 1, and so the game does not continue forever.

Note that in the symmetric case $p = q = \frac{1}{2}$ the formula fails. This is because then quadratic for r has a double root $r = 1$ and we obtain only one solution of the form r^k instead of two. In this case the second solution is kr^k, and we obtain the solution

$$q_k = 1 - \frac{k}{a} = \frac{a-k}{k} \tag{13.5}$$

This is the ratio of the opponents stake to the player's stake, a rather intuitive result.

Example 1. Nancy is spending the weekend in Reno. She decides to risk her entire bankroll of $\$10,000$ in order to try to make her expenses of $\$1000$, but she will quit once she is $\$1000$ ahead. She makes only bets at the dice table which she has a probability 0.492929 of winning, betting $\$100$ on each roll. What is the probability that she will make expenses rather than blow her entire roll?

Solution. This is a gambler's ruin problem. Since she is betting in multiples of $\$100$, she has $k = 10,000/100 = 100$ units, while her "opponent" has $1000/100 = 10$ units, since he is considered ruined, and play stops, when she reaches $\$1000$ in winnings. With $p = 0.492929$, we have $r = q/p = 1.028690$, so the probability of her *ruin* is

$$q_{100} = \frac{r^{110} - r^{100}}{r^{110} - 1} = 0.25786 \simeq 0.26$$

so she has a probability of approximately 0.74 of success.

Example 2. In Example 1, suppose Nancy makes only $\$50$ bets?

Solution. This effectively doubles the number of gambling units, so $k = 200$ and $a = 220$. The results in a probability of ruin of 0.43 and a probability of success of only 0.57. Increasing the size of the bets - decreasing the number of gambling units - decreases the ruin probability of the player with unfavorable odds. ■

13.2 Duration of the Game.

For problem (c.), let the game start from k, and let T_k be the number of plays until one of the players is ruined. Let $D_k = E(T_k)$. Then D_k is the average length or *duration* of a game.

Let W_1 = "*Win on first play*" and L_1 = "*Lose on first play*". Conditioning on the first play gives :

$$E(T_k) = E(T_k \mid W_1)P(W_1) + E(T_k \mid L_1)P(L_1) \tag{13.6}$$

But since the trials are independent,

$$E(T_k \mid W_1) = D_{k+1} + 1$$

is just the expected duration starting at $k + 1$ plus 1 for the play already made. Thus (13.6) says that

$$D_k = (D_{k+1} + 1)\, p + (D_{k-1} + 1)\, q = pD_{k+1} + qD_{k-1} + 1 \tag{13.7}$$

We also obviously have

$$D_0 = D_a = 0 \tag{13.8}$$

since if one player has no stake at the beginning, the game is over before it starts.

Equation (13.7) is an inhomogeneous equation because of the 1. If we try a solution of the form

$$D_k = ck$$

we find, on plugging in, that

$$c = \frac{1}{q - p} \tag{13.9}$$

works. If we subtract off this solution, we find that

$$Q_k = D_k - \frac{k}{q - p}$$

satisfies

$$Q_k = pQ_{k+1} + qQ_{k-1}$$

which is the sane as (13.1). Thus

$$Q_k = C \left(\frac{q}{p}\right)^k + D$$

and

$$D_k = \frac{k}{q-p} + C\left(\frac{q}{p}\right)^k + D. \tag{13.10}$$

Determining the constants C and D to satisfy (13.8) gives the final answer

$$D_k = \frac{k}{q-p} + \left(\frac{a}{p-q}\right)\frac{1-\left(\frac{q}{p}\right)^k}{1-\left(\frac{q}{p}\right)^a}. \tag{13.11}$$

For $p = q = 1/2$, the result is

$$D_k = k(a-k). \tag{13.12}$$

Example 1. In Examples 1 and 2 of the preceding section, how many rolls on average will Nancy's play last?

Solution. With $d = q - p = 0.014142$ we have, for bets of $\$100$,

$$D_{100} = \frac{100}{d} + \left(\frac{110}{d}\right)\frac{1-r^{100}}{1-r^{110}} = 1298.57.$$

For bets of $\$50$, $D = 5320$. ■

13.3 An Infinitely Rich Adversary.

Let us consider now the case in which the adversary B has infinite resources.

The probability q_k that A is ruined can be found by taking the limit as $a \to \infty$ in (13.3). This gives

$$q_k = \begin{cases} \left(\frac{q}{p}\right)^k & \text{if } p > q, \quad \text{i.e. } p > \frac{1}{2} \\ 1 & \text{if } p \le q, \quad \text{i.e. } p < \frac{1}{2} \end{cases}$$

Thus if the game is unfavorable to A, he will be ruined with probability 1, whereas if it is favorable, he will have a positive probability

$$1 - \left(\frac{q}{p}\right)^k$$

of *never* being ruined, which increases the larger his initial stake. For the duration, we have from (13.10)

$$D_k = \begin{cases} \infty & \text{if } p > q, \quad \text{i.e. } p > \frac{1}{2} \\ \frac{k}{q-p} & \text{if } p < q, \quad \text{i.e. } p < \frac{1}{2} \end{cases}$$

Thus in the favorable case, the expected duration is infinite. This is to be expected, since the length of the game is even infinite with positive probability.

What about the symmetric case $p = q = \frac{1}{2}$? So far the results have been pretty much what one would guess, but here we have something interesting. Taking $a \to \infty$ in (13.5) and (13.12) gives for the probability of ruin

$$q_k = 1$$

and for the duration

$$D_k = \infty.$$

That is, *A is ruined with probability 1, but the expected wait for that to happen is infinite!* The interpretation is that there are games in which A gets so far ahead so that it takes a very long time for him to be wiped out. These very large values dominate the arithmetic mean, and drive it to infinity in the limit. Understand that the *length of the walk is in every case finite*, it is just that the *average over a large number n of trials* gets larger and larger as n tends to infinity.

In terms of random walks, this is a random walk on the nonnegative integers $\{0, 1, 2, ...\}$ in which the walk stops once it reaches $k = 0$. The results say that if $p \leq \frac{1}{2}$, then no matter where he starts, the Walker will eventually reach the origin, but if $p > \frac{1}{2}$, there is a drift to the right and the Walker may never reach the origin. The expected duration of the walk is finite if $p < \frac{1}{2}$ but infinite if $p \geq \frac{1}{2}$.

13.4 Random Walk on the Integers.

Consider next a random walk on the full integers

$$Z = \{..., -2, -1, 0, 1, 2, ...\}$$

with probability p of moving to the right. There are no exceptional points in this case; if the Walker hits the origin, he moves right with probability p, and left with probability q.

Suppose that the Walker starts from the origin $k = 0$ The probability that he returns is

$$\begin{aligned} & P(\text{return}) \\ = \; & P(\text{1st step is right})P(\text{ruin from } k = 1) + P(\text{1st step is left})P(\text{ruin from } k = -1) \end{aligned}$$

If $p > \frac{1}{2}$, this gives

$$P(\text{return}) = p\left(\frac{q}{p}\right)^1 + q \cdot 1 = 2q < 1.$$

Once he has returned, the event of a second return is independent of what has gone on before and again has probability $2q$. So

$$P(\text{two returns }) = (2q)^2$$

Similarly,

$$P(n \text{ returns }) = (2q)^n .$$

as $n \to \infty$. The probability that he returns infinitely often is therefore zero, since

$$P(\text{return infinitely often }) \leq P(n \text{ returns }) = (2q)^n \to 0$$

Thus for an infinite walk with $p > \frac{1}{2}$, the Walker will return to the origin only a finite number of times, but will eventually disappear off to the right. For $p < \frac{1}{2}$, the result is the same, except that he disappears to the left.

For $p = \frac{1}{2}$, the probability of return is 1, and hence $P(n \text{ returns }) = 1$ and

$$P(\text{return infinitely often }) = \lim P(\text{returns } n \text{ times}) = 1.$$

Thus for a symmetric random walk, the particle returns to the origin infinity often with an infinite average wait in between each return.

13.5 *Brownian Motion.

We now want to consider a continuous version of the random walk.

Consider a symmetric random walk on a lattice of points separated by a distance that is on the points

$$\{..., -\Delta x, 0, \Delta x, 2\Delta x, 3\Delta x, ...\}$$

Suppose that the particle jumps to an adjacent point at a time interval Δt. Let X_n be ± 1 with probability $1/2$. The position of the particle at time $t = n\,\Delta t$ is then

$$X_n(t) = \Delta x\,(X_1 + X_2 + \cdots + X_n)\,.$$

By the Central Limit Theorem, $X_n(t)$ is approximately $N(0, \sigma^2)$ where

$$\sigma^2 = n\,(\Delta x)^2 = t\frac{(\Delta x)^2}{\Delta t}.$$

If we pass to the limit as $n \to \infty$, we will get something if

$$\frac{(\Delta x)^2}{\Delta t}$$

converges to a limit. We will therefore take

$$(\Delta x)^2 = \Delta t$$

so that in the limit $X(t) = \lim X_n(t)$ is $N(0, t)$.

Theorem 1. *For $t > s \geq 0$, the increment $X(t) - X(s)$ is independent of $X(s)$ and has distribution $N(0, t - s)$.*

The family of r.v. $X(t)$ is called the *Wiener process* or *Brownian motion.*

13.6 Problems.

(1.) Prove equation (13.3).

(2.) Prove equation (13.4).

(3.) Prove equation (13.5) by taking the limit as $q/p \to 1$.

(4.) Prove equation (13.5) by assuming a solution to (13.1) for $p = 1/2$ of the form $C + Dk$.

(5.) Solve Example 1 of section 1 if Nancy bets with $\$1000$ chips. Find the expected duration as well.

(6.) Prove equation (13.9).

(7.) Prove equation (13.11).

(8.) Prove equation (13.12).

(9.) Find the duration of a crap game. The rules of craps may be found in problem 10 of Chapter 3..

Appendix A Integrals.

A.1 The Gaussian Integral.

In this Appendix, we find the value of certain definite integrals which appear in the text. The first is the Gaussian integral, which is the normalization integral for the standard normal density.

Theorem 1.

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

First proof. Let

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

then

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2} dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2/2} dy \right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-y^2/2} dx dy \\ &= \int_0^{2\pi} e^{-r^2/2} r dr d\theta = 2\pi \int_0^{\infty} e^{-r^2/2} r dr = 2\pi. \end{aligned}$$

Second proof. We have

$$\left(\frac{I}{2} \right)^2 = \left(\int_0^{\infty} e^{-x^2/2} dx \right) \left(\int_0^{\infty} e^{-y^2/2} dy \right) = \int_0^{\infty} e^{-x^2/2} \left(\int_0^{\infty} e^{-y^2/2} dy \right) dx$$

In the inner y-integral, let $y = xt$ and $dy = xdt$ to obtain

$$\begin{aligned} \left(\frac{I}{2} \right)^2 &= \int_0^{\infty} e^{-x^2/2} \left(\int_0^{\infty} e^{-t^2 x^2/2} x dt \right) dx = \int_0^{\infty} \left(\int_0^{\infty} e^{-\left(1+t^2\right) x^2/2} x dx \right) dt \\ &= \int_0^{\infty} \frac{1}{1+t^2} dt = \frac{\pi}{2}. \blacksquare \end{aligned}$$

A.2 The Gamma Function.

The Gamma function is an extension of $n!$ to non-integer values of n.

Definition 1. For $p > 0$, define

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx$$

Note If $p \leq 0$, the integral is undefined due to the singularity at the origin.

Theorem 2. *(Properties of $\Gamma(p)$).*

(a.) $\Gamma(1) = 1$.

(b.) $\Gamma(p+1) = p\Gamma(p)$.

(c.) If n is an integer, then $\Gamma(n) = n!$.

(d.) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Proof. Part (a.) is trivial. For (b.), integration by parts gives

$$\Gamma(p+1) = \int_0^\infty x^p e^{-x} dx = \int_0^\infty x^p d\left[-e^{-x}\right] = \left[-x^p e^{-x}\right]_0^\infty + \int_0^\infty p x^{p-1} e^{-x} dx = p\Gamma(p)$$

since the boundary term vanishes.

Part (c.) follows from (b.) by induction.

For (d.), $\Gamma(\frac{1}{2})$ reduces to the Gaussian integral. Let $x = s^2/2$ and $dx = sds$ in the definition of $\Gamma(\frac{1}{2})$ to obtain

$$\begin{aligned} \Gamma(\frac{1}{2}) &= \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = \int_0^\infty \frac{\sqrt{2}}{s} e^{-s^2/2} sds \\ &= \sqrt{2} \int_0^\infty e^{-s^2/2} ds = \sqrt{2} \cdot \frac{\sqrt{2\pi}}{2} = \sqrt{\pi}. \square \end{aligned}$$

Example 1. Find $\Gamma(\frac{7}{2})$.

Solution.

$$\Gamma(\frac{7}{2}) = \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \Gamma(\frac{1}{2}) = \frac{15}{8}\sqrt{\pi}. \blacksquare$$

Example 2. Evaluate $\int_0^\infty x^5 e^{-x} dx$.

Solution.

$$\int_0^\infty x^5 e^{-x} dx = \Gamma(6) = 5! = 120. \blacksquare$$

Example 3. Evaluate $\int_0^\infty x^5 e^{-x^2/2} dx$.

Solution.

$$\int_0^\infty x^5 e^{-x^2/2} dx = \int_0^\infty x^4 e^{-x^2/2} x dx = \int_0^\infty (2s)^2 e^{-s} ds = 4 \int_0^\infty s^2 e^{-s} ds = 4 \cdot \Gamma(3) = 4 \cdot 2! = 8. \blacksquare$$

In a similiar manner, on obtains

Corollary 1.

$$\int_0^\infty x^p e^{-x^2/2} dx = 2^{(p-1)/2}\Gamma(\frac{p+1}{2}).$$

A.3 The Beta Function.

The Beta function is the normalization integral for the Beta density.

Definition 2. For $a, b > 0$, define

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

Theorem 3. For $a, b > 0$,

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

Proof. We have

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty x^{a-1}y^{b-1}e^{-(x+y)}dxdy$$

Letting $s = x + y$ with $y \le s < \infty$ and interchanging the order of integration gives

$$\int_0^\infty \int_s^\infty (s-y)^{a-1}\,y^{b-1}e^{-s}dsdy = \int_0^\infty \int_0^y (s-y)^{a-1}\,y^{b-1}e^{-s}dyds$$

Now let $y = st$, $dy = sdt$ with $0 \le t \le 1$ to get

$$\begin{aligned} & \int_0^\infty \int_0^1 s^{a-1}\,(1-t)^{a-1}\,s^{b-1}t^{b-1}e^{-s}sdtds \\ = & \int_0^\infty s^{a+b-1}e^{-s}\left(\int_0^1 (1-t)^{a-1}\,t^{b-1}dt\right)ds = B(a,b)\Gamma(a+b). \end{aligned}$$

Example 1. Evaluate $\int_0^1 x^3\,(1-x)^4\,dx$
Solution.

$$\int_0^1 x^3\,(1-x)^4\,dx = B(4,5) = \frac{\Gamma(4)\Gamma(5)}{\Gamma(5+4)} = \frac{3!4!}{8!} = \frac{1}{280}.$$

There are several other forms of the Beta Function.

Theorem 4.

(a.)

$$B(a,b) = 2\int_0^{\pi/2} (\sin t)^{2a-1} (\cos t)^{2b-1}\ dt.$$

(b.)

$$2^{a+b-1}B(a,b) = \int_{-1}^{1} (1-x)^{a-1}(1+x)^{b-1}dx.$$

(*Hint*: Let $x = 2s - 1$.)

(c.)

$$B(a,b) = \int_0^{\infty} \frac{s^{a-1}}{(1+s)^{a+b}}ds.$$

The proofs, which are achieved by making various changes of variable, are left to the problems.

For reference, we include here the following integral, obtained in section 8.8, in connection with *Student's t-distribution*.

Theorem 5.

$$\int_{-\infty}^{\infty} \frac{dx}{(1+x^2)^{p+1/2}} = \frac{1}{\sqrt{\pi}}\frac{\Gamma(p-\frac{1}{2})}{\Gamma(p)}.$$

A.4 Differentiating Indefinite Integrals.

We want to discuss differentiating integrals of the type

$$F(x) = \int_{a(x)}^{b(x)} f(t)\ dt$$

This requires two results

(1.) *The Fundamental Theorem of Calculus*. If $f(x)$ is continuous, the derivative of

$$F(x) = \int_a^x f(t)\ dt$$

is

$$F'(x) = f(t).$$

(2.) *The Chain Rule*. If $f(x)$ and $g(x)$ are differentiable, then the derivative of $F(x) = g(f(x))$ is

$$F'(x) = g'(f(x))f'(x).$$

Example 1. Differentiate

$$F(x) = \int_0^x t^2 e^{-t}\, dt$$

Solution. By the Fundamental Theorem,

$$F'(x) = x^2 e^{-x}.\blacksquare$$

Example 2. Differentiate

$$G(x) = \int_0^{x^3} t^2 e^{-t}\, dt$$

Solution. Write

$$G(x) = F(x^3)$$

where $F(x)$ is as in Example 1. By the Chain Rule,

$$G\prime(x) = F'(x^3)\frac{dx^3}{dx} = \left(x^3\right)^2 e^{-x^3} \cdot 3x^2 = 3x^8 e^{-x^3}.\blacksquare$$

A.5 Differentiation under the Integral Sign.

The derivative of an integral depending on a parameter may be found by differentiation under the integral sign. Formally, if

$$F(t) = \int_a^b f(x,t)\, dx.$$

then

$$F'(t) = \int_a^b \frac{\partial}{\partial t} f(x,t)\, dx.$$

Example 1. Find the derivative of

$$F(t) = \int_{-\infty}^{\infty} e^{-tx^2/2}\, dx.$$

Solution. We have

$$F'(t) = \int_{-\infty}^{\infty} \frac{\partial}{\partial t} e^{-tx^2}\, dx = -\int_{-\infty}^{\infty} x e^{-tx^2}\, dx.\blacksquare$$

A.6 Problems.

(1.) Show that

$$B(a,b) = 2\int_0^{\pi/2} (\sin t)^{2a-1} (\cos t)^{2b-1}\, dt.$$

(2.) Show that

$$\int_{-1}^{1}(1-x)^{a-1}(1+x)^{b-1}dx = 2^{a+b-1}B(a,b).$$

(*Hint*: Let $x = 2s-1$.)

(3.) Show that

$$B(a,b) = \int_0^{\infty} \frac{s^{a-1}}{(1+s)^{a+b}}ds.$$

(*Hint*: Let $x = 1/(s+1)$.)

(4.) Evaluate the following integrals.

(a.) $\int_0^{\infty} x^5 e^{-x}dx$

(b.) $\int_0^{\infty} x^3 e^{-4x}dx.$

(c.) $\int_0^{\infty} x^4 e^{-x^2}dx.$

(d.) $\int_0^1 x^4 (1-x)^5 \, dx.$

(e.) $\int_2^4 (x-2)^3 (4-x)^4 \, dx.$

(f.) $\int_{-1}^{1}(1-x)^3(1+x)^5dx.$

(g.) $\int_0^{\pi/2} (\sin t)^3 (\cos t)^5 \ dt.$

(h.) $\int_0^{\infty} s^2/(1+s)^6 \, ds.$

(i.) $\int_{-\infty}^{\infty} \frac{dx}{(1+x^2)^4}$.

(j.) $\int_{-\infty}^{\infty} \frac{dx}{\sqrt{(1+x^2)^3}}$

(5.) Differentiate the following.

(a.) $\int_0^x \frac{\sin t}{t}dt.$

(b.) $\int_x^2 e^{-t^2}dt.$

(c.) $\int_0^{x^2} e^{-t^2}dt.$

(d.) $\int_x^{x^2} \sin(t^2)dt.$

(6.) If

$$F(t) = \int_{a(t)}^{b(t)} f(t,x) \, dx,$$

show that

$$F'(t) = f(t, b(t))b'(t) - f(t, a(t))a'(t) + \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(t, x)\, dx$$

(7.) Use problem 6 to differentiate

$$F(t) = \int_{t^2}^{t^3} \frac{1}{1 + x^2 t^2} dx.$$