MATH 2P81
PROBABILITY
Lecture Notes

Contents

Part I

DISCRETE PROBABILITY

Chapter 1 COUNTING AND RELATED TOPICS

Permuting objects

In how many possible ways can we arrange

▶ n Distinct Objects◀

(such as a, b, c, d,....) in a row?

We start with n empty boxes which we fill, one by one, from left to right. We have n choices to fill the first box, once it's done there are $n-1$ choices of how to fill the second box, $n-2$ choices for the third box, ... until we come to the last box, having only 1 object (letter) left to put in. The choices obviously multiply, as each 'word' created at any stage of the process is unique (no duplicates). Thus the answer is $n \times (n-1) \times (n-2) \times ..... \times 3 \times 2 \times 1 \stackrel{def.}{=} n!$

What if some of these object are

▶Indistinguishable◀

such as, for example *aaabbc*. How many *distinct* permutations of these letters are there, i.e. how many distinct words can we create by permuting *aaabbc*? We can start by listing *all* 6! permutations, and then establishing how many times each distinct word appears on this list (the amount of its 'duplicity' – one should really say 'multiplicity'). Luckily enough, the 'duplicity' of each distinct word proves to be the same. We can thus simply divide 6! by this common duplicity to get the final answer.

To get the duplicity of a particular word, such as, for example *baacba* we first attach a unique index to each letter: $b_1a_1a_2c_1b_2a_3$ and then try to figure out the number of permutations of these, now fully distinct, symbols, which keeps the actual *word* (*baacba*) intact. This is obviously achieved by permuting the a's among themselves, the b's among themselves, etc. We can thus create 3! (number of ways of permuting the a's) *times* 2! (permuting the b's) combinations which are distinct in the original 6!-item list, but represent the same word now. (We have multiplied 3! by 2! since every permutation of the a's combines with every permutation of the b's to provide a *unique* combination of the indexed letters).

The answer is thus

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

(we have included 1! to indicate that there is only one permutation of the single c, to make the formula complete). The resulting expression is so important to us that we introduce a new symbol

$$\frac{6!}{3!2!1!} \stackrel{def.}{=} \binom{6}{3,2,1}$$

which we read: 6 choose 3 choose 2 choose 1 (note that the bottom numbers must add up to the top number). It is obvious that the same argument holds for any other unique word of the *aaabbc* type.

It should now be obvious that, in the case of permuting n_1 a's, n_2 b's, n_3 c's,.....n_k z's, we will get

$$\frac{N!}{n_1!n_2!n_3!.....n_k!} \stackrel{def.}{=} \binom{N}{n_1, n_2, n_3, ..., n_k} \qquad (*)$$

distinct words (where $N = \sum_{i=1}^{k} n_i$ which is the total word length). These numbers are called MULTINOMIAL COEFFICIENTS (later, we will learn why).

Selecting objects

The **basic question** of this section is:

In how many ways can we select r out of n distinct objects (letters)?

This question is actually *ambiguous*, in two ways:

1. Are we allowed to select the same letter more than once?
2. Do we care about the order in which the selection is made? ■

Depending on the answer, we end up with four distinct results.

EXAMPLE: Suppose $n = 3$ (letters a, b, c), and $r = 2$. Then, if:

- Order is important but we must not repeat letters − the answer is 6 (ab, ac, ba, bc, ca, cb).
- Unordered selection, without replicates − answer: 3 ($ab \equiv ba$, $ac \equiv ca$, $bc \equiv ca$) [note that unordered selections can be always arranged alphabetically; insisting on that enables us to avoid accidental duplication].
- Order is important, each letter can be selected repeatedly − answer: 9 (aa, ab, ac, ba, bb, bc, ca, cb, cc).
- Unordered selection, allowed duplicating letters − answer: 6 (aa, $ab \equiv ba$, $ac \equiv ca$, bb, $bc \equiv cb$, cc). ■

Can we figure out the general formula (with any r and n) for each one of these four possibilities? Let us try it.

▶Ordered Selection, No Duplication◀

(Selecting a chair, treasurer and secretary out of ten members of a club).

Since the result should be ordered, we can start with r empty boxes, and fill them one by one, counting (and multiplying) the choices:

1^{st} box	2^{nd} box	3^{rd} box	.....	(r-1)th box	rth box
$n\times$	$(n-1)\times$	$(n-2)\times$	.....	$(n-r+2)\times$	$(n-r+1)$

The result is thus

$$n \times (n-1) \times (n-2) \times .... \times (n-r+1) = \frac{n!}{(n-r)!} \stackrel{def.}{=} P_r^n \tag{1}$$

called the NUMBER OF PERMUTATIONS.

With $n = 3$ and $r = 2$ this gives $\frac{3!}{(3-2)!} = 6$ (check).

Just to practice: $P_4^{10} = 10 \times 9 \times 8 \times 7 = 5040$ (start with 10, for the total of 4 factors).

▶Unordered Selection, Without Duplication◀

(Selecting a committee of three, out of ten members represented by 10 distinct letters).

If we examine the previous list of possibilities (we did not really build it, but we all must be able to *visualize* it), we notice that each unique selection of r letters is repeated exactly $r!$ times (it will be there with all its permutations, since these were considered distinct). All we have to do is to remove this duplicity by dividing the previous answer by $r!$, obtaining

$$\frac{P_r^n}{r!} = \frac{n!}{(n-r)!r!} \stackrel{def.}{=} C_r^n \tag{2}$$

(NUMBER OF COMBINATIONS). Later on, these will also be called BINOMIAL COEFFICIENTS. Note the symmetry of this formula: selecting 7 people out of 10 can be done in the same number of ways as selecting 3 (and telling them: you did *not* make it). With $n = 3$ and $r = 2$ we get $\frac{3!}{2!1!} = 3$ (check).

Just to practice: $C_{13}^{17} = C_4^{17} = \frac{17 \times 16 \times 15 \times 14}{4 \times 3 \times 2 \times 1}$ (same number of factors, when you include 1).

▶Ordered Selection, Duplication of Letters Allowed◀

(Building a five-letter word, using an alphabet of 26 letters).

Again, fill r empty boxes with one letter each. This time we have a choice of n letters *every time*. So the answer is $n \times n \times n \times ..... \times n$ (r times), namely

$$n^r \tag{3}$$

With $n = 3$ and $r = 2$ we get $3^2 = 9$ (check).

▶Unordered Selection, Allowing Duplication◀

(Choose ten pieces of fruit from a shelf full of apples, pears, oranges and bananas).

This is the most difficult case (we cannot use the previous list, as the duplicity of a specific unordered selection varies from case to case), so we first solve our specific example, and then generalize the result.

We start with our $r = 10$ boxes (to contain one piece of fruit each), but to assure an *unordered* selection, we insist that apples go first, pears second, and so on. To determine how may boxes get an apple, we place a bar after the last 'apple' box, similarly with pears, etc. For example: □□|□|□□□□□|□□ means getting 2 apples, 1 pear, 5 oranges and 2 bananas. Note that we can place the bars *anywhere* (with respect to the boxes), such as: □□□| |□□□□□□□| (we don't like pears and bananas). Also note that it will take exactly $3 = n - 1$ bars to complete our 'shopping list'. Thus any permutation of $3 = n - 1$ bars and $10 = r$ boxes corresponds to a particular selection (at the same time, a distinct permutation represents a distinct choice $\Rightarrow$ there is a one-to-one correspondence between these permutations and a complete list of fruit selections). We have already solved the problem of distinct permutations (the answer is $C_3^{13} = 286$), so that is the number of options we have now. The general formula is obviously

$$C_{n-1}^{r+n-1} \equiv C_r^{r+n-1} \tag{4}$$

With $n = 3$ and $r = 2$ this gives $C_2^4 = 6$ (check).

We have thus derived a formula for each of the four possibilities. Your main task is to be able to correctly decide *which* of these *to use* in each particular situation.

EXAMPLE: Let us re-derive (*) by taking the following approach: To build an N-letter word out of n_1 a's, n_2 b's, ...n_k z's, we start with N empty boxes, then choose n_1 of these to receive the letter a, having done that we choose (from the remaining $N - n_1$ boxes) n_2 boxes for the letter b, and so on (multiplying the number of choices to get the final answer). Which of our four formulas do we use (at each stage of the selection)? Well, we have to select n_i *distinct* boxes (i.e. duplication is *not* allowed), and we don't care about the order (selecting Boxes 1, 4 and 7 for the letter a is the same as selecting Boxes 4, 7 and 1). We thus use Formula (2), to get:

$$\binom{N}{n_1} \times \binom{N - n_1}{n_2} \times ..... \binom{n_k}{n_k}$$

How come this differs from our original answer (*). Only seemingly, when expanded, the new formula gives

$$\frac{N!}{n_1!(N-n_1)!} \times \frac{(N-n_1)!}{n_2!(N-n_1-n_2)!} \times ... \times 1 = \frac{N!}{n_1!n_2!....n_k!} \qquad \text{(check)}$$

Note that $\binom{N}{n_1, n_2}$ yields the regular binomial coefficient, usually written as $\binom{N}{n_1}$. This notational exception is allowed only in the case of $k = 2$. ■

Now, let us put these formulas to work, first to derive the so called

Binomial expansion

You probably all know that

$$\begin{aligned}(x+y)^2 &= x^2+2xy+y^2\\ (x+y)^3 &= x^3+3x^2y+3xy^2+y^3\\ &\cdots\cdots\cdots\\ (x+y)^n &= \binom{n}{0}x^n+\binom{n}{1}x^{n-1}y+\binom{n}{2}x^{n-2}y^2+....\binom{n}{n}y^n\end{aligned}$$

the general (last-line) expansion usually written as

$$(x+y)^n=\sum_{i=0}^{n}\binom{n}{i}x^{n-i}y^i \qquad \text{(BE)}$$

We have to remember that $0!\overset{def.}{=}1$, so that $\binom{n}{0}=\binom{n}{n}=1$. A whole row of these coefficients can be easily constructed as follows (Pascal's triangle):

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
=\binom{5}{0} & & =\binom{5}{1} & & =\binom{5}{2} & & =\binom{5}{3} & & =\binom{5}{4} & & =\binom{5}{5}
\end{array}$$

To prove BE, we need only basic algebra:

$(x+y)^n=(x+y)(x+y).....(x+y)$ [n times] = [distributive law]

$$\begin{array}{l}=xxx...x\\ +yxx...x\\ +xyx...x\\ \quad.....\\ +yyy...y\end{array}\qquad\left\{\begin{array}{l}\text{This is a list of } \textit{all}\ n\text{-letter words}\\ \text{made up of letters } x \text{ and } y.\\ \text{Formula (3) tells us that there are}\\ 2^n \text{ of them.}\end{array}\right.$$

Let us now combine the terms of this sum which are algebraically identical: x^n [only one word will have all x's] $+x^{n-1}y\times\binom{n}{1}$ [this many words will have one x] $+....+x^{n-i}y^i\times\binom{n}{i}$ [this many n-letter words have exactly i x's] $+....+y^n$. This is the binomial expansion.

In summary: the essential part of the proof was knowing how many n-letter words have exactly i x's in them. Formula (*) with $k=2$ supplied the answer. □

▶Binomial-Expansion Extras◀

When we need the binomial expansion, it is usually *not* with just x and y, but with something like:

$$(1-3x^2)^n=1-3nx^2+9\binom{n}{2}x^4-27\binom{n}{3}x^6+.....+(-3x^2)^n$$

This indicates that all our formulas will be useful to us only when we learn how to *apply* them (just memorizing them would be useless).

The binomial expansion can be **extended** to a **non-integer** n (I will call it β, just to distinguish) when $y = 1$:

$$(1+x)^\beta = 1 + \beta x + \binom{\beta}{2}x^2 + \binom{\beta}{3}x^3 + ...$$

where $\binom{\beta}{2} = \frac{\beta(\beta-1)}{2}$, $\binom{\beta}{3} = \frac{\beta(\beta-1)(\beta-2)}{3!}$, etc. This time the expansion is infinite and its proof requires Maclaurin formula (not just simple algebra) − try it.

EXAMPLES:

$$\begin{aligned}(1+x)^{-3} &= 1 - 3x + 6x^2 - 10x^3 + ... \\ (1+a)^{\frac{3}{2}} &= 1 + \frac{3}{2}a + \frac{3}{8}a^2 - \frac{1}{16}a^3 + \frac{3}{128}a^4 + ...\end{aligned}$$

Understand the construction of the individual coefficients, such as $\binom{-3}{3} = \frac{(-3)(-4)(-5)}{6} = -10$ and $\binom{\frac{3}{2}}{4} = \frac{\frac{3}{2}\times\frac{1}{2}\times\frac{-1}{2}\times\frac{-3}{2}}{24} = \frac{3}{128}$ ■

Multinomial expansion

is an extension of the binomial expansion, having more than 2 terms inside parentheses. We will derive it using three terms, the extension to any other number of terms is then quite obvious.

We want to generalize the well known: $(x+y+z)^2 = x^2+y^2+z^2+2xy+2xz+2yz$ to: $(x+y+z)^n = (x+y+z)(x+y+z)...(x+y+z)$ [n factors] = [distributive law] $xxx...x + yxx...x + ... + zzz...z$ [all 3^n n-letter words built out of x, y and z] = [collecting algebraically identical contributions:] $x^n + \binom{n}{n-1}x^{n-1}y + ... + \binom{n}{n-5,3,2}x^{n-5}y^3z^2 + ... + z^n$ [the coefficients representing the number of words with the corresponding number of x's, y's and z's] =

$$\sum_{\substack{i,j,k\geq 0 \\ i+j+k=n}} \binom{n}{i,j,k} x^i y^j z^k$$

where the summation is over all possible selections of non-negative exponents which add up to n.

How many **terms** are there in this summation? Our Formula (4) tells us that it should be $\binom{n+2}{2}$ − the three exponents are chosen in the apple-pear-orange like manner. This is because when choosing the x's, y's and z' we don't care about the order, and we are allowed to duplicate.

Similarly

$$(x+y+z+w)^n = \sum \binom{n}{i,j,k,\ell} x^i y^j z^k w^\ell$$

where the summation is over all non-negative integers having the correct sum n. There are altogether $\binom{n+3}{3}$ terms in the last expansion.

EXAMPLES:

- $(x+y+z)^3 = x^3+y^3+z^3+3x^2y+3x^2z+3xy^2+3xz^2+3y^2z+3yz^2+6xyz$ [has $10 = \binom{3+2}{2}$ terms − check].

- Typical exam question: Find the coefficient of ut^3 in the expansion of $(u+2-4t)^5$.

  Solution: The only term containing ut^3 is: $\binom{5}{1,1,3}(u)^1(2)^1(-4t)^3 = -2560ut^3$.

  Answer: -2560.

- Another exam-like question: $(1+2x-5x^2)^{17}$ is a 34-degree polynomial in x. When expressed as such (i.e. when expanded, and terms with like powers of x are combined), what will be the coefficient of x^4?

  Solution: $\binom{17}{i,j,k}(1)^i(2x)^j(-5x^2)^k$ is the general term of this expansion. Let us make a table of the exponents which contribute to x^4:

i	j	k
13	4	0
14	2	1
15	0	2

  . This translates to: $\binom{17}{4}(2x)^4 + \binom{17}{14,2,1}(2x)^2(-5x^2) + \binom{17}{2}(-5x^2)^2 = 680x^4$.

  Answer: 680. ■

Related issues

Our formulas will enable us to settle yet another issue:

►Partitioning◄

of n distinct objects (people) into several groups (teams) of given size (not necessarily the same for each group).

Suppose, for example, that nine people are to be divided into three teams of 2, 3, and 4 members. In how many ways can this be done? This is solved by realizing that there is a one-to-one correspondence between these and permutations of *aabbbcccc*. Let us look at one such case:

1	2	3	4	5	6	7	8	9
b	a	c	c	a	b	c	b	c

where the *position* of each letter correspond to a specific person, and the letter itself indicates which team he joins. Our table would thus assign Persons 2 and 5 to the first team (of two people), Persons 1, 6 and 8 to the second team (of three people), and Persons 3, 4, 7 and 9 to the third team (of four people). The number of possible assignments must be therefore equal to the number of such permutations, which is $\binom{9}{2,3,4} = 1260$.

There is a bit of a problem when some groups are of the same size, such as dividing 9 people into three groups of 3 people each. The routine answer gives $\binom{9}{3,3,3} = 1680$, but does this consider the following two assignments distinct or identical: $1,2,3 \mid 4,5,6 \mid 7,8,9$ and $4,5,6 \mid 1,2,3 \mid 7,8,9$? Anyone who is following our line of reasoning should clearly see that the formula does consider these two as *distinct* (because *aaabbbccc* and *bbbaaaccc* are two distinct *words*, and that's how the formula works). If, for whatever reason, we want to consider such

team assignments identical (and the choice is *ours*), we have to 'fix' the formula, thus:

$$\frac{\binom{9}{3,3,3}}{3!} = 280$$

Or, in the more general case:

$$\frac{\binom{20}{2,2,2,3,3,4,4}}{3!2!2!} = 61108047000$$

I hope everyone understands the logic of the last answer.

Finally, a ▶Circular Arrangement◀

means placing n distinct objects (people) around a circular table (rather than in a row). If we only care about who sits next to whom (differentiating left and right), but not about the orientation of the whole arrangement within the room, we have $(n-1)!$ possible ways of seating the people. We can see this by starting with n empty chairs, placing $Mr.\ A$ in one of them (he has no neighbors yet) and then having $n-1$ choices to fill the chair to his right, $n-2$ choices to fill the next chair, ... until we have one person waiting to become his left-hand neighbor..

End-of-chapter examples

1. A college team plays a series of 10 games which they can either win (W), lose (L) or tie (T).

   (a) How many possible outcomes can the series have (differentiating between WL and LW, i.e. order is important).
   Answer: $3^{10} = 59049$.

   (b) How many of these have exactly 5 wins, 4 losses and 1 tie?
   Answer: $\binom{10}{5,4,1} = 1260$.

   (c) Same as (a) if we *don't* care about the order of wins, losses and ties?
   Answer: $\binom{12}{2} = 66$ (only one of these will have 5 wins, 4 losses and 1 tie).

2. A student has to answer 20 true-false questions.

   (a) In how many distinct ways can this be done?
   Answer: $2^{20} = 1048576$.

   (b) How many of these will have exactly 7 correct answers?
   Answer: $\binom{20}{7} = 77520$.

   (c) *At least* 17 correct answers?
   (Here, these is *no 'shortcut'* formula, we have to do this individually, one by one, adding the results): $\binom{20}{17} + \binom{20}{18} + \binom{20}{19} + \binom{20}{20} = \binom{20}{3} + \binom{20}{2} + \binom{20}{1} + \binom{20}{0} = 1351$.

(d) *Fewer* than 3? (*excludes* 3): $\binom{20}{0} + \binom{20}{1} + \binom{20}{2} = 211$.

3. In how many ways can 3 Americans, 4 Frenchmen, 4 Danes and 2 Canadians be seated (here we are particular about nationalities, but not about individuals)

   (a) in a row.

   Answer (same as the number of permutations of $AAAFFFFDDDDCC$): $\binom{13}{3,4,4,2} = \binom{13}{3}\binom{10}{4}\binom{6}{4} = 900900$.

   (b) In how many of these will people of the same nationality sit together?

   Answer: We just have to arrange the four nationalities, say a, f, d and c: $4! = 24$.

   (c) Repeat (a) with circular arrangement:

   Answer (13 of the original arrangements are duplicates now, as $AAAFFFFDDDDCC$, $AAFFFFDDDDCCA$, ..., $CAAAFFFFDDDDC$ are identical): $\frac{900900}{13} = 69300$.

   (d) Repeat (b) with circular arrangement:

   Answer (circular arrangement of nationalities): $3! = 6$.

4. Four couples (Mr&Mrs A, Mr&Mrs $B, ...$) are to be seated at a round table.

   (a) In how many ways can this be done?

   Answer: $7! = 5040$.

   (b) How many of these have all spouses sit next to each other?

   Solution: First we have to arrange the families with respect to each other. This can be done in 3! ways. Then, having two seats reserved for each couple, we have to decide the mutual position of every wife and husband ($2 \times 2 \times 2 \times 2$).

   Answer: $3! \times 2^4 = 96$. (Later on, our main task will be converting these to probabilities. If the seating is done randomly, the probability of keeping the spouses together will be then $\frac{96}{5040} = 1.905\%$).

   (c) How many of these have the men and women alternate?

   Solution: Place Mr A into one chair, then select his right-hand neighbor (must be a woman) in 4 ways, select her extra neighbor (3 ways), and so on.

   Answer: $4 \times 3 \times 3 \times 2 \times 2 \times 1 \times 1 = 144$ (corresponds to 2.86%).

   (d) How many of these have the men (and women) sit together?

   Solution: This is analogous to (b). We have to arrange the two groups (men and women) with respect to each other first. But, in the circular arrangement, this can be done in one way only! Then, we have to take care of arranging the 4 men and four women within the four chairs allocated to them. This can be done in 4! ways each.

   Answer: $(4!)^2 = 576$ (correspond to 11.43%).

5. In how many ways can we put 12 books into 3 shelves? This question is somehow ambiguous: do we want to treat the books as distinct or identical, and if we do treat them as distinct, do we care about the order in which they are placed within a shelf? The choice is ours, let's try it each way:

   (a) It the books are treated as 12 identical copies of the same novel, then the only decision to be made is: how many books do we place on each shelf (the shelves are obviously distinct, and large enough to accommodate all 12 books if necessary).

   The answer follows from Formula (4) with $n = 3$ and $r = 12$ - for each book we have to select a shelf, but the order does not matter ($1, 3$ and $3, 1$ puts one book each on Shelf 1 and 3), and duplication is allowed: $\binom{14}{2} = 91$

   (b) If we treat the books as distinct and their order within each shelf important, we solve this in two stages:

   First we decide how many books we place in each shelf, which was done in (a), *then* we choose a book to fill, one by one, each allocated slot (here we have $12 \times 11 \times 10 \times .... \times 2 \times 1$ choices).

   Answer: $91 \times 12! = 43,589,145,600$.

   (c) Finally, if the books are considered all distinct, but their arrangement within each shelf is irrelevant, we simply have to decide which shelf will each book go to [similar to (a), order important now].

   This can be done in $3 \times 3 \times 3 \times .... \times 3 = 3^{12} = 531441$ number of ways.

6. Twelve men can be seated in a row in $12! = 479001600$ number of ways (trivial).

   (a) How many of these will have Mr A and Mr B sit next to each other?

   Solution: Mr A and Mr B have to be *first* treated as a single item, for a total of 11 items. These can be permuted in 11! number of ways. *Secondly*, we have to place Mr A and Mr B in the two chairs already allocate to them, in 2! ways.

   Answer: $2 \times 11! = 79833600$.

   (b) How many of the original arrangements will have Mr A and Mr B sit *apart*?

   This set consists of those which did not have them sit next to each other, i.e. $(a) - (b) = 12! - 2 \times 11! = 399168000$.

   (c) How many of the original arrangements will have *exactly* 4 people sit *between* Mr A and Mr B?

   Solution: *First*, we allocate two chairs for Mr A and Mr B, thus: ■□□□□■□□□□□□, □■□□□□■□□□□□,....,□□□□□□■□□□□■, altogether in 7 possible ways (here we count using our fingers − no fancy formula). *Secondly*, we seat the people. We have 10! choices for filling the □□...□ chairs, times 2 choices for how to fill ■ and ■.

   Answer: $7 \times 2 \times 10! = 50803200$.

7. Security council of the UN has 15 permanent members, US, Russia, GB, France and China among them. These can be seated in a row in 15! possible arrangements.

   (a) How many of these have France and GB sit together but (at the same time) US and Russia sit apart?

   Solution: We break the problem into two parts:

   i. France and GB sit together in $2 \times 14! = 174,356,582,400$ of the original 15! arrangements (we understand the logic of this answer from the previous question).

   ii. France and GB sit together and (at the same time) US and Russia sit together in $2 \times 2 \times 13! = 24,908,083,200$ arrangements (similar logic: first we create two groups of two, one for France/GB, the other for US/Russia and permute the resulting 13 *items*, then we seat the individual people).

   The answer is obviously the difference between the two: $2 \times 14! - 2^2 \times 13! = 149,448,499,200$. (To make the answer more meaningful, convert it to probability).

8. Consider the standard deck of 52 cards (4 suits: hearts, diamonds, spades and clubs, 13 'values': 2, 3, 4...10, Jack, Queen, King, Ace). Deal 5 cards from this deck. This can be done in $\binom{52}{5} = 2598960$ distinct ways (trivial).

   (a) How many of these will have *exactly* 3 diamonds?

   Solution: First select 3 diamonds and two 'non-diamonds', then combine these together, in $\binom{13}{3} \times \binom{39}{2} = 211926$ number of ways.

   (b) *Exactly* 2 aces?

   Same logic: $\binom{4}{2} \times \binom{48}{3} = 103776$.

   (c) Exactly 2 aces and 2 diamonds?

   This is slightly more complicated because there is a card which is both an ace and a diamond. The deck must be first divided into four parts, the ace of diamonds (1 card) the rest of the aces (3), the rest of the diamonds (12), the rest of the cards (36). We then consider two cases, either the ace of diamonds is included, or not. The two individual answers are added, since they are mutually incompatible (no 'overlap'): $\binom{1}{1}\binom{3}{1}\binom{12}{1}\binom{36}{2} + \binom{1}{0}\binom{3}{2}\binom{12}{2}\binom{36}{1} = 29808$.

9. In how many ways can we deal 5 cards each to 4 players?

   (a) Answer: $\binom{52}{5} \times \binom{47}{5} \times \binom{42}{5} \times \binom{37}{5} = 1.4783 \times 10^{24}$

   (b) So that each gets exactly one ace?

   Answer (consider dealing the aces and the non-aces separately): $\binom{4}{1}\binom{3}{1}\binom{2}{1}\binom{1}{1} \times \binom{48}{4}\binom{44}{4}\binom{40}{4}\binom{36}{4} = 3.4127 \times 10^{21}$

   (c) None gets any ace:

   Answer: $\binom{48}{5}\binom{43}{5}\binom{38}{5}\binom{33}{5} = 1.9636 \times 10^{23}$

(d) Mr A gets 2 aces, the rest get none.
Answer: $\binom{4}{2} \times \binom{48}{3}\binom{45}{5}\binom{40}{5}\binom{35}{5} = 2.7084 \times 10^{22}$

(e) (Any) one player gets 2 aces, the other players get none.
Solution: The previous answer is correct whether it is Mr A, B, C or D who gets the 2 aces (due to *symmetry*), all we have to do is to add the four (identical) numbers, because the four corresponding sets cannot overlap, i.e. are mutually incompatible or EXCLUSIVE).
Answer: $4 \times 2.7084 \times 10^{22} = 1.0834 \times 10^{23}$

(f) Mr. A gets 2 aces.
Answer: $\binom{4}{2}\binom{48}{3} \times \binom{47}{5}\binom{42}{5}\binom{37}{5} = 5.9027 \times 10^{22}$. Note that when computing the probability of this happening, the $\binom{47}{5}\binom{42}{5}\binom{37}{5}$ part cancels out (we can effectively deal 5 cards to him and stop).

(g) Mr. C gets 2 aces.
Solution: If he is the third player to be dealt his cards, we can either do this to long and impractical way (taking into account how many aces have been dealt to Mr A and Mr B), thus: $\binom{48}{5}\binom{43}{5} \times \binom{38}{3}\binom{4}{2} \times \binom{37}{5} + \binom{48}{4}\binom{4}{1}\binom{44}{5} \times \binom{39}{3}\binom{3}{2} \times \binom{37}{5} + \binom{48}{5}\binom{43}{4}\binom{4}{1} \times \binom{39}{3}\binom{3}{2} \times \binom{37}{5} + \binom{48}{4}\binom{4}{1}\binom{44}{4}\binom{3}{1} \times \binom{40}{3} \times \binom{37}{5} + \binom{48}{3}\binom{4}{2}\binom{45}{5} \times \binom{40}{3} \times \binom{37}{5} + \binom{48}{5}\binom{43}{3}\binom{4}{2} \times \binom{40}{3} \times \binom{37}{5} = 5.9027 \times 10^{22}$, or be smart and argue that, due to the symmetry of the experiment, the answer must be the same as for Mr. A.

(h) At least one player gets 2 aces (regardless of what the others get).
This is quite a bit more difficult, to the extend that we must postpone solving it.

10. (Game of **Poker**): 5 cards are dealt from an ordinary deck of 52. The total number of possible outcomes (5-card hands) is $\binom{52}{5} = 2598960$ (trivial). How many of these contain exactly

(a) One pair, i.e. two identical values (and no other duplication of values).
Solution: This is done in two stages, first we select the suit to be represented by a pair and three distinct suits to be represented by a singlet each: $\binom{13}{1} \times \binom{12}{3}$, then we select two individual cards from the first suit: $\binom{4}{2}$ and one card each from the other 3 suits: 4^3.
Answer: $\binom{13}{1}\binom{12}{3}\binom{4}{2}4^3 = 1098240$.

(b) Two pairs.
Following the same logic: $\binom{13}{2}\binom{11}{1} \times \binom{4}{2}^2 \times 4 = 123552$

(c) A triplet: $\binom{13}{1}\binom{12}{2} \times \binom{4}{3} \times 4^2 = 54912$

(d) Full house (a pair and a triplet): $\binom{13}{1}\binom{12}{1} \times \binom{4}{3} \times \binom{4}{2} = 3744$

(e) Four of a kind: $\binom{13}{1}\binom{12}{1} \times \binom{4}{4}\binom{4}{1} = 624$

(f) A straight (five consecutive values − ace can be considered both as the highest and the lowest value, i.e. Ace, 2, 3, 4, 5 is a straight).

Solution: There are 10 possibilities as to the sequence of values (starting from Ace...5, up to 10...Ace), once this is chosen, one has to select the individual cards: $4 \times 4 \times 4 \times 4 \times 4$.

Answer: $10 \times 4^5 = 10240$.

(g) Flush (five cards of the same suit).

Solution: 4 ways of selecting the suit, $\binom{13}{5}$ ways of selecting the individual cards from it.

Answer: $4 \times \binom{13}{5} = 5148$.

(h) We should note that a hand can be *both a straight and a flush* (a first overlap encountered so far).

We have again 10 possibilities for the values, but only 4 ways of selecting the cards (they must be of the same suit). The number of these hands is thus $10 \times 4 = 40$.

(i) None of the above.

Solution: First we select five *distinct* values, disallowing the 10 cases resulting in a straight: $\binom{13}{5} - 10$, then we select one card of each chosen value, disallowing a flush, which happens in only 4 of these cases: $4^5 - 4$.

Answer: $\left(\binom{13}{5} - 10\right) \times (4^5 - 4) = 1302540$.

One can verify that adding all these answers, except for (h) which needs to be *subtracted* (why?), results in the correct total of 2598960 (check).

11. Roll a die five times. The number of possible (ordered) outcomes is $6^5 = 7776$ (trivial). How many of these will have:

(a) One pair of identical values(and no other duplicates).

Solution: First we choose the value which should be represented twice and the three values to go as singles: $\binom{6}{1} \times \binom{5}{3}$, then we decide how to place the 5 selected numbers in the five blank boxes, which can be done in $\binom{5}{2,1,1,1}$ ways (equal to the number of *aabcd* permutations).

Answer: $\binom{6}{1} \times \binom{5}{3} \times \binom{5}{2,1,1,1} = 3600$.

(b) Two pairs.

The same logic gives: $\binom{6}{2}\binom{4}{1} \times \binom{5}{2,2,1} = 1800$.

(c) A triplet: $\binom{6}{1}\binom{5}{2} \times \binom{5}{3,1,1} = 1200$.

(d) 'Full house' (a triplet and a pair): $\binom{6}{1}\binom{5}{1} \times \binom{5}{3,2} = 300$.

(e) 'Four of a kind': $\binom{6}{1}\binom{5}{1} \times \binom{5}{4,1} = 150$.

(f) 'Five of a kind': $\binom{6}{1} \times \binom{5}{5} = 6$.

(g) Nothing.

Solution: We again fill the empty boxes, one by one, avoiding any duplication: $6 \times 5 \times 4 \times 3 \times 2 = 720$.

Note that all these answers properly add up to 7776 (check).

12. Let us try the same thing with 15 rolls of a die ($6^{15} = 4.7018 \times 10^{11}$ outcomes in total). How many of these will have:

    (a) A quadruplet, 2 triplets, 2 pairs and 1 singlet:
    $\binom{6}{1}\binom{5}{2}\binom{3}{2}\binom{1}{1} \times \binom{15}{4,3,3,2,2,1} = 6.8108 \times 10^{10}$

    (b) 3 triplets and 3 pairs: $\binom{6}{3}\binom{3}{3} \times \binom{15}{3,3,3,2,2,2} = 1.5135 \times 10^{10}$.

    We will not try to complete this exercise; the full list would consist of 110 possibilities.

Chapter 2 RANDOM EXPERIMENTS

A few examples

1. Rolling a die

2. Rolling 2 (n in general) dice (or, equivalently, one die twice, or n times)

3. Selecting 2 people out of 4 (k objects out of n in general)

4. Flipping a coin until a head appears

5. Rotating a wheel with a pointer

6. Flipping a tack ($\perp$)

Sample space

is a collection of all possible outcomes of an experiment. The individual (*complete*) outcomes are called **simple events**. For the six examples above, we get:

1. The outcomes can be uniquely represented by the number of dots shown on the top face. The sample space is thus the following set of six elements: $\{1, 2, 3, 4, 5, 6\}$.

2. With two dice, we have a decision to make: do we want to consider the dice as indistinguishable (to us, they usually are) and have the sample space consist of *unordered* pairs of numbers, or should we mark the dice (red and green say) and consider an *ordered* pair of numbers as an outcome of the experiment (the first number for red, the second one for green die)? The choice is ours; we are allowed to consider as much or as little detail about the experiment as we need, but there two constraints:

   (a) We have to make sure that our sample space has enough information to answer the questions at hand (if the question is: what is the probability that the red die shows a higher number than the green die, we obviously need the *ordered* pairs).

   (b) Subsequently, we learn how to assign probabilities to individual outcomes of a sample space. This task can quite often be greatly simplified by a convenient design of the sample space.. It just happens that, when rolling two dice, the simple events (pairs of numbers) of the sample space have the same simple probability of $\frac{1}{36}$ when they are ordered; assigning correct probabilities to the unordered list would be extremely difficult. That is why, for this kind of experiment (rolling a die any fixed number of times), we always choose the sample space to consist of an *ordered* set of numbers (whether the question requires it or not).

In the case of two dice, we will thus use the following (conveniently organized) sample space:

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,1	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

and correspondingly for more than 2 dice (we will no longer be able to write it down explicitly, but we should be able to *visualize* the result). Note that a *single* simple event consists of *two* (or more) numbers. As explained earlier, we will never try to simplify this sample space by removing the order; there is one simplification one *can* make though, if the question is concerned only with *sixes* versus *non-sixes*: we can reduce the sample space of 36 simple events to: $\{66, 6O, O6, OO\}$ where O stands for any *other* number but 6. Assigning probabilities will be a touch more difficult now, but it will prove to be manageable.

3. Selecting 2 (*distinct*) people out of 4. Here (unless the question demands it), we can ignore the order of the selection, and simplify the sample space to: $\{AB, AC, AD, BC, BD, CD\}$ [unordered pairs], with $\binom{4}{2} = 6$ *equally likely* outcomes (simple events). Selecting k out of n objects will similarly result in $\binom{n}{k}$ equally likely possibilities. Another typical experiment of this kind is dealing 5 cards out of 52.

4. The new feature of this example (waiting for the first head) is that the sample space is *infinite*: $\{H, TH, TTH, TTTH, TTTTH, TTTTTH, ...\}$. Eventually, we must learn to differentiate between the DISCRETE (countable) infinity, where the individual simple events can be labeled 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, ... in an exhaustive manner, and the CONTINUOUS infinity (real numbers in any interval). The current example is obviously a case of discrete infinity, which implies that the simple events *cannot* be equally likely (they would all have the probability of $\frac{1}{\infty} = 0$, implying that their sum is 0, an obvious contradiction). But we can easily manage to assign correct and meaningful probabilities even in this case (as discussed later).

5. The rotating wheel has also an infinite sample space (an outcome is identified with the final position − angle − of the pointer, measured from some fixed direction), this time being represented by all real numbers from the interval $[0, 2\pi)$ [assuming that angles are measured in radians]. This infinity of simple events is of the continuous type, with some interesting consequences. Firstly, from the symmetry of the experiment, all of its outcomes must be *equally likely*. But this implies that the probability of each single outcome is *zero*! Isn't this a contradiction as well? The answer is *no;* in this case the number of outcomes is no longer countable, and therefore the infinite sum (actually, an integral) of their *zero* probabilities *can* become nonzero (we need them to add up to 1). The final puzzle is: how do we put all these zero probabilities together to answer a simple question such as: what is the probability that the

pointer will stop in the $[0, \frac{\pi}{2}]$ interval? This will require introducing a new concept of the so called PROBABILITY DENSITY (*probability* of an interval, *divided* by the *length* of the interval). We will postpone this until the second part of this course.

6. What exactly is new about the tack and its two simple outcomes: $\{\perp, \not\perp\}$? Here, for the first time, we will not be able to introduce probabilities based on any symmetry argument, these will have to be established empirically by flipping the tack *many* times, finding the *proportion* of times it lands in the $\perp$ position and calling *this* the probability of $\perp$ (to be quite correct, the exact probability of $\perp$ is the *limit* of these experiments, when their number approaches infinity). That effectively implies that the probability of any such event can never be known *exactly;* we deal with this by replacing it by a PARAMETER p, which we substitute for the exact probability in all our formulas. Eventually we may learn (if we manage to reach those chapters) how to test HYPOTHESES concerning the value of p (such as, for example, $p = 0.7$).

Events

The technical definition of an event is: any SUBSET of the sample space. These are usually denoted by capital letters from the beginning of the alphabet: A, B, C,...

EXAMPLES (using the experiment of rolling two dice):

1. Let A be the event that the total number of dots equal 8.

   This of course consists of the subset: $\{(2,6), (3,5), (4,4), (5,3), (6,2)\}$ [five simple events].

2. B defined by requiring that neither of the two numbers be a six.

   This correspond to the subset:

1,1	1,2	1,3	1,4	1,5
2,1	2,2	2,3	2,4	2,5
3,1	3,1	3,3	3,4	3,5
4,1	4,2	4,3	4,4	4,5
5,1	5,2	5,3	5,4	5,5

3. C: first number smaller than second.

   Subset:

1,2	1,3	1,4	1,5	1,6
	2,3	2,4	2,5	2,6
		3,4	3,5	3,6
			4,5	4,6
				5,6

Set Theory (review)

Our definition of events as subsets of the sample space indicates that it may help to recall what we already know about sets, subsets, etc. Unfortunately, on occasion Statistics uses its own, different terminology for some of the set-theory definitions; it may help to build the corresponding '**dictionary**':

The old notion of:	is (are) now called:
Universal set Ω	Sample space
Elements of Ω (its individual 'points')	Simple events (complete outcomes)
Subsets of Ω	Events
Empty set $\emptyset$	Null event

We continue to use the word INTERSECTION (notation: $A \cap B$, representing the collection of simple events common to both A *and* B), UNION ($A \cup B$, simple events belonging to either A *or* B or *both*), and COMPLEMENT ($\overline{A}$, simple events *not* in A). One should be able to visualize these using Venn diagrams, but when dealing with more than 3 events at a time, one can tackle problems only with the help of

Boolean Algebra (another review)

Both $\cap$ and $\cup$ (individually) are COMMUTATIVE and ASSOCIATIVE, meaning $A \cap B = B \cap A$ and $(A \cap B) \cap C = A \cap (B \cap C)$, and the same when $\cap \to \cup$. Being associative implies that $A \cap B \cap C$ does not require any parentheses to be meaningful (same with $\cup$).

Intersection is **distributive** over union: $A \cap (B \cup C \cup ...) = (A \cap B) \cup (A \cap C) \cup ...$[try to prove it, using A, B, C only, through Venn diagrams].

Similarly, union is distributive over intersection: $A \cup (B \cap C \cap ...) = (A \cup B) \cap (A \cup C) \cap ...$ [try proof]. This is unlike the regular algebra of adding and multiplying numbers [addition is *not* distributive over multiplication: $a+(b \cdot c) \neq (a+b) \cdot (a+c)$], obviously the two algebras 'behave' differently.

Here is a handful of rather **trivial rules** which one can easily verify: $A \cap \Omega = A$, $A \cap \emptyset = \emptyset$, $A \cap A = A$, $A \cup \Omega = \Omega$, $A \cup \emptyset = A$, $A \cup A = A$, $A \cap \overline{A} = \emptyset$, $A \cup \overline{A} = \Omega$, $\overline{\overline{A}} = A$.

Also, when $A \subset B$ (A is a SUBSET of B, meaing that every element of A also belongs to B), we get: $A \cap B = A$ (the smaller event) and $A \cup B = B$ (the bigger event).

And two not so trivial laws (both called **DeMorgan's**): $\overline{A \cap B} = \overline{A} \cup \overline{B}$, and $\overline{A \cup B} = \overline{A} \cap \overline{B}$. These can be verified easily by Venn diagrams; both can be extended to any number of events:

$$\overline{A \cap B \cap C \cap ...} = \overline{A} \cup \overline{B} \cup \overline{C} \cup ...$$

and vice versa (i.e. $\cap \leftrightarrow \cup$).

And a simple definition: A and B are called (mutually) **exclusive** or DISJOINT when $A \cap B = \emptyset$ (i.e. there is no overlap between the two events, they have no simple events in common).

Probability of events

Having a sample space consisting of individual **simple events**, we would now like to assign each of these a sensible **probability** (relative frequency of its occurrence in a *long* run). It's obvious that each of these probabilities must be a non-negative number.

To find a probability of any **other event** A (not necessarily simple), we then add the probabilities of the simple events A consists of. This immediately implies that probabilities must follow a few basic rules:

$$\begin{aligned} \Pr(A) &\geq 0 \\ \Pr(\emptyset) &= 0 \\ \Pr(\Omega) &= 1 \end{aligned}$$

(the relative frequency of all Ω is obviously 1).

We should mention that $\Pr(A) = 0$ does not necessarily imply that $A = \emptyset$, some nonempty events may have a zero probability (we have already seen examples of these); they are 'officially' called IMPOSSIBLE events (a very misleading name, I will call them **zero-probability events**).

▶Other Formulas◀

$\Pr(A \cup B) = \Pr(A) + \Pr(B)$ but *only* when $A \cap B = \emptyset$ (*disjoint*). This implies that $\Pr(\overline{A}) = 1 - \Pr(A)$ as a special case.

This implies that $\Pr(A \cap \overline{B}) = \Pr(A) - \Pr(A \cap B)$ [obvious also from the corresponding Venn diagram].

For any A and B (possibly overlapping) we have

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

which can be verified from a Venn diagram (a probability of an event can be visualized as its *area*).

Using Boolean algebra we can **extend** this to: $\Pr(A \cup B \cup C) = \Pr(A) + \Pr(B \cup C) - \Pr\{A \cap (B \cup C)\} = \Pr(A) + \Pr(B) + \Pr(C) - \Pr(B \cap C) - \Pr\{(A \cap B) \cup (A \cap C)\} = \Pr(A) + \Pr(B) + \Pr(C) - \Pr(A \cap B) - \Pr(A \cap C) - \Pr(B \cap C) + \Pr(A \cap B \cap C)$.

And, by induction, we can get the **fully general**

$$\begin{aligned} \Pr(A_1 \cup A_2 \cup A_3 \cup ... \cup A_k) &= \sum_{i=1}^{k} \Pr(A_i) - \sum_{i<j}^{k} \Pr(A_i \cap A_j) + \sum_{i<j<\ell}^{k} \Pr(A_i \cap A_j \cap A_\ell) - ... \\ &\quad \pm \Pr(A_1 \cap A_2 \cap A_3 \cap ... \cap A_k) \end{aligned}$$

(the plus sign for k odd, the minus sign for k even). The formula computes the probability that *at least one* of the A_i events happens.

It is interesting to note that the probability of getting *exactly one* of the A_i events (i.e. either an element from $A_1 \cap \overline{A}_2 \cap \overline{A}_3 \cap ... \cap \overline{A}_k$, or $\overline{A}_1 \cap A_2 \cap \overline{A}_3 \cap ... \cap \overline{A}_k$, ... or $\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3 \cap ... \cap A_k$) is similarly computed by:

$$\begin{aligned} &\sum_{i=1}^{k} \Pr(A_i) - 2 \sum_{i<j}^{k} \Pr(A_i \cap A_j) + 3 \sum_{i<j<\ell}^{k} \Pr(A_i \cap A_j \cap A_\ell) - ... \\ &\pm k \Pr(A_1 \cap A_2 \cap A_3 \cap ... \cap A_k) \end{aligned}$$

We demonstrate the application of these formulas in the following, rather non-trivial EXAMPLE:

Suppose that k distinct letters (to different friends) have been written, each with a corresponding (uniquely addressed) envelope. Then, for some strange reason, the letters are placed in the envelopes purely randomly (after a thorough shuffling). The sample space of this experiment is thus a list of all permutations of k objects,

123
132
213
231
312
321

when $k = 3$ (we will assume that 123 represents the correct placement of all three letters). In general, there are $k!$ of these, all of them *equally likely* (due to symmetry, i.e. none of these arrangements should be more likely than any other). There are three simple-looking questions:

1. What is the probability of all letters being placed correctly?

   Solution (fairly trivial): Only one out of $k!$ random arrangements meets the criterion, thus the answer is $\frac{1}{k!}$ (astronomically small for k beyond 10).

2. What is the probability that *none* of the k letters are placed correctly?

   Solution is this time a lot more difficult. First we have to realize that it is relatively easy to figure out the probability of any given letter being placed *correctly*, and also the probability of any combination (*intersection)* of these, i.e. two specific letters correctly placed, three letters correct..., etc. [this kind of approach often works in other problems as well; intersections are usually easy to deal with, unions are hard but can be converted to intersections].

   Let us verify this claim. We use the following notation: A_1 means that the first letter is placed correctly (regardless of what happens to the rest of them), A_2 means the second letter is placed correctly, etc. $\Pr(A_1)$ is computed by counting the number of permutations which have 1 in the correct *first* position, and dividing this by $k!$. The number of permutations which have 1 fixed is obviously $(k-1)!$ [we are effectively permuting 2, 3, ... k, altogether $k-1$ objects]. $\Pr(A_1)$ is thus equal to $\frac{(k-1)!}{k!} = \frac{1}{k}$. The probability of A_2, A_3, etc. can be computed similarly, but it should be clear from the symmetry of the experiment that all these probabilities must be the same, and equal to $\Pr(A_1) = \frac{1}{k}$ (why should any letter have a better chance of being placed correctly than any other?). Similarly, let us compute $\Pr(A_1 \cap A_2)$, i.e. probability of the first *and* second letter being placed correctly (regardless of the rest). By again counting the corresponding number of permutations (with 1 and 2 fixed), we arrive at $\frac{(k-2)!}{k!} = \frac{1}{k(k-1)}$. This must be the same for any other pair of letters, e.g. $\Pr(A_3 \cap A_7) = \frac{1}{k(k-1)}$, etc. In this manner we also get $\Pr(A_1 \cap A_2 \cap A_3) = \Pr(A_3 \cap A_7 \cap A_{11}) = \frac{1}{k(k-1)(k-2)}$, etc.

So now we know how to deal with *any* intersection. All we need to do is to express the event 'all letters misplaced' using intersections *only*, and evaluate the answer, thus:

$$\Pr(\overline{A_1} \cap \overline{A_2} \cap ... \cap \overline{A_k}) \text{ [all letters misplaced]} =$$
$$\Pr(\overline{A_1 \cup A_2 \cup ... \cup A_k}) \text{ [DeMorgan]} =$$
$$1 - \Pr(A_1 \cup A_2 \cup ... \cup A_k) =$$
$$1 - \sum_{i=1}^{k} \Pr(A_i) + \sum_{i<j}^{k} \Pr(A_i \cap A_j) + ... \mp \Pr(A_1 \cap A_2 \cap ... \cap A_k) =$$
$$1 - k \cdot \tfrac{1}{k} + \tbinom{k}{2} \cdot \tfrac{1}{k(k-1)} - \tbinom{k}{3} \tfrac{1}{k(k-1)(k-2)} + ... \mp \tfrac{1}{k!} =$$
$$1 - 1 + \frac{1}{2!} - \frac{1}{3!} + ... \mp \frac{1}{k!}$$

For $k = 3$ this implies $1 - 1 + \frac{1}{2} - \frac{1}{6} = \frac{1}{3}$ (check, only 231 and 312 out of six permutations). For $k = 1$, 2, 4, 5, 6, and 7 we get: 0 (check, one letter cannot be misplaced), 50% for two letters (check), 37.5% (four letters), 36.67% (five), 36.81% (six), 36.79% (seven), after which the probabilities do not change (i.e., surprisingly, we get the same answer for 100 letters, a million letters, etc.).

Can we identify the **limit** of the $1 - 1 + \frac{1}{2!} - \frac{1}{3!} + ...$ sequence? Yes, of course, this is the expansion of $e^{-1} = .36788$.

3. Similarly, the probability of exactly one letter being placed correctly is $k \cdot \frac{1}{k} - 2\binom{k}{2}\frac{1}{k(k-1)} + 3\binom{k}{3}\frac{1}{k(k-1)(k-2)} + ... \mp k \cdot \frac{1}{k!} = 1 - 1 + \frac{1}{2!} + ... \mp \frac{1}{(k-1)!}$ (the previous answer short of its last term!). This equals to 1, 0, 50%, 37.5%, ... for $k = 1, 2, 3, 4, ...$ respectively, and has the same limit. ■

The **main point** of the whole section is the following

►Summary◄

Probability of any (Boolean) expression involving events A, B, C, ... can be *always* converted probabilities involving the individual events and their simple (non-complemented) *intersections* ($A \cap B$, $A \cap B \cap C$, etc.) *only*.

Proof: When the topmost operation of the expression (and, subsequently, any of the resulting subexpressions) is a union, we remove it by the $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$ rule (or its generalization). When the highest operation is a complement, we get rid of it by $\Pr(\overline{A}) = 1 - \Pr(A)$. After that, $\cap$ must be at the top of our evaluation tree. If, at the next level, there is at least one complement, we remove it (or them, one by one) by $\Pr(A \cap \overline{B}) = \Pr(A) - \Pr(A \cap B)$. Similarly we deal with a next-level union by applying $\Pr\{A \cap (B \cup C)\} = \Pr(A \cap B) + \Pr(A \cap C) - \Pr(A \cap B \cap C)$. In this manner we can remove all levels below $\cap$ until, ultimately, nothing but simple intersections remain. □

Let's go over a few easy EXAMPLES:

1. $\Pr\{(A\cap B)\cup\overline{B\cap C}\} = \Pr\{A\cap B\} + \Pr\{\overline{B\cap C}\} - \Pr\{A\cap B\cap\overline{B\cap C}\} = \Pr\{A\cap B\} + 1 - \Pr\{B\cap C\} - \Pr\{A\cap B\} + \Pr\{A\cap B\cap B\cap C\} = 1 - \Pr\{B\cap C\} + \Pr\{A\cap B\cap C\}$. This can be also deduced from the corresponding Venn diagram, bypassing the algebra (a legitimate way of doing things).

2. $\Pr\{(A\cap B)\cup\overline{C\cup D}\} = \Pr\{A\cap B\} + \Pr\{\overline{C\cup D}\} - \Pr\{(A\cap B)\cap\overline{C\cup D}\} = \Pr\{A\cap B\} + 1 - \Pr\{C\cup D\} - \Pr\{A\cap B\} + \Pr\{(A\cap B)\cap(C\cup D)\} = 1 - \Pr\{C\cup D\} + \Pr\{(A\cap B\cap C)\cup(A\cap B\cup D)\} = 1 - \Pr\{C\} - \Pr\{D\} + \Pr\{C\cap D\} + \Pr\{A\cap B\cap C\} + \Pr\{A\cap B\cap D\} - \Pr\{A\cap B\cap C\cap D\}$.

3. Four players are dealt 5 cards each. What is the probability that at least one player gets exactly 2 aces (a chapter ago, we could not solve this problem).

   Solution: Let A_1 be the event that the first player gets exactly 2 aces, A_2 means that the second player has exactly 2 aces, etc. The question amounts to finding $\Pr(A_1\cup A_2\cup A_3\cup A_4)$. By our formula, this equals $\sum_{i=1}^{4}\Pr(A_i) - \sum_{i<j}^{4}\Pr(A_i\cap A_j) + 0$ [the intersection of 3 or more of these events is empty − there are only 4 aces]. For $\Pr(A_1)$ we get $\frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} = 3.993\%$ [the denominator counts the total number of five-card hands, the numerator counts only those with exactly two aces] with the same answer for $\Pr(A_2), ...\ \Pr(A_4)$ [the four players must have equal chances]. Similarly $\Pr(A_1\cap A_2) = \frac{\binom{4}{2,2,0}\binom{48}{3,3,42}}{\binom{52}{5,5,42}} = 0.037\%$ [the denominator represents the number of ways of dealing 5 cards each to two players, the numerator counts only those with 2 aces each − recall the 'partitioning' formula], and the same probability for any other pair of players.

   Final answer: $4\Pr(A_1) - 6\Pr(A_1\cap A_2) = 15.75\%$.

4. There are 100,000 lottery tickets marked 00000 to 99999. One of these is selected at random. What is the probability that the number on it contains 84 [consecutive, in that order] at least once.

   Solution: Let's introduce four events: A means that the first two digits of the ticket are 84 (regardless of what follows), B: 84 is found in the second and third position, C: 84 in position three and four, and D: 84 in the last two positions. Obviously we need $\Pr(A\cup B\cup C\cup D) = \Pr(A) + \Pr(B) + \Pr(C) + \Pr(D) - \Pr(A\cap C) - \Pr(A\cap D) - \Pr(B\cap D) + 0$ [the remaining possibilities are all null events - the corresponding conditions are incompatible, see the Venn diagram].

   The answer is $4\times\frac{1000}{100{,}000} - 3\times\frac{10}{100{,}000} = 0.04 - 0.0003 = 3.97\%$ [the logic of each fraction should be obvious − there are 1000 tickets which belong to A, 10 tickets which meet conditions A and C, etc.]. ■

Probability tree & Conditional probability

Consider a random experiment which is done in several STAGES such as, for example, selecting 3 marbles (one by one, without replacement − these are the three 'stages' of this experiment), from a box containing (originally) 3 red and 5 blue marbles. The easiest way to display possible outcomes of this experiment is to draw a so called **probability tree**, with the individual branches representing possible outcomes at each stage of the experiment. This will be done in class; it is effectively a graphical representation of

$$rrr$$
$$rrb$$
$$rbr$$
$$rbb$$
$$brr$$
$$brb$$
$$bbr$$
$$bbb$$

(the sample space, each line being one simple event). In the graph, it is one complete path (from beginning till end) which represents a simple event (each can be also identified by its end point).

It is easy to assign **probabilities** to individual BRANCHES of this tree; the initial selection of a red marble r has the probability of $\frac{3}{8}$, once this is done the probability of the next marble being blue (the $r \to b$ branch) is $\frac{5}{7}$ [5 blue marbles out of 7, one red is out], after that selecting r again (the $rb \to r$ branch) has the probability of $\frac{2}{6}$ [2 red marbles left out of 6]. Note that the probabilities at each 'fork' have to add up to one.

We introduce the following notation: R_1 means a red marble is selected first (in terms of our sample space, this event consists of: $\{rrr, rrb, rbr, rbb\}$), R_2 means a red marble is selected in the second draw (regardless of the outcome of Draw 1 and 3): $\{rrr, rrb, brr, brb\}$, and similarly we define R_3, B_1 (a blue marble first), B_2, and B_3.

▶Two Issues to Settle◀

- What are the simple probabilities of individual branches found above (by counting how many marbles of each color are left at that stage). The first one ($\frac{3}{8}$, of the initial choice of r) is obviously $Pr(R_1)$. The second one ($\frac{5}{7}$, of the $r \to b$ branch) can *not* be simply $\Pr(B_2)$, since there are other ways of getting a blue marble in the second draw. To give it its proper name, we have to introduce the so called **conditional probability** of an event B, given that another event A has already happened, notation: $\Pr(B|A)$. This is a very natural notion in a multi-stage experiment when the outcome of B is decided based on the outcome of the previous stage(s). It is thus obvious that $\frac{5}{7}$ represents, by this definition, $\Pr(B_2|R_1)$. Similarly $\frac{2}{6}$ is $\Pr(R_3|R_1 \cap B_2)$ [third marble being red *given* that the first one was red *and* the second one blue].

- How do we compute probabilities of simple events (and thus events in general) of this sample space (we recall that a simple event is a complete 'path' (e.g. rbr). Clearly: if this experiment is repeated (infinitely) many times, $\frac{3}{8}$ of them will result in r as the first marble, *out of these* $\frac{5}{7}$ will have b as the second marble, and *out of these* $\frac{2}{6}$ will finish with r as the third marble. $\frac{5}{7}$ out of $\frac{3}{8}$ is $\frac{15}{56}$ $(= \frac{3}{8} \cdot \frac{5}{7})$ and $\frac{2}{6}$ out of $\frac{15}{56}$ is $\frac{5}{56}$ $(= \frac{3}{8} \cdot \frac{5}{7} \cdot \frac{2}{6})$. We can formalize this by

$$\Pr(rbr) = \Pr(R_1 \cap B_2 \cap R_3) = \Pr(R_1) \cdot \Pr(B_2|R_1) \cdot \Pr(R_3|R_1 \cap B_2)$$

  In a similar manner we can assign probabilities to all 'points' (simple events) of the sample space, thus:

Path:	rrr	rrb	rbr	rbb	brr	brb	bbr	bbb
Probability:	$\frac{1}{56}$	$\frac{5}{56}$	$\frac{5}{56}$	$\frac{10}{56}$	$\frac{5}{56}$	$\frac{10}{56}$	$\frac{10}{56}$	$\frac{10}{56}$

  Note that it is convenient to keep the same (common) denominator of these probabilities. This helps us realize which simple events are more likely that others, and by what factor; it also simplifies subsequent computations.

  Conclusion: The '*natural*' probabilities of a multistage experiment are the conditional probabilities of individual branches. All other probabilities must be build from these, using the product rule to first find probabilities of simple events.

The rule which was used to compute the probability of the $R_1 \cap B_2 \cap R_3$ intersection is called the

►Product Rule◄

and it can be generalized to *any* two, three, etc. events thus:

$$\begin{aligned}
\Pr(A \cap B) &= \Pr(A) \cdot \Pr(B|A) \\
\Pr(A \cap B \cap C) &= \Pr(A) \cdot \Pr(B|A) \cdot \Pr(C|A \cap B) \\
\Pr(A \cap B \cap C \cap D) &= \Pr(A) \cdot \Pr(B|A) \cdot \Pr(C|A \cap B) \cdot \Pr(D|A \cap B \cap C) \\
&\vdots
\end{aligned}$$

EXAMPLE: 4 players are dealt 13 cards each from an ordinary deck (of 52 cards). What is the probability that each player will get exactly one ace?

Old solution: The sample space consists of all possible ways of partitioning 52 cards into 4 groups of equal size, $\binom{52}{13,13,13,13}$ in number. To give each player exactly one ace, we have to similarly partition the aces and non-aces, and multiply the answers. There are $\binom{4}{1,1,1,1} \cdot \binom{48}{12,12,12,12}$ ways of doing this; divided by the above *total* number of ways (all equally likely) gives the answer: 10.55%.

New solution: If A, B, C, and D represent Mr.A, Mr.B, Mr.C and Mr.D getting exactly one ace (respectively), we employ the product rule: $\Pr(A \cap$

$B \cap C \cap D) = \Pr(A)\Pr(B|A)\Pr(C|A \cap B)\Pr(D|A \cap B \cap C) = \frac{\binom{4}{1}\binom{48}{12}}{\binom{52}{13}} \cdot \frac{\binom{3}{1}\binom{36}{12}}{\binom{39}{13}} \cdot \frac{\binom{2}{1}\binom{24}{12}}{\binom{26}{13}} \cdot \frac{\binom{1}{1}\binom{12}{12}}{\binom{13}{13}}$ [visualize the experiment done sequentially, Mr.A is dealt the first 13 cards, Mr.B the next 13, etc.]. This of course gives the same answer (note that the last of the four factors is equal to 1 − once you have dealt one ace each to Mr.A, B, and C, Mr.D *must* get his one ace with the conditional probability of 1). ■

In general, we can **define** the

►Conditional Probability◄

for *any* two events (*regardless* whether they came from a multistage experiment or not) by

$$\Pr(B|A) \equiv \frac{\Pr(A \cap B)}{\Pr(A)}$$

(the product rule in reverse).

The **meaning** of any such conditional probability is as follows: given (someone who has observed the *complete* outcome of the experiment is telling us) that A has happened (but we are left in the dark about the rest of the experiment), then $\Pr(B|A)$ represents the conditional probability of B happening as well. This is as if the whole sample space has shrunk to A only, and the probability of any other event had to be re-computed accordingly. The general definition thus applies to *any* A and B of *any* random experiment (not necessarily multistage).

Not surprisingly, *all* **formulas** which hold true in the original sample space are *still valid* in the new 'reduced' sample space A, i.e. *conditionally*, e.g.: $\Pr(\overline{B}|A) = 1 - \Pr(B|A)$, $\Pr(B \cup C|A) = \Pr(B|A) + \Pr(C|A) - \Pr(B \cap C|A)$, etc. (make all probabilities of any old formula conditional on A). Both of these will be proved in class via Venn diagrams (you can try it on your own), but the idea works for any other formula as well.

Note that with this general definition it is possible to compute the conditional probability of, say R_1 *given* B_3 (i.e. *guessing* the outcome of the *first* stage *based* on the outcome of the *last* stage − a total reversal of what we have been doing so far). Subsequently, this is utilized in the so called Bayes' theorem.

And a simple EXAMPLE of establishing a value of conditional probability in general:

Suppose the experiment consists of rolling two dice (red and green), A is: 'the total number of dots equals 6', B is: 'the red die shows an even number'. Compute $\Pr(B|A)$.

Solution: Let us use the old sample space, indicating simple events of A by ◯, of B by ×, and of the $A \cap B$ overlap by ⊗ :

				◯	
×	×	×	⊗	×	×
		◯			
×	⊗	×	×	×	×
◯					
×	×	×	×	×	×

$\Pr(B|A)$ is clearly the number of the overlap simple events $\otimes$ divided by the number of simple events in A (either $\bigcirc$ or $\otimes$), as these are all equally likely.

Answer: $\Pr(B|A) = \frac{2}{5}$ [or $\frac{\frac{2}{36}}{\frac{5}{36}}$, if you want to insist on using the $\frac{\Pr(A\cap B)}{\Pr(A)}$ formula]. ■

Note that in general $\Pr(B|A) \neq \Pr(A|B)$, as these two conditional probabilities correspond to totally different situations, and have no reason to be even compared. [In the former example $\Pr(A|B) = \frac{2}{18} = \frac{1}{9}$, to demonstrate the point].

Partition of a sample space

(nothing to do with our previous *partitioning* of a group of people into several teams). This new notion of a PARTITION represents chopping the sample space into several smaller events, say A_1, A_2, A_3, ..., A_k, so that they

(i) don't overlap (i.e. are all mutually exclusive): $A_i \cap A_j = \emptyset$ for any $1 \leq i, j \leq k$

(ii) cover the whole Ω (i.e. 'no gaps'): $A_1 \cup A_2 \cup A_3 \cup ... \cup A_k = \Omega$. ■

It should be obvious that the 'finest' partition is the *collection* of all simple events, and the 'crudest' partition is Ω itself. The most interesting partitions will of course be the in-between cases. One such example is A and $\overline{A}$ (where A is an arbitrary event).

Partitions can be quite useful when computing a probability of yet another event B. This task can be often simplified by introducing a *convenient* partition, and utilizing the following

►Formula of Total Probability◄

$$\Pr(B) = \Pr(B|A_1) \cdot \Pr(A_1) + \Pr(B|A_2) \cdot \Pr(A_2) + ... + \Pr(B|A_k) \cdot \Pr(A_k)$$

which can be readily verified by a Venn diagram when we realize that $\Pr(B|A_1) \cdot \Pr(A_1) = \Pr(B \cap A_1)$, $\Pr(B|A_2) \cdot \Pr(A_2) = \Pr(B \cap A_2)$, etc. The difficulty of applying the formula to a specific situation relates to the fact that the question will normally specify B only; the actual partition must be introduced by us, intelligently, as a part of the solution.

EXAMPLE: Two players are dealt 5 cards each. What is the probability that they will have the same number of aces?

Solution: We partition the sample space according to how many aces the first player gets, calling the events A_0, A_1, ..., A_4. Let B be the event of our question (both players having the same number of aces). Then, by the formula of total probability: $\Pr(B) = \Pr(A_0)\Pr(B|A_0) + \Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2) + \Pr(A_3)\Pr(B|A_3) + \Pr(A_4)\Pr(B|A_4) = \frac{\binom{4}{0}\binom{48}{5}}{\binom{52}{5}} \cdot \frac{\binom{4}{0}\binom{43}{5}}{\binom{47}{5}} + \frac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}} \cdot \frac{\binom{3}{1}\binom{44}{4}}{\binom{47}{5}} + \frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} \cdot \frac{\binom{2}{2}\binom{45}{3}}{\binom{47}{5}} + \frac{\binom{4}{3}\binom{48}{2}}{\binom{52}{5}} \cdot 0 + \frac{\binom{4}{4}\binom{48}{0}}{\binom{52}{5}} \cdot 0 = 49.33\%$ ■

▶Bayes Rule◀

This time we start with an EXAMPLE:

Consider four 'black' boxes: two of them (call them Type I) have 1 green and 2 red marbles inside, one (Type II) has 1 green and 1 red marble, and one (Type III) has 2 green and 1 red marble. Let one of these boxes be selected at random, and a marble drawn from it. The probability tree of this experiment looks like this (the fraction in parentheses is the conditional probability of the corresponding branch – this will be done properly in class):

1^{st} branch	2^{nd} branch	Pr		
$(\frac{2}{4})$ I $\rightarrow$	$(\frac{2}{3})$ r	$\frac{8}{24}$	✓	◯
$\searrow$	$(\frac{1}{3})$ g	$\frac{4}{24}$		
$(\frac{1}{4})$ II $\rightarrow$	$(\frac{1}{2})$ r	$\frac{3}{24}$	✓	
$\searrow$	$(\frac{1}{2})$ g	$\frac{3}{24}$		
$(\frac{1}{4})$ III $\rightarrow$	$(\frac{1}{3})$ r	$\frac{2}{24}$	✓	
$\searrow$	$(\frac{2}{3})$ g	$\frac{4}{24}$		

Let I, II, and III represent the events of selecting Type I, II, or III box; then (I, II, III) is an obvious partition of the sample space. Similarly, if R and G represent selecting a red and green marble, respectively (regardless of the box), then (R, G) is yet another partition of our sample space.

1. Compute $\Pr(R)$:

   Using the total-probability formula: $\Pr(R) =$
   $\Pr(R|\text{I})\cdot\Pr(\text{I})+\Pr(R|\text{II})\cdot\Pr(\text{II})+\Pr(R|\text{III})\cdot\Pr(\text{III}) = \frac{8}{24}+\frac{3}{24}+\frac{2}{24} = 54.17\%$.
   This is the same as checking off (✓) the simple events which contribute to R, and adding their probabilities. The formula just spells out the logic of this simple and natural procedure.

   Similarly, we can compute $\Pr(G) = \frac{11}{24}$.

2. Important: Find $\Pr(\text{I}\,|R)$:

   This at first appears as a rather unusual question (to find the conditional probability of an outcome of the *first* stage of our experiment, *given* the result of the *second* stage – note the chronological reversal!). Yet, in the next example we demonstrate that this is quite often what is needed.

   Solution: We use the formal definition of conditional probability: $\Pr(\text{I}|R) = \frac{\Pr(\text{I}\cap R)}{\Pr(R)} = \frac{\frac{8}{24}}{\frac{13}{24}} = \frac{8}{13} = 61.54\%$. This is the probability of having selected Type I box (we cannot tell - they all look identical) given that a red marble was drawn. Note that this conditional probability is higher than the original $\Pr(\text{I}) = 50\%$ (do you understand why?). And, yet another marble drawn from the same box may help us even more (especially if it's also red!). ■

The **procedure** for computing $\Pr(\text{I}\,|R)$ can be **general**ized as follows: check off (✓) all simple events contributing to R, out of these check off (perhaps using a different symbol, ◯ in our case) those which also contribute to I (i.e. of the $\text{I}\cap R$

overlap) Then divide the total probability of the later by the total probability of the former.

This constitutes what I call the **Bayes rule** (your textbook presents it as a formula, which we can bypass). We always encounter it in the context of a multistage experiment to be dealt with by drawing the corresponding probability tree.

Let us go over one more EXAMPLE of its application:

> Let 0.5% of a population in a certain area have tuberculosis. There is a medical test which can detect this condition in 95% of all (infected) cases, but at the same time the test is (falsely) positive for 10% of the healthy people [all in all, the test is at least 90% accurate].
>
> The question is: A person is selected randomly and tested. The test is positive (indicating a presence of TB). What is the probability that the person actually has it [our guess probably is: at least 90%, but we are in for a big surprise].
>
> Solution: This is a very simple two-stage experiment; in the first stage (considering how the experiment is *actually* performed) the person is selected, resulting in either 'sick' s or 'healthy' h individual[the actual outcome is hidden from us as sick and healthy people look the same to us] with the probability of 0.005 and 0.995, respectively; in the second stage the person is tested, resulting in either positive p (TB!) or negative n result (the corresponding probabilities are of the conditional type, depending on the first-stage outcome), thus:
>
> $$\begin{array}{rll} (0.005)\ s \rightarrow (0.95)\ p & & 0.00475\ \checkmark\ \bigcirc \\ \qquad\qquad \searrow (0.05)\ n & & 0.00025 \\ (0.995)\ h \rightarrow (0.10)\ p & & 0.09950\ \checkmark \\ \qquad\qquad \searrow (0.90)\ n & & 0.89550 \end{array}$$
>
> If S denotes the 'sick person' event and P stands for the 'positive test' event, we obviously need to compute $\Pr(S|P)$.
>
> Using the Bayes rule, this is equal to $\frac{0.00475}{0.00475+0.09950} = 4.556\%$ [still fairly small, even though almost 10 times bigger than before the test]. Note that $\checkmark$ marks simple events of P, and $\bigcirc$ those of $S \cap P$. ■

Independence

►Of Two Events◄

is a rather **natural** notion: if the experiment is done in such a manner that A (happening or not) cannot influence the probability of B, B is *independent* of A. Formally, this means that $\Pr(B|A) = \Pr(B)$ [knowing that A has happened does not change the probability of B]. Mathematically, this is equivalent to: $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$, and also to $\Pr(A|B) = \Pr(A)$. Thus, B being independent of A implies that A is independent of B, which means that independence of two events is always *mutual* ($A \bowtie B$ will be our informal notation for independence of A and

B). The same condition is also equivalent to $\Pr(A \cap \overline{B}) = \Pr(A) \cdot \Pr(\overline{B})$ [prove!], etc. Thus $A \bowtie B \Leftrightarrow A \bowtie \overline{B} \Leftrightarrow \overline{A} \bowtie B \Leftrightarrow \overline{A} \bowtie \overline{B}$.

We should mention that the condition of independence may sometimes be met 'accidently' by two events which *do* seem to *influence* each other. Technically, they will also be considered independent, but such **artificial** independence is of not much use to us. We will concentrate on independence which we can clearly deduce from the nature of the experiment, such as: an outcome of one die cannot influence the outcome of another die; but also: an outcome of a die cannot influence its future outcome(s) − a die has no memory. Avoid the common mistake of confusing independence with being mutually exclusive − two events which are independent must have a non-zero overlap (of a specific size); on the other hand exclusive events are strongly dependent, since $\Pr(A|B) = 0$ [and not $\Pr(A)$].

The notion of independence can be extended to 3 or more events.

The natural, **mutual independence** of

►Three Events◄

requires them to be independent *pairwise*, *plus*: $\Pr(A \cap B \cap C) = \Pr(A) \cdot \Pr(B) \cdot \Pr(C)$.

And again, this is the *same as* A, B and $\overline{C}$ being mutually independent, etc. (eight distinct ways of putting it).

Mutual independence also *implies* that any event build from A and B (e.g. $A \cup \overline{B}$) must be independent of C.

In **general** ► k Events◄

are mutually independent if the probability of any intersection of these (or their complements) is equal to the corresponding product of individual probabilities [$2^k - 1 - k$ conditions when not considering complements!].

The main point of **natural independence** is that all of these conditions are there, *automatically*, for us to utilize, just for realizing that the events have no way of influencing each other's outcome.

Mutual independence of A, B, C, D, ... also **implies** that any event build of A, B, ... must be independent of any event build out of C, D, ... [as long as the two sets are *distinct*].

Proof: We have already seen that any event can be replaced by its complement without effecting independence. The mutual independence of A, B and C implies that $A \cap B$ and C are independent [$\Pr\{(A \cap B) \cap C\} = \Pr(A)\Pr(B)\Pr(C) = \Pr(A \cap B)\Pr(C)$], and also that $A \cup B$ and C are independent [$\Pr\{(A \cup B) \cap C\} = \Pr(A \cap C) + \Pr(B \cap C) - \Pr(A \cap B \cap C) = \Pr(A)\Pr(C) + \Pr(B)\Pr(C) - \Pr(A)\Pr(B)\Pr(C) = (\Pr(A) + \Pr(B) - \Pr(A)\Pr(B))\Pr(C) = \Pr(A \cup B)\Pr(C)$]. The rest follows by induction. □

The final and most important

▶Implication of Independence◀

To compute the probability of a Boolean expression (itself an event) involving only mutually independent events, it is sufficient to know the events' *individual* probabilities. This is clear from the fact that the probability of any composite event can be expressed in terms of probabilities of the individual-event intersections, and these in turn can now be converted to products of individual probabilities (the actual computation may be further simplified by various 'shortcuts').

EXAMPLE: Let $\Pr(A) = 0.1$, $\Pr(B) = 0.2$, $\Pr(C) = 0.3$ and $\Pr(D) = 0.4$. Compute $\Pr[(A \cup B) \cap \overline{C \cup D}]$.

Solution: $= \Pr(A \cup B) \cdot [1 - \Pr(C \cup D)] = [0.1 + 0.2 - 0.02] \cdot [1 - 0.3 - 0.4 + 0.12] = 11.76\%$ ■

End-of-chapter examples

- Express $\Pr\{(A \cup \overline{C \cap D}) \cap \overline{B \cup \overline{D}}\}$ in terms of the individual probabilities $\Pr(A)$, $\Pr(B)$, ... assuming that the four events are independent.

  Solution: $= \Pr\{(A \cup \overline{C} \cup \overline{D}) \cap \overline{B} \cap D\} = \Pr\{(A \cup \overline{C} \cup \overline{D}) \cap D\} \cdot [1 - \Pr(B)] = \Pr\{(A \cap D) \cup (\overline{C} \cap D) \cup \emptyset\} \cdot [1 - \Pr(B)] = \{\Pr(A) \Pr(D) + [1 - \Pr(C)] \Pr(D) - \Pr(A)[1 - \Pr(C)] \Pr(D)\} \cdot [1 - \Pr(B)]$.

- Let us return to Example 2 of the previous chapter (lottery with 100,000 tickets) and compute the probability that a randomly selected ticket has an 8 and a 4 on it (each at least once, in any order, and not necessarily consecutive).

  Solution: Define A: no 8 at any place, B: no 4. We need $\Pr(\overline{A} \cap \overline{B})$ [at least one 8 *and* at least one 4] $= \Pr(\overline{A \cup B})$ [DeMorgan] $= 1 - \Pr(A \cup B) = 1 - \Pr(A) - \Pr(B) + \Pr(A \cap B)$. Clearly $A \equiv A_1 \cap A_2 \cap ... \cap A_5$, where A_1: 'no 8 in the first place', A_2: 'no 8 in the second place', etc. A_1, A_2, ..., A_5 are mutually independent (selecting a random 5 digit number is like rolling an 10-sided die five times), thus $\Pr(A) = \Pr(A_1) \cdot \Pr(A_2) \cdot ... \cdot \Pr(A_5) = (\frac{9}{10})^5$. Similarly, $\Pr(B) = (\frac{9}{10})^5$. Now, $A \cap B \equiv C_1 \cap C_2 \cap ... \cap C_5$ where C_1: not 8 nor 4 in the first spot, C_2: not 8 nor 4 in the second, etc.; these of course are also independent, which implies $\Pr(A \cap B) = (\frac{8}{10})^5$.

  Answer: $1 - 2(\frac{9}{10})^5 + (\frac{8}{10})^5 = 14.67\%$.

- The same question, but this time we want at least one 8 followed (sooner or later) by a 4 (at least once). What makes this different from the original question is that 8 and 4 now don't have to be consecutive.

  Solution: We *partition* the sample space according to the *position* at which 8 appears *for the first time*: B_1, B_2, ..., B_5, plus B_0 (which means there is no 8). Verify that this *is* a partition. Now, if A is the event of our question (8 followed by a 4), we can apply the formula of total probability thus: $\Pr(A) = \Pr(A|B_1) \cdot \Pr(B_1) + \Pr(A|B_2) \cdot \Pr(B_2) + \Pr(A|B_3) \cdot \Pr(B_3) + \Pr(A|B_4) \cdot \Pr(B_4) + \Pr(A|B_5) \cdot \Pr(B_5) + \Pr(A|B_0) \cdot \Pr(B_0)$. Individually, we deal with these in the following manner (we use the third term as an example): $\Pr(B_3) =$

$(\frac{9}{10})^2(\frac{1}{10})$ [no 8 in the first slot, no 8 in the second, 8 in third, and anything after that; then multiply due to independence], $\Pr(A|B_3) = 1-(\frac{9}{10})^2$ [given the first 8 is in the third slot, get at least one 4 after; easier through complement: $1 - \Pr(\text{no 4 in the last two slots})$].

Answer: $\Pr(A) = [1-(\frac{9}{10})^4] \cdot \frac{1}{10} + [1-(\frac{9}{10})^3] \cdot \frac{9}{10}\frac{1}{10} + [1-(\frac{9}{10})^2] \cdot (\frac{9}{10})^2\frac{1}{10} + [1-\frac{9}{10}] \cdot (\frac{9}{10})^3\frac{1}{10} + 0 \cdot (\frac{9}{10})^4\frac{1}{10} = 8.146\%$

- Out of 10 dice, 9 of which are regular but one is 'crooked' (6 has a probability of 0.5), a die is selected at random (we cannot tell which one, they all *look* identical). Then, we roll it twice. The sample space of the complete experiment, including probabilities, is

$r66$	$0.9 \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{9}{360}$	✓	◯
$r6\bar{6}$	$0.9 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{45}{360}$	✓	
$r\bar{6}6$	$0.9 \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{45}{360}$		◯
$r\bar{6}\bar{6}$	$0.9 \cdot \frac{5}{6} \cdot \frac{5}{6} = \frac{225}{360}$		
$c66$	$0.1 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{9}{360}$	✓	◯
$c6\bar{6}$	$0.1 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{9}{360}$	✓	
$c\bar{6}6$	$0.1 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{9}{360}$		◯
$c\bar{6}\bar{6}$	$0.1 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{9}{360}$		

  We will answer three question:

  1. Given that the first roll resulted in a six (Event S_1), what is the (conditional) probability of getting a six again in the second roll (Event S_2)?

     Solution: In our sample space we mark off the simple events contributing to S_1 (by ✓) and to S_2 (by ◯) and compute $\dfrac{\Pr(S_1 \cap S_2)}{\Pr(S_1)}$ (by adding the corresponding probabilities).

     Answer: $\frac{9+9}{9+9+45+9}$ (the common denominator of 360 cancels out) $= 25\%$.
  2. Are S_1 and S_2 independent?

     Let us check it out, carefully! $\Pr(S_1 \cap S_2) \stackrel{?}{=} \Pr(S_1) \cdot \Pr(S_2)$.

     Solution: $\frac{18}{360} (= \frac{1}{20}) \neq \frac{72}{360} \cdot \frac{72}{360} (= \frac{1}{25})$.

     Answer: No.
  3. Given that both rolls resulted in a six, what is the (conditional) probability of having selected the crooked die?

     Answer: $\frac{9}{9+9} = 50\%$.

- Ten people have been arrested as suspects in a crime one of them must have committed. A lie detector will (incorrectly) incriminate an innocent person with a 5% probability, it can (correctly) detect a guilty person with a 90% probability.

  1. One person has been tested so far and the lie detector has its red light flashing (implying: 'that's him'). What is the probability that he is the criminal?

Solution: Using c for 'criminal', i for 'innocent' r for 'red light flashing' and g for 'green', we have the following sample space:

$$\begin{array}{ll} cr & \frac{1}{10}\cdot\frac{9}{10}=0.090 \quad \checkmark \ \bigcirc \\ cg & \frac{1}{10}\cdot\frac{1}{10}=0.010 \\ ir & \frac{9}{10}\cdot\frac{1}{20}=0.045 \quad \checkmark \\ ig & \frac{9}{10}\cdot\frac{19}{20}=0.855 \end{array}$$

Answer: $\Pr(C|R)=\frac{0.090}{0.090+0.045}=\frac{2}{3}$ (far from certain!).

2. All 10 people have been tested and exactly one incriminated. What is the probability of having the criminal now?

   A simple event consists now of a complete record of these tests (the sample space has of 2^{10} of these), e.g. $rggrggrggg$. Assuming that the first item represents the criminal (the sample space must 'know' who the criminal is), we can assign probabilities by simply multiplying since the tests are done *independently* of each other. Thus, the simple event above will have the probability of $0.9\times 0.95^2\times 0.05\times 0.95^2\times 0.05\times 0.95^3$, etc. Since only one test resulted in r, the only simple events of relevance (the idea of a 'reduced' sample space) are:

$$\begin{array}{ll} rggggggggg & 0.9\times 0.95^9 \\ grgggggggg & 0.1\times 0.95^8\times 0.05 \\ \ldots\ldots\ldots & \\ gggggggggr & 0.1\times 0.95^8\times 0.05 \end{array}$$

   Given that it was one of these outcomes, what is the probability it was actually the first one?

   Answer: $\dfrac{0.9\times 0.95^9}{0.9\times 0.95^9+9\times 0.1\times 0.95^8\times 0.05}=95\%$ (now we are a lot more certain − still not 100% though!).

- Two men take one shot each at a target. Mr. A can hit it with the probability of $\frac{1}{4}$, Mr. B's chances are $\frac{2}{5}$ (he is a better shot). What is the probability that the target is hit (at least once)?

  Here, we have to (*on our own*) assume independence of the two shots.

  Solution (using an obvious notation): $\Pr(A\cup B)=\Pr(A)+\Pr(B)-\Pr(A\cap B)=\frac{1}{4}+\frac{2}{5}-\frac{1}{10}=55\%$.

  Alternately: $\Pr(A\cup B)=1-\Pr(\overline{A\cup B})=1-\Pr(\overline{A}\cap\overline{B})=1-\frac{3}{4}\cdot\frac{3}{5}=55\%$ [replacing Pr(at least one hit) by 1 − Pr(all misses)].

- What is more likely, getting at least one 6 in four rolls of a die, or getting at least one *double* 6 in twenty four rolls of a *pair* of dice?

  Solution: Let's work it out. The first probability can be computed as 1−Pr(no sixes in 4 rolls) $=1-(\frac{5}{6})^4$ [due to independence of the individual rolls] $=51.77\%$. The second probability, similarly, as $1-\Pr$(no double six in 24 rolls of a pair) $=1-(\frac{35}{36})^{24}=49.14\%$ [only one outcome out of 36 results in a double six].

  Answer: Getting at least one 6 in four rolls is more likely.

- Four people are dealt 13 cards each. You (one of the players) got one ace. What is the probability that your partner has the other three aces? (Go back three questions to get a hint).

  We can visualize the experiment done sequentially, with you being the first player and your partner the second one [even if the cards were actually dealt in a different order, that cannot change probabilities, right?]. The answer is a *natural* conditional probability, i.e. the actual condition (event) is decided in the first stage [consider it completed accordingly]. The second stage then consists of dealing 13 cards out of *39*, with *3* aces remaining.

  Answer: $\frac{\binom{3}{3}\binom{36}{10}}{\binom{39}{13}} = 3.129\%$.

  The moral: conditional probability is, in some cases, the 'simple' probability.

- A, B, C are mutually independent, having (the individual) probabilities of 0.25, 0.35 and 0.45, respectively. Compute $\Pr[(A \cap \overline{B}) \cup C]$.

  Solution: $= \Pr(A \cap \overline{B}) + \Pr(C) - \Pr(A \cap \overline{B} \cap C) = 0.25 \times 0.65 + 0.45 - 0.25 \times 0.65 \times 0.45 = 53.94\%$.

- Two coins are flipped, followed by rolling a die as many times as the number of heads shown. What is the probability of getting fewer than 5 dots in total?

  Solution: Introduce a partition A_0, A_1, A_2 according to how many heads are obtained. If B stands for 'getting fewer than 5 dots', the total-probability formula gives: $\Pr(B) = \Pr(A_0)\Pr(B|A_0) + \Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2) = \frac{1}{4} \times 1 + \frac{2}{4} \times \frac{4}{6} + \frac{1}{4} \times \frac{6}{36} = 62.5\%$.

  The probabilities of A_0, A_1, and A_2 followed from the sample space of two flips: $\{hh, ht, th, tt\}$; the conditional probabilities are clear for $\Pr(B|A_0)$ and $\Pr(B|A_1)$, $\Pr(B|A_2)$ requires going back to 36 outcomes of two rolls of a die and counting those having a total less than 5: $\{11, 12, 13, 21, 22, 31\}$.

- Consider the previous example. Given that there were exactly 3 dots in total, what is the conditional probability that the coins showed exactly one head?

  Solution: We are given the outcome of the second stage to guess at the outcome of the first stage. We need the Bayes rule, and the following (*simplified*) sample space:

$\mathbf{03}$	$\frac{1}{4} \cdot 0$	✓	
$\mathbf{0\overline{3}}$	$\frac{1}{4} \cdot 1$		
$\mathbf{13}$	$\frac{1}{2} \cdot \frac{1}{6}$	✓	◯
$\mathbf{1\overline{3}}$	$\frac{1}{2} \cdot \frac{5}{6}$		
$\mathbf{23}$	$\frac{1}{4} \cdot \frac{2}{36}$	✓	
$\mathbf{2\overline{3}}$	$\frac{1}{4} \cdot \frac{34}{36}$		

  where the first entry is the number of heads, and the second one is the result of rolling the die, simplified to tell us only whether the total dots equaled **3,** or did not ($\overline{\mathbf{3}}$). $\Pr(\mathbf{1}|\mathbf{3}) = \frac{\frac{1}{12}}{\frac{1}{12} + \frac{1}{72}} = 85.71\%$. Note that here, rather atypically, we used bold digits as *names* of events.

- Jim, Joe, Tom and six other boys are randomly seated in a row. What is the probability that at least two of the three friends will sit next to each other?

  Solution: Let's introduce A: 'Jim and Joe sit together', B: 'Jim and Tom sit together', C: 'Joe and Tom sit together'. We need $\Pr(A \cup B \cup C) = \Pr(A) + \Pr(B) + \Pr(C) - \Pr(A \cap B) - \Pr(A \cap C) - \Pr(B \cap C) + \Pr(A \cap B \cap C)$. There is 9! random arrangements of the boys, $2 \times 8!$ will meet condition A (same with B and C), $2 \times 7!$ will meet both A and B (same with $A \cap C$ and $B \cap C$), *none* will meet all three.

  Answer: $3 \times \frac{2 \times 8!}{9!} - 3 \times \frac{2 \times 7!}{9!} = 58.33\%$.

- (From a former exam – these are usually a touch easier): Shuffle a deck of 52 cards. What is the probability that the four aces will end up next to each other (as a group of four consecutive aces)?

  Answer: $\frac{4! \times 49!}{52!} = 0.0181\%$ $(= \frac{1}{5525})$ [for small probabilities, the last number − telling us that this will happen, on the average, only in 1 out of 5525 attempts – conveys more information than the actual percentage].

- Consider a 10 floor government building with all floors being equally likely to be visited. If six people enter the elevator (individually, i.e. independently) what is the probability that they are all going to (six) different floors?

  Solution: The experiment is in principle identical to rolling a 9-sided die (there are nine floors to be chosen from, exclude the main floor!) six times (once for each person − this corresponds to selecting his/her floor). The sample space thus consists of 9^6 equally likely outcomes (each looking like this: 2 4 8 6 9 4 − ordered selection, repetition allowed). Out of these, only $9 \times 8 \times 7 \times 6 \times 5 \times 4 = P_6^9$ consist of all distinct floors.

  Answer: $\frac{P_6^9}{9^6} = 11.38\%$.

- (*Extension* of the previous example). What if the floors are not equally likely [they issue licences on the 4^{th} floor, which has therefore a higher probability of $\frac{1}{2}$ to be visited by a 'random' arrival − the other floors remain equally likely with the probability of $\frac{1}{16}$ each].

  Solution: The sample space will be the same, but the individual probabilities will no longer be identical; they will now equal to $(\frac{1}{2})^i(\frac{1}{16})^{6-i}$ where i is how many times 4 appears in the selection [2 4 8 6 9 4 will have the probability of $(\frac{1}{2})^2(\frac{1}{16})^4$, etc.]. We have to single out the outcomes with all six floors different and *add* their probabilities. Luckily, there are only two types of these outcomes: (i) those *without* any 4: we have P_6^8 of these, each having the probability of $(\frac{1}{16})^6$, and (ii) those with a *single* 4: there are $6 \times P_5^8$ of these, each having the probability of $(\frac{1}{2})(\frac{1}{16})^5$.

  Answer: $P_6^8(\frac{1}{16})^6 + 6P_5^8(\frac{1}{2})(\frac{1}{16})^5 = 2.04\%$ (the probability is a lot smaller now).

- Within the next hour 4 people in a certain town will call for a cab. They will choose, randomly, out of 3 existing (equally popular) taxi companies. What is the probability that no company is left out (each gets at least one job)?

Solution: This is again a roll-of-a-die type of experiment (this time we roll 4 times − once for each customer − and the die is 3-sided − one side for each company). The sample space will thus consist of 3^4 equally likely possibilities, each looking like this: 1321. How many of these contain all three numbers? To achieve that, we obviously need one duplicate and two singles. There are 3 ways to decide which company gets two customers. Once this decision has been made (say 1 2 2 3), we simply permute the symbols [getting $\binom{4}{2,1,1}$ distinct 'words'].

Answer: $\dfrac{3 \times \binom{4}{2,1,1}}{3^4} = \frac{4}{9} = 44.44\%$.

- There are 10 people at a party (no twins). Assuming that all 365 days of a year are equally likely to be someone's birth date [not quite, say the statistics, but we will ignore that] and also ignoring leap years, what is the probability of:

  1. All these ten people having different birth dates?

     Solution: This, in principle, is the same as choosing 6 different floors in an elevator (two examples ago).

     Answer: $\dfrac{P_{10}^{365}}{365^{10}} = 88.31\%$.

  2. Exactly two people having the same birth date (and no other duplication).

     Solution: This is similar to the previous example where we needed exactly one duplicate. By a similar logic, there are 365 ways to choose the date of the duplication, $\binom{10}{2}$ ways of placing these into 2 of the 10 empty slots, and P_8^{364} of filling out the remaining 8 slot with distinct birth dates.

     Answer: $\dfrac{365 \times \binom{10}{2} \times P_8^{364}}{365^{10}} = 11.16\%$ (seems reasonable).

  These two answers account for 99.47% of the total probability. Two or three duplicates, and perhaps one triplicate would most likely take care of the rest; try it!

- A simple padlock is made with only ten distinct keys (all equally likely). A thief steals, independently, 5 of such keys, and tries these to open your lock. What is the probability that he will succeed?

  Solution: Again, a roll-of-a-die type of experiment (10 sides, 5 rolls). The question is in principle identical to rolling a die to get at least one six. This, as we already know, is easier through the corresponding complement.

  Answer: $1 - (\frac{9}{10})^5 = 40.95\%$. ■

Chapter 3 RANDOM VARIABLES – DISCRETE CASE

If each (complete) outcome [simple event] of a random experiment is *assigned a single real number* (usually an integer), this (assignment) is called a **random variable** (RV). Using the same experiment we can define any number of random variables, and call them X, Y, Z, etc. (capital letters from the end of the alphabet).

EXAMPLE: Using the experiment of rolling two dice, we can define X as the *total* number of dots, and Y as the *larger* of the two numbers. This means assigning numbers to individual simple events in the following fashion:

X:

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

Y:

1	2	3	4	5	6
2	2	3	4	5	6
3	3	3	4	5	6
4	4	4	4	5	6
5	5	5	5	5	6
6	6	6	6	6	6

Note the **difference** between *events* and *random variables*: an event is effectively an assignment, to each outcome, of either 'yes' (✓, meaning: I am in) or 'no' (blank, meaning: I am out). E.g. Event A: 'the total number of dots is even' will be represented by:

✓		✓		✓	
	✓		✓		✓
✓		✓		✓	
	✓		✓		✓
✓		✓		✓	
	✓		✓		✓

Probability distribution of a random variable

is a **table** (or formula) summarizing the information about

1. possible outcomes of the RV (numbers, arranged from the smallest to the largest)
2. the corresponding probabilities. ■

Thus, for example, our X and Y have the following (probability) distributions:

$X =$	2	3	4	5	6	7	8	9	10	11	12
Pr:	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

and

$Y =$	1	2	3	4	5	6
Pr:	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

The probabilities of each distribution must of course add up to 1 (checking this is a lot easier if you use the same denominator).

Later on we will find it more convenient to express the same information using **formulas** instead of tables (we will stick to tables for as long as we can, i.e. for the rest of this chapter). Thus, for example the distribution of X can be specified by: $f_X(i) = \frac{6-|i-7|}{36}$ with $i = 2, 3, ...12$ where $f_X(x)$ is the so called PROBABILITY FUNCTION of X. Similarly: $f_Y(i) = \frac{2i-1}{36}$ with $i = 1, 2, ...6$ (these being the potential values of Y, $f_Y(i)$ computing the corresponding probability).

Formulas become more convenient when dealing with RVs having too many (sometimes infinitely many) values. Thus, for **example**, if we go back to the experiment of flipping a coin till a head appears, and define X as the total number of tosses (anything which translates an outcome of an experiment into a single number is a RV), we have a choice of either

$X =$	1	2	3	4	.....	i	.....
Pr:	$\frac{1}{2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	.....	$\frac{1}{2^i}$	.....

or $f_X(i) = \frac{1}{2^i}$ with $i = 1, 2, 3,....$ [implying: up to infinity]. In this case, one would usually prefer the formula to an unwieldy table.

Sometimes it's useful to have a graph (**histogram**) of a distribution. The probabilities are usually displayed as vertical bars or (connected) rectangles. We get a nice graphical view of what's likely and what is not.

▶Distribution Function of a Random Variable◀

At this point the names may get a bit confusing. Learn to differentiate between a *probability distribution* [a more generic term implying a complete information about a RV, usually in a form of a table or a *probability function* $f(i)$] and a *distribution function* (I like to call it 'capital F'), which is **defined** by: $F_X(k) = \Pr(X \leq k)$ i.e. effectively a table (or a formula) providing *cumulative* (i.e. total) probabilities of intervals of values, from the smallest up to and including k.

Thus, using one of our previous **example**s:

$Y =$	1	2	3	4	5	6
F_Y:	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{9}{36}$	$\frac{16}{36}$	$\frac{25}{36}$	1

It is obvious that the values of $F(k)$ can only increase with increasing k, and that the last one must be equal to 1. When there is no last value (i.e. k can go up to infinity), it is the $\lim_{k\to\infty} F_X(k)$ which must equal to 1.

EXAMPLE: The total number of tosses in the flip-a-coin experiment has the following distribution function: $F_X(k) = \sum_{i=1}^{k} (\frac{1}{2})^i = \frac{1}{2} \cdot \frac{1-\left(\frac{1}{2}\right)^k}{1-\frac{1}{2}} = 1 - (\frac{1}{2})^k$, for $k = 1, 2, 3,...$ (its argument is a 'dummy' variable, it makes no difference whether we call it i, j, k, or anything else − your textbook calls it x, but I don't like that notation). Obviously, $\lim_{k\to\infty} F_X(k) = 1$ (check). ■

Multivariate distribution

of **several** *random variables* (we have already mentioned that more than one RV can be defined for the same experiment). We start with the

▶Distribution of Two Random Variables◀

A two dimensional table which, for every combination of the RVs' values specifies the respective probability, is called their (bivariate) **joint distribution.**

The same information can be usually given (in a more 'compact' form) by the corresponding **probability function** $f(i, j)$ and the range of possible i (the first RV's) and j (the second RV's) values. Unlike you textbook, I usually like to include the names of the two RVs as subscripts, thus: $f_{XY}(i, j)$.

EXAMPLE: A coin is flipped three times. X: total number of tails, Y: number of heads up to the first tail. We first display the sample space with the values of X and Y:

Prob:	Outcome:	X:	Y:
$\frac{1}{8}$	HHH	0	3
$\frac{1}{8}$	HHT	1	2
$\frac{1}{8}$	HTH	1	1
$\frac{1}{8}$	HTT	2	1
$\frac{1}{8}$	THH	1	0
$\frac{1}{8}$	THT	2	0
$\frac{1}{8}$	TTH	2	0
$\frac{1}{8}$	TTT	3	0

The joint distribution of X and Y follows:

$Y =$ / $X =$	0	1	2	3
0	0	0	0	$\frac{1}{8}$
1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	0
2	$\frac{2}{8}$	$\frac{1}{8}$	0	0
3	$\frac{1}{8}$	0	0	0

(*)

Thus, for example $\Pr(X = 2 \cap Y = 0) = \frac{2}{8}$, etc.

Trying to express this distribution via $f_{XY}(i, j)$ would be rather difficult, and the resulting function very unwieldy (to say the least) – there is no point attempting it. ■

The **joint distribution function** is defined as $F_{XY}(i, j) = \Pr(X \leq i \cap Y \leq j)$. It's not going to be used by us much.

Marginal distribution of X (and, similarly, of Y) is, effectively the *ordinary* (UNIVARIATE) distribution of X (as if Y has never been defined). It can be obtained from the bivariate (joint) distribution by adding the probabilities in each row [over all possible Y-values, using the total probability formula: $\Pr(X = 0) = \Pr(X = 0 \cap Y = 0) + \Pr(X = 0 \cap Y = 1) + \Pr(X = 0 \cap Y = 2) + \Pr(X = 0 \cap Y = 3)$; in this context one must realize that $Y = 0$, $Y = 1$, $Y = 2$, ... are *events*, furthermore,

they constitute a *partition* of the sample space]. The results are conveniently displayed at the table's margin (thus the name):

$Y=$ / $X=$	0	1	2	3	
0	0	0	0	$\frac{1}{8}$	$\frac{1}{8}$
1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	0	$\frac{3}{8}$
2	$\frac{2}{8}$	$\frac{1}{8}$	0	0	$\frac{3}{8}$
3	$\frac{1}{8}$	0	0	0	$\frac{1}{8}$

implying that the marginal (i.e. 'ordinary') distribution of X is

$X=$	0	1	2	3
Pr:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Similarly, one can find the marginal distribution of Y by adding the probabilities in each column.

A bivariate distribution is often given to us via the corresponding joint probability function. One of the two ranges has usually the '*marginal*' form (the limits are constant), the other range is '*conditional*' (i.e. both of its limits may depend on the value ot the ohter random variable). The best way is to 'translate' this information into an explicit table whenever possible.

EXAMPLE: Consider the following bivariate probability function of two random variables X and Y:

$$f_{XY}(i,j) = c \cdot (2i + j^2) \qquad \text{where} \qquad \begin{array}{l} 0 \le i \le 2 \\ i \le j \le 4 - i \end{array}$$

Find the value of c, the marginal distribution of Y and (based on this) $\Pr(Y \le 2)$.

Solution: We translate the above information into the following table

$Y=$	1	2	3	4
$X=0$	$1c$	$4c$	$9c$	$16c$
1	$3c$	$6c$	$11c$	0
2	0	$8c$	0	0

which clearly implies that $c = \frac{1}{58}$, the marginal distribution of Y is

$Y=$	1	2	3	4
Pr	$\frac{4}{58}$	$\frac{18}{58}$	$\frac{20}{58}$	$\frac{16}{58}$

and $\Pr(Y \le 2) = \frac{22}{58} = 37.93\%$. ■

Independence of X and Y is almost always a consequence of X and Y being defined based on two distinct parts of the experiment which, furthermore, cannot influence each other's outcome(e.g. rolling a die 4 times, let X be the total number of dots in the first two rolls, and Y be the total number of dots in the last two

rolls). Normally, we should be able to tell, based on this, that X and Y *must* be independent, and utilize the consequences of any such *natural* independence.

Formally, X and Y being independent means that $\Pr(X = i \cap Y = j) = \Pr(X = i) \times \Pr(Y = j)$ for *every* possible combination of i and j (each joint probability is a *product* of the two corresponding marginal probabilities). [We can readily see that, in the above example, X and Y are *not* independent, since $0 \neq \frac{1}{8} \times \frac{4}{8}$]. This implies that, when X and Y are independent, their marginal distributions enable us to compute each and every of their *joint* probabilities by a simple multiplication (we usually don't need to construct the corresponding, now fully *redundant,* join probability table).

All of these concepts can be extended to

►Three or More Random Variables◄

In such a case we usually don't like working with (3 or more-dimensional) tables; we will have to rely on formulas.

$$\Pr(X = i \cap Y = j \cap Z = k) \equiv f_{XYZ}(i, j, k)$$

defines the (joint) **probability function**; it *must* be accompanied by stipulating the *permissible ranges* of i, j and k (f given without this information would be meaningless).

Based on this, we are able to instantly determine whether the corresponding RVs are **independent** or not, since their independence requires that:

- $f(i, j, k)$ can be written as a product of a function of i times a function of j times a function of k,

  and

- the i, j, and k ranges are (algebraically) independent of each other (i.e. both the lower and upper limit of *each* range are fixed numbers, *not* functions of the other two variables).

When either of these two conditions is violated (even if it is only by one item), the two RVs are *dependent.*

EXAMPLES:

1. $f_{XY}(i, j) = \frac{i+j}{24}$, where $1 \leq i \leq 3$ and $1 \leq j \leq i$, clearly implies that X and Y are *not* independent (here, both conditions are broken). To deal with this bivariate distribution, my advice is to 'translate' it into an explicit table of joint probabilities, whenever possible (try it with this one).

2. $f_{XYZ}(i, j, k) = \frac{i \cdot j \cdot k}{108}$, where $1 \leq i \leq 3$, $1 \leq j \leq 3$ and $1 \leq k \leq 2$. Yes, both conditions are met, therefore X, Y and Z *are* independent. [It is then very easy to establish the individual marginals, e.g.: $f_X(i) = c \cdot i$ with $1 \leq i \leq 3$, where c is a constant which makes the probabilities add up to 1 ($\frac{1}{6}$ in this case)]. ■

Finally, an **important note** about ranges: The are two distinct ways of specifying a two-dimensional region: we can start by the 'marginal' range of i values and follow it by the 'conditional' range of j values, or we can do it the other way around (both descriptions are equally correct, but often appear quite distinct). When constructing a marginal distribution, the summation *must* be done over the '*conditional*' range of the *other* variable; thus, when working with formulas [instead of explicit tables], one must always make sure that the ranges are in the appropriate form (and be able to do the 'translation' when they are not).

EXAMPLE: $f(i,j) = \frac{(i+1)\cdot(j+1)}{60}$, where $0 \le i \le 2$ and $i \le j \le i+2$ [to understand the example, try translating this into the corresponding table of probabilities]. Note that the corresponding two variables are *not* independent. The other way of expressing the ranges (this time, it is rather tricky) is: $0 \le j \le 4$ and $\max(0,j) \le i \le \min(2,j)$ [I hope you understand the max/min notation − verify that this is so!]. ■

Similarly, there are 6 (in general) distinct ways of stipulating the ranges of i, j and k (we may start with the marginal range of i, follow with the j-range given i, and finally the k-range given both i and j; obviously having 3! choices as to the order). This is very important for us to understand, especially when reaching the continuous distributions [students always have difficulties at this point].

▶Conditional Distribution◀

of X, given an (observed) value of Y.

Using the old notion for conditional probabilities, we know that

$$\Pr(X = i \,|\, Y = \mathbf{j}) = \frac{\Pr(X = i \cap Y = \mathbf{j})}{\Pr(Y = \mathbf{j})}$$

All we have to do is to introduce a **new notation** for these, namely: $f_{X|Y=\mathbf{j}}(i)$ where i varies over its *conditional* range (given the value of $\mathbf{j}$; we use a different print type to emphasize that j has a specific, fixed value).

These probabilities (for all such i values) constitute a new, special probability distribution called CONDITIONAL DISTRIBUTION which has, nevertheless, all the **properties** of an *ordinary* distribution. They usually arise in a situation when a specific value of one RV has already been observed, but one is still waiting for the outcome of the other.

EXAMPLE: Using Table (*) of our original example, we can easily construct

$Y \,\|X=2$	0	1
Prob:	$\frac{2}{3}$	$\frac{1}{3}$

by taking the probabilities in the $X = 2$ row and dividing each of them by their total (the corresponding *marginal* probability of $X = 2$). Note that values with zero probability have been discarded.

Similarly:

$X \,\|Y=0$	1	2	3
Prob:	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

■

Things get more tricky when dealing with **three** (or more) **Random Variables**. One can define a conditional distribution of *one* of them, given a value of each of the other two, say:

$$\Pr(X = i \,|Y = \mathbf{j} \cap Z = \mathbf{k}) = \frac{f_{XYZ}(i, \mathbf{j}, \mathbf{k})}{f_{YZ}(\mathbf{j}, \mathbf{k})}$$

with i varying over all values permitted by $\mathbf{j}$ and $\mathbf{k}$ (fixed),

or a conditional (and joint) distribution of *two* of them, given a value of the third:

$$\Pr(X = i \cap Y = j \,|Z = \mathbf{k}) = \frac{f_{XYZ}(i, j, \mathbf{k})}{f_Z(\mathbf{k})}$$

with i and j varying over all pairs of values allowed by $\mathbf{k}$.

Mutual **independence implies** that all *conditional* distributions are *identical* to the corresponding *marginal* distribution. For example, when X, Y and Z are mutually independent, $\Pr(X = i \,|Y = \mathbf{j}) \equiv \Pr(X = i)$, $\Pr(X = i \,|Y = \mathbf{j} \cap Z = \mathbf{k}) = \Pr(X = i)$, etc. [$X$ has the same distribution, whatever the value of the other variable(s)].

The rule to remember: Under mutual independence it is legitimate to simply *ignore* (remove) the condition(s).

Transforming random variables

It should be obvious that, if X is a random variable, any transformation of X (i.e. an expression involving X, such as $\frac{X}{2} + 1$) defines a new random variable (say Z) with its own new distribution. This follows from our basic definition of a random variable (the experiment returns a single number).

EXAMPLE: If X has a distribution given by

$X =$	0	1	2	3
Prob:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

then, to build a distribution of $Z = \frac{X}{2} + 1$, one simply replaces the first-row values of the previous table, thus:

$Z =$	1	$\frac{3}{2}$	2	$\frac{5}{2}$
Prob:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

.

Similarly, if the new RV is $U = (X - 2)^2$ [one can define any number of new RVs based on the same X], using the same approach the new table would look:

$U =$	4	1	0	1
Prob:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Here we of course don't like the duplication of values and their general 'disorder', so the same table should always be presented as:

$U =$	0	1	2	3	4
Prob:	$\frac{3}{8}$	$\frac{4}{8}$	0	0	$\frac{1}{8}$

[the values 2 and 3 have been inserted, with zero probabilities, to make the table more 'regular' − doing this is optional]. ■

The most important such case is the so called **linear transformation** of X, i.e.

$$Y = aX + b$$

where a and b are two constants. Note that the *shape* (in terms of a histogram) of the Y-distribution is the *same* as that of X, only the horizontal scale has different tick marks now. The new random variable is effectively the old random variable on a new scale (such as expressing temperature in Celsius to define X and in Fahrenheit to 'transform it' to Y). This is why linear transformations are particularly easy to deal with, as we will see later.

Similarly, we can

►Transform Two Random Variables◄

into a single one by using any mathematical expression (function) of the two.

EXAMPLE: If X and Y have the distribution of our old bivariate example, and $W = |X - Y|$, we can easily construct the (univariate) distribution of W by first building a table which shows the *value* of W for each X, Y combination:

$Y =$ / $X =$	0	1	2	3
0	0	1	2	3
1	1	0	1	2
2	2	1	0	1
3	3	2	1	0

and then collect the probabilities of each *unique* value of W, from the smallest to the largest, thus:

$W =$	0	1	2	3
Prob:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{2}{8}$

. This is the resulting distribution of W. ■

Transforming RVs is a lot more fun in the continuous case (a few months from now).

Chapter 4 EXPECTED VALUE OF A RANDOM VARIABLE

also called its **mean** or **average**, is a number which corresponds (empirically) to the average value of the random variable when the experiment is repeated, independently, infinitely many times (i.e. it is the *limit* of such averages). We can compute it based on the RV's distribution by realizing that the individual *probabilities* (second row) represent the limit of the OBSERVED FREQUENCIES of the corresponding values (first row).

For **example**, the probability of $\frac{1}{6}$ (of getting a six when rolling a die) is telling us that, in a long run, one sixth of all outcomes will have that value (exact only in the infinite limit), etc. One also has to remember that averaging (i.e. the very simple 'add all values and divide by their number') can be simplified by $\frac{1 \times k_1 + 2 \times k_2 + ... + 6 \times k_6}{k_1 + k_2 + ... + k_6}$ where k_1 is the number of outcomes which resulted in a 1 (dot), k_2 is the number of those with 2 (dots), etc. [the numerator still gives the simple sum of all observed values, and the denominator their number]. This can be rewritten as $1 \times r_1 + 2 \times r_2 + ... + 6 \times r_6$ where $r_1 = \frac{k_1}{k_1+k_2+...+k_6}$ is the relative frequency of outcome 1, $r_2 = \frac{k_2}{k_1+k_2+...+k_6}$ is the relative frequency of outcome 2, etc. In the infinite limit the relative frequencies become the corresponding probabilities, $r_1 \to f(1)$, $r_2 \to f(2)$, etc. [recall that $f(1) \equiv \Pr(X = 1), ...$]. The expected (average) value of a RV X is thus $1 \times f(1) + 2 \times f(2) + ... + 6 \times f(6)$ or, in **general**,

$$\mathbb{E}(X) = \sum_{i=0}^{n} i \times f(i)$$

where the summation is from the smallest value (usually 0 or 1) to the largest possible value, say n. Note that this simply means multiplying the numbers of the first row (of a distribution table) by the corresponding numbers of the second row, and adding the results.

$\mathbb{E}(X)$ is the usual **notation** for the expected value of X. Sometimes, for the same thing, we also use the following *alternate* ('shorthand') notation: μ_X (μ is the Greek letter 'mu', not to be confused with u). We will often refer to the process of taking the expected value of a RV (or an expression involving RVs) as 'averaging' [since *weighted averaging* it is].

EXAMPLES:

1. When X is the number of dots in a single roll of a die, this gives $\mathbb{E}(X) = \frac{1+2+3+4+5+6}{6} = 3.5$. Note that this is the exact *center* (of symmetry) of the distribution. This observation is true for any symmetric distribution, which enables us to bypass the computation in such cases. Also note that the result (3.5) is *not* one of the possible values. Thus the name (expected value) is rather misleading, it is *not* the value we would *expect* to get in any roll.

2. Let Y be the larger (max) of the two numbers when rolling two dice [we constructed its distribution in the previous chapter]. $\mathbb{E}(Y) = 1 \times \frac{1}{36} + 2 \times$

$\frac{3}{36}+3\times\frac{5}{36}+4\times\frac{7}{36}+5\times\frac{9}{36}+6\times\frac{11}{36}=\frac{1+6+15+28+45+66}{36}=\frac{161}{36}=4.47\overline{2}$. In a probability histogram, where probabilities are represented by (heavy) bars, this would correspond to their 'center of mass' (if the x-axis is seen as a weightless platform, this is the point at which the structure could be supported without tilting). This enables us to roughly estimate the mean and detect a possible computational error when it happens (if nothing else, make sure that the answer is within the RV's range)!

3. Let U have the following (arbitrarily chosen) distribution:

$U=$	0	1	2	4
Prob:	0.3	0.2	0.4	0.1

$\mathbb{E}(U)=0.2+0.8+0.4=1.4$. ■

When ▶Transforming◀

a random variable, what happens to **the mean**?

The main thing to remember is that in general (unless the transformation is linear − to be discussed later) the mean does *not* transform accordingly, and has to be computed anew, i.e. $\mathbb{E}[g(X)]\neq g\left(\mathbb{E}[X]\right)$ [same as $g(\mu_X)$ in 'shorthand'], the simplest but rather important example being: $\mathbb{E}(X^2)\neq\left(\mathbb{E}(X)\right)^2$ $[=\mu_X^2]$!!!

EXAMPLE: Related to the previous problem, we define $W=|U-2|^3$. It would be a big *mistake* to assume that $\mathbb{E}[|U-2|^3]\stackrel{?}{=}|\mathbb{E}(U)-2|^3=0.6^3=0.216$, it is *not*. There are two ways of computing the *correct* expected value, as follows:

- We can simply build the distribution of the new RV (the way we learned in the previous chapter), and then use the basic procedure of computing its mean:

$W=$	0	1	8
Prob:	0.4	0.2	0.4

$\Rightarrow\mathbb{E}(W)=0.2+3.2=3.4$.

- We can use the old distribution, adding an extra row for the new RV's values:

$W=$	8	1	0	8
$U=$	0	1	2	4
Prob:	0.3	0.2	0.4	0.1

and then perform the following 'weighted' averaging of W: $\mathbb{E}(W)=8\times0.3+1\times0.2+0\times0.4+8\times0.1$ resulting in the same answer of 3.4. ■

The **equivalence** of the two techniques is true in general, we can summarize it by the following formula:

$$\mathbb{E}[W=g(U)]=\sum_{All\ j}j\times f_W(j)=\sum_{All\ i}g(i)\times f_U(i)$$

where $g(U)$ is the actual transformation [$g(U)\equiv|U-2|^3$ in our example].

i.e. those of the type $Y = aX + b$ (where a and b are two constants), the situation is different; we can prove easily that

$$\mathbb{E}(aX + c) = a\mathbb{E}(X) + c$$

Proof: $\mathbb{E}(aX+c) = \sum_{All\ i}(ai+c)f_X(i) = a\sum_{All\ i} i \times f_X(i) + c\sum_{All\ i} f_X(i) = a\mathbb{E}(X) + c$ □

EXAMPLE: $\mathbb{E}(2U - 3) = 2 \times 1.4 - 3 = -0.2$ [where U is the variable of the previous sections]. Verify this by a direct technique (you have a choice of two) and note the significant simplification achieved by this formula. ■

Expected values related to a bivariate distribution

When a bivariate distribution (of two RVs) is given, the easiest way to compute the **individual** expected values (of X and Y) is through the marginals.

EXAMPLE: Based on

		$X=$ 1	2	3	
$Y=$	0	0.1	0	0.3	0.4
	1	0.3	0.1	0.2	0.6
		0.4	0.1	0.5	

we compute $\mathbb{E}(X) = 1 \times 0.4 + 2 \times 0.1 + 3 \times 0.5 = 2.1$ and $\mathbb{E}(Y) = 0 \times 0.4 + 1 \times 0.6 = 0.6$. ■

We know that any **function of** X **and** Y (e.g. their simple product $X \cdot Y$) is a new random variable with its own distribution and therefore its mean. How do we compute $\mathbb{E}(X \cdot Y)$? Again, we have two ways of doing this, either by building the distribution of $Z = X \cdot Y$ and using the usual formula, or by multiplying the (joint) probabilities of the bivariate table by the corresponding value of $X \cdot Y$ and adding the results (over the whole table, i.e. both rows and columns), thus (using the previous example): $\mathbb{E}(X \cdot Y) = 1 \times 0 \times 0.1 + 2 \times 0 \times 0 + 3 \times 0 \times 0.3 + 1 \times 1 \times 0.3 + 2 \times 1 \times 0.1 + 3 \times 1 \times 0.2 = 1.1$.

This means that in **general** we have

$$\mathbb{E}\left[g(X,Y)\right] = \sum_{Rows}\sum_{Columns} g(i,j) \times f_{XY}(i,j)$$

More EXAMPLES [based on the previous bivariate distribution]:

1. $\mathbb{E}\left[(X-1)^2\right] = 0^2 \times 0.4 + 1^2 \times 0.1 + 2^2 \times 0.5 = 2.1$ [here we used the X-marginal, bypassing the 2-D table].

2. $\mathbb{E}\left[\frac{1}{1+Y^2}\right] = \frac{1}{1+0^2} \times 0.4 + \frac{1}{1+1^2} \times 0.6 = 0.7$ [similarly, use the Y-marginal].

3. $\mathbb{E}\left[\frac{(X-1)^2}{1+Y^2}\right]$ [don't try to multiply the last two results, that would be *wrong*]. Here it may help to first build the corresponding table of the $\frac{(X-1)^2}{1+Y^2}$ values: $\boxed{\begin{array}{ccc} 0 & 1 & 4 \\ 0 & \frac{1}{2} & 2 \end{array}}$, then multiply each item of this table by the corresponding item of the probability table and add the results: $1.2 + 0.05 + 0.4$ [discarding zero values] $= 1.65$. ■

▶Linear Case◀

Please note that in general we again *cannot* equate $\mathbb{E}[g(X,Y)]$ with $g(\mu_X, \mu_Y)$, *unless* the function is **linear** (in both X and Y) i.e. $g(X,Y) \equiv aX + bY + c$. Then we have:

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$$

Proof: $\mathbb{E}[aX + bY + c] = \sum_i \sum_j (a \times i + b \times j + c) f_{XY}(i,j) = a \sum_i i \times f_X(i) + b \sum_j j \times f_Y(j) + c = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$. Note that i is treated as a constant by the j summation and vice versa. □

EXAMPLE: Using the previous bivariate distribution, $\mathbb{E}(2X - 3Y + 4)$ is simply $2 \times 2.1 - 3 \times 0.6 + 4 = 6.4$ ■

The previous **formula** easily extends to any number of variables:

$$\mathbb{E}[a_1X_1 + a_2X_2 + ... + a_kX_k + c] = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + ... + a_k\mathbb{E}(X_k) + c$$

Note that *no* assumption of *independence* was made about these variables!

▶Independence Related Issues◀

Can independence *help* when computing *some* of our expected values? The answer is yes, the expected value of a *product* of RVs equals the product of the individual expected values, when these RVs are **independent**:

$$X \bowtie Y \quad \Rightarrow \quad \mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

where $\bowtie$ is our notation for (*pairwise*) independence.

Proof: $\mathbb{E}(X \cdot Y) = \sum_i \sum_j i \times j \times f_X(i) \times f_Y(j) = \left(\sum_i i \times f_X(i)\right) \times \left(\sum_j j \times f_Y(j)\right) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$ □

The statement can actually be made more **general**:

$$X \bowtie Y \quad \Rightarrow \quad \mathbb{E}[g_1(X) \cdot g_2(Y)] = \mathbb{E}[g_1(X)] \cdot \mathbb{E}[g_2(Y)]$$

where g_1 and g_2 are any two (univariate) functions (Proof would be practically the same).

Moments of a single random variable

We now return to the case of a single RV and define its so called MOMENTS as follows: $\mathbb{E}(X^n)$ where n is an integer is called the n^{th} **simple moment** of X (or, of the corresponding distribution). The mean $\mathbb{E}(X)$ is thus the *first* simple moment (yet another name for the same thing!), $\mathbb{E}(X^2)$ is the *second* simple moment (remember, it is *not* equal to the first moment squared!), etc. The zeroth moment, $\mathbb{E}(X^0 \equiv 1)$, is always identically equal to 1.

Similarly we can define the so called **central moments** (your textbook calls them 'moments with respect to the mean', but that's too long for us!) as $\mathbb{E}\left[(X-\mu_X)^n\right]$, where $\mu_X \equiv \mathbb{E}(X)$ [as we know already]. Thus, the first central moment $\mathbb{E}(X-\mu_X) = \mu_X - \mu_X \equiv 0$ [always identically equal to zero, and of not much use].

The second central moment $\mathbb{E}\left[(X-\mu_X)^2\right] = \mathbb{E}\left[X^2 - 2X\mu_X + \mu_X^2\right] = \mathbb{E}(X^2) - 2\mu_X^2 + \mu_X^2 = \mathbb{E}(X^2) - \mu_X^2 \geq 0$ (averaging non-negative quantities cannot result in a negative number; also, the last expression is more convenient computationally than the first) is of such importance that it goes under yet another name, it is called the RV's **variance,** notation: $Var(X)$. Its purpose is to measure the spread (width) of the distribution by finding a typical, 'average' deviation of the observations from its 'center' μ_X. Note that averaging the deviation $X-\mu_X$ directly would have given us zero, as the positive and negative values cancel out. One could propose averaging $|X-\mu_X|$ to correct that problem, but this would create all sorts of difficulties, both computational and 'theoretical' (as we will see later). So we have averaged $(X-\mu_X)^2$, which also gets rid of the cancellation problem (and more 'elegantly' so), but results in the average *squared* deviation [if X is length, its variance will be in square inches − wrong units].

This can be fixed by simply taking the square root of the variance [which finally gives us this 'typical' deviation from the mean] and calling it the **standard deviation** of X, notation: $\sigma_X = \sqrt{Var(X)}$ [this is the Greek letter 'sigma']. One can make a rough estimate of σ from the graph of the distribution (the interval $\mu-\sigma$ to $\rho+\sigma$ should contain the 'bulk' of the distribution − anywhere from 50 to 90%); this rough rule should detect any gross mistake on our part.

Finally, SKEWNESS is defined as $\frac{\mathbb{E}\left[(X-\mu_X)^3\right]}{\sigma^3}$ [it measures to what extent is the distribution non-symmetric, or better yet: left (positively) or right (positively) 'skewed'], and KURTOSIS as $\frac{\mathbb{E}\left[(X-\mu_X)^4\right]}{\sigma^4}$ [it measures the degree of 'flatness', 3 being a typical value, higher for 'peaked', smaller for 'flat' distributions]. The last two quantities (unlike the variance) are of only a marginal importance to us.

EXAMPLES:

1. If X is the number of dots when rolling one die, $\mu_X = \frac{7}{2}$ [we computed that already], $Var(X) = \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$ [we used the 'computational' formula $\mathbb{E}(X^2) - \mu_X^2$, verify that $\mathbb{E}\left[(X-\mu_X)^2\right]$ results in the same answer, but it is clumsier to use]. This implies that the standard deviation $\sigma_X = \sqrt{\frac{35}{12}} = 1.7078$. Note that 3.5 ± 1.708 contains 66.7% of the distribution. Skewness, for a symmetric distribution, must be equal to 0, kurtosis can

be computed based on $\mathbb{E}[(X-\mu)^4] = \frac{(-2.5)^4+(-1.5)^4+(-0.5)^4+0.5^4+1.5^4+2.5^4}{6} = 14.729 \Rightarrow \text{kurtosis} = \frac{14.729}{(\frac{35}{12})^2} = 1.7314$ ['flat'].

2. Consider the distribution of one of our previous examples: $\begin{array}{c|cccc} U = & 0 & 1 & 2 & 4 \\ \hline \text{Prob:} & 0.3 & 0.2 & 0.4 & 0.1 \end{array}$. $\mu_U = 1.4$ [already computed], $Var(U) = 0^2 \times 0.3 + 1^2 \times 0.2 + 2^2 \times 0.4 + 4^2 \times 0.1 - 1.4^2 = 1.44 \Rightarrow \sigma_U = \sqrt{1.44} = 1.2$. From $\mathbb{E}[(U-\mu_U)^3] = (-1.4)^3 \times 0.3 + (-0.4)^3 \times 0.2 + 0.6^3 \times 0.4 + 2.6^3 \times 0.1 = 1.008$, the skewness is $\frac{1.008}{1.2^3} = .58333$ [long right tail], and from $\mathbb{E}[(U-\mu_U)^4] = (-1.4)^4 \times 0.3 + (-0.4)^4 \times 0.2 + 0.6^4 \times 0.4 + 2.6^4 \times 0.1 = 5.7792$ the kurtosis equals $\frac{5.7792}{1.2^4} = 2.787$ ■

When X is transformed to define $Y = g(X)$, we already know that there is no general 'shortcut' for computing $\mathbb{E}(Y)$. This (even more so) applies to the **variance** of Y, which also needs to be computed 'from scratch'. But, we did manage to simplify the expected value of a **linear transformation** of X (of the $Y = aX + c$ type). Is there any *simple* conversion of $Var(X)$ into $Var(Y)$ in this (linear) case?

The answer is 'yes', and we can easily derive the corresponding formula: $Var(aX + c) = \mathbb{E}[(aX+c)^2] - (a\mu_X + c)^2 = \mathbb{E}[a^2X^2 + 2aX + a^2] - (a\mu_X + c)^2 = a^2\mathbb{E}(X^2) - a^2\mu_X^2$ [the rest cancel] $= a^2 Var(X)$ [note that c drops out entirely as expected, it corresponds to the change of origin only – 'sliding' the distribution by a fixed amount c, which does not change its width]. This implies that

$$\sigma_{aX+c} = |a|\sigma_X$$

(don't forget that $\sqrt{a^2} = |a|$, not a).

Moments – the bivariate case

When dealing with a joint distribution of two RVs we can always compute the *individual* (**single-variable) moments** (means, variances, etc.) based on the corresponding the *marginal* distributions.

Are there any other **(joint) moments**? Yes, a whole multitude of them. And similarly to the univariate case, we can separate them into *simple* (joint) moments $\mathbb{E}(X^n \cdot Y^m)$ and *central* moments: $\mathbb{E}[(X-\mu_X)^n \cdot (Y-\mu_Y)^m]$ where n and m are integers. Thus, for example, $\mathbb{E}(X^2 \cdot Y^3)$ computes the second-third simple moment of X and Y.

▶Covariance◀

The most important of these is the *first-first central* moment, called the **covariance** of X and Y:

$$Cov(X,Y) = \mathbb{E}[(X-\mu_X)\cdot(Y-\mu_Y)] \equiv \mathbb{E}(X \cdot Y) - \mu_X \cdot \mu_Y$$

[the last is the 'computational' formula]. Covariance is obviously a 'symmetric' notion, i.e. $Cov(X,Y) = Cov(Y,X)$. It becomes *zero* when X and Y are *independent* [this is an immediate consequence of what independence implies for products, as we learned already]:

$$X \bowtie Y \Rightarrow Cov(X,Y) = 0$$

Note that this cannot be reversed: zero covariance does *not* necessarily imply independence.

Based on the covariance one can define the

►Correlation Coefficient◄

of X and Y by: $\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$ [this is the Greek letter 'rho']. The absolute value of this coefficient cannot be greater than 1.

Proof: $\mathbb{E}\left\{[(X-\mu_X)+\lambda(Y-\mu_Y)]^2\right\} \geq 0$ for any value of λ [an arbitrary parameter]. Expanded, this implies $Var(X) + 2\lambda Cov(X,Y) + \lambda^2 Var(Y) \geq 0$. The minimum of the left hand side [considering Var and Cov fixed] is at $\lambda = -\frac{Cov(X,Y)}{Var(Y)}$ [by simple differentiation]. Substituting this λ gives: $Var(X) - \frac{Cov(X,Y)^2}{Var(Y)} \geq 0 \Rightarrow \rho_{XY}^2 \leq 1$ □

EXAMPLE: Using one of our previous distributions

		$X=$ 1	2	3	
$Y=$	0	0.1	0	0.3	0.4
	1	0.3	0.1	0.2	0.6
		0.4	0.1	0.5	

we have $\mu_X = 2.1$, $\mu_Y = 0.6$ [done earlier] $Var(X) = 5.3 - 2.1^2 = .89$, $Var(Y) = 0.6 - 0.6^2 = 0.24$, $Cov(X,Y) = 0.3 + 0.2 + 0.6 - 2.1 \times 0.6 = -0.16$ [as likely to be negative as positive] and $\rho_{XY} = \frac{-0.16}{\sqrt{0.89 \times 0.24}} = -0.3462$ ■

►Linear Combination of Random Variables◄

One can simplify a **variance** of a **linear combination of two** RVs [so far we have a formula for $aX+c$ only]. Let's try it: $Var(aX+bY+c) = \mathbb{E}\left[(aX+bY+c)^2\right] - (a\mu_X + b\mu_B + c)^2 = a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(X \cdot Y) - a^2\mu_X^2 - b^2\mu_Y^2 - 2ab\mu_X\mu_Y$ [the c-terms cancel] =

$$a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y)$$

Independence would make the last term *zero*.

This result can be easily **extended** to a linear combination of *any number* of random variables:

$$Var(a_1X_1 + a_2X_2 + ...a_kX_k + c) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + ... + a_k^2 Var(X_k) + 2a_1a_2Cov(X_1,X_2) + 2a_1a_3Cov(X_1,X_3) + ... + 2a_{k-1}a_kCov(X_{k-1},X_k)$$

Mutual *independence* (if present) would make the last row of $\binom{k}{2}$ *covariances disappear* (as they are all equal to zero).

And finally a formula for a **covariance** of one *linear combination* of RVs against another:

$$Cov\left(a_1X_1 + a_2X_2 + ..., b_1Y_1 + b_2Y_2 + ...\right) =$$
$$a_1b_1Cov(X_1,Y_1) + a_1b_2Cov(X_1,Y_2) + a_2b_1Cov(X_2,Y_1) + a_2b_2Cov(X_2,Y_2) + ...$$

[each term from the left hand side against each term on the right - the '*distributive law* of covariance']. Note that in the last formula the X and Y variables don't need to be all distinct; whenever we encounter something like $Cov(U,U)$, we know how to deal with it [by our definition: $Cov(X,X) \equiv Var(X)$].

Correlation coefficient: Using these formulas we can easily prove that $\rho_{aX+c,bY+d} = \frac{Cov(aX+c,bY+d)}{\sigma_{aX+c}\cdot\sigma_{bY+d}} = \frac{a\cdot b\cdot Cov(X,Y)}{|a|\cdot|b|\cdot\sigma_X\cdot\sigma_Y} = \pm\rho_{XY}$ [+ when a and b have the same sign, $-$ when they have opposite signs].

Another important special case results when we independently *sample* the same bivariate X-Y distribution n times, calling the individual results X_1 and Y_1, X_2 and Y_2, ..., X_n and Y_n. These constitute the so called

►Random Independent Sample◄ of *size* n

Note that in this case $X_1 \bowtie X_2$, $Y_1 \bowtie Y_2$, $X_1 \bowtie Y_2$, $Y_1 \bowtie X_2, ...$ but X_1 and Y_1,....remain *dependent*. Obviously then $Var(X_1 + X_2 + ... + X_n) = Var(X_1) + Var(X_2) + ... + Var(X_n) \equiv n\,Var(X)$ and, similarly, $Var(\sum_{i=1}^{n} Y_i) = n\,Var(Y)$. On the other hand $Cov(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i) = Cov(X_1,Y_1) + Cov(X_2,Y_2) + ... + Cov(X_n,Y_n) = n\,Cov(X,Y)$.

All this implies that the **correlation coefficient** between the two totals $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} Y_i$ equals $\frac{n\,Cov(X,Y)}{\sqrt{n\,Var(X)}\cdot\sqrt{n\,Var(Y)}} \equiv \rho_{X,Y}$ (the correlation between *individual* X and Y observations). The *same* is true for the corresponding **sample means** $\frac{\sum_{i=1}^{n} X_i}{n}$ and $\frac{\sum_{i=1}^{n} Y_i}{n}$ (why?).

Moment generating function

Technically, it is **defined** as the following expected value [but of a very *special* type, so to us, MGF is really 'something else']:

$$M_X(t) \equiv \mathbb{E}\left[e^{t\,X}\right]$$

where t is an arbitrary (real) parameter [$M_X(t)$ will be our usual *notation* for a MGF].

The main **purpose** for introducing a MGF is this: when *expanded* in t, it yields:

$$M_X(t) = 1 + t\,\mathbb{E}(X) + \frac{t^2}{2}\mathbb{E}(X^2) + \frac{t^3}{3!}\mathbb{E}(X^3) + ...$$

the individual coefficients of the t-powers being the *simple moments* of the distribution, each divided by the corresponding factorial. Quite often this is the *easiest way* of calculating them! Note that this is *equivalent* to:

$$\mathbb{E}(X^k) = M_X^{(k)}(t=0)$$

or, in words, to get the k^{th} simple moment differentiate the corresponding MGF k times (with respect to t) and set t equal to zero.

EXAMPLE: Let us consider the distribution of the number of rolls when waiting for the first head: $f(i) = (\frac{1}{2})^i$, $i = 1, 2, 3,....$To compute its expected value and standard deviation, we first need $\mathbb{E}(X) = \sum_{i=1}^{\infty} i \times (\frac{1}{2})^i = 1 \times \frac{1}{2} + 2 \times (\frac{1}{2})^2 + 3 \times (\frac{1}{2})^3 + ...$ and $\mathbb{E}(X^2) = \sum_{i=1}^{\infty} i^2 \times (\frac{1}{2})^i$. These infinite sums are neither of the geometric, nor of the exponential type (the only two we know how to handle), so unless we want to spend a lot of time learning about infinite sums (we don't), we have to resort to something else. Let us see how difficult it is to build the corresponding MGF: $M(t) = \mathbf{E}\left[e^{t\,X}\right] = \sum_{i=1}^{\infty} e^{t\,i} \times (\frac{1}{2})^i = \sum_{i=1}^{\infty} \left(\frac{e^t}{2}\right)^i = A + A^2 + A^3 + A^4 + ...$ where $A = \frac{e^t}{2}$. This is a simple geometric sum which has the value of $\frac{A}{1-A} = \dfrac{e^t}{2 - e^t}$. (Displaying its graph would be *meaningless*, its *shape* has nothing to do with the distribution.) From this $M(t)$ we can find the first two simple moments by:

$$M'(t) = \frac{d}{dt}\frac{e^t}{2 - e^t} = \frac{2e^t}{(2 - e^t)^2}\bigg|_{t=0} = 2$$

$$M''(t) = \frac{d}{dt}\frac{2e^t}{(2 - e^t)^2} = \frac{4e^t + 2e^{2t}}{(2 - e^t)^3}\bigg|_{t=0} = 6$$

The mean of the distribution is thus equal to 2 [we expected that: on the average it should take two rolls to get a head − after all, if you keep on rolling, half of the outcomes will be heads], the standard deviation is $\sqrt{6 - 2^2} = \sqrt{2} = 1.414$ ■

▶Important Results◀ concerning MGF

For two *independent* RVs we have: $M_{X+Y}(t) = \mathbb{E}\left[e^{t\,(X+Y)}\right]$ [this follows from the basic definition] $= \mathbb{E}\left[e^{t\,X} \cdot e^{t\,Y}\right] = \mathbb{E}\left[e^{t\,X}\right] \cdot \mathbb{E}\left[e^{t\,Y}\right]$ [due to independence] $= M_X(t) \cdot M_Y(t)$, i.e. the **MGF** of a **sum** of two **independent** RVs is the *product* of their individual MGFs. This result can be extended to *any number* of mutually independent RVs:

$$M_{X+Y+Z}(t) = M_X(t) \cdot M_Y(t) \cdot M_Z(t)$$

etc. This result is more *significant* than it appears because it is very *difficult* to find the distribution of an independent sum directly.

EXAMPLE: Find the distribution of the total number of dots in *three* rolls of a die. We have the distributions for one and two rolls, to 'add' these, we have to first build the joint distribution of the two

	2	3	4	5	6	7	8	9	10	11	12
1	$\frac{1}{216}$	$\frac{2}{216}$	$\frac{3}{216}$	$\frac{4}{216}$	$\frac{5}{216}$	$\frac{6}{216}$	$\frac{5}{216}$	$\frac{4}{216}$	$\frac{3}{216}$	$\frac{2}{216}$	$\frac{1}{216}$
2	$\frac{1}{216}$	$\frac{2}{216}$	$\frac{3}{216}$	$\frac{4}{216}$	$\frac{5}{216}$	$\frac{6}{216}$	$\frac{5}{216}$	$\frac{4}{216}$	$\frac{3}{216}$	$\frac{2}{216}$	$\frac{1}{216}$
3	$\frac{1}{216}$	$\frac{2}{216}$	$\frac{3}{216}$	$\frac{4}{216}$	$\frac{5}{216}$	$\frac{6}{216}$	$\frac{5}{216}$	$\frac{4}{216}$	$\frac{3}{216}$	$\frac{2}{216}$	$\frac{1}{216}$
4	$\frac{1}{216}$	$\frac{2}{216}$	$\frac{3}{216}$	$\frac{4}{216}$	$\frac{5}{216}$	$\frac{6}{216}$	$\frac{5}{216}$	$\frac{4}{216}$	$\frac{3}{216}$	$\frac{2}{216}$	$\frac{1}{216}$
5	$\frac{1}{216}$	$\frac{2}{216}$	$\frac{3}{216}$	$\frac{4}{216}$	$\frac{5}{216}$	$\frac{6}{216}$	$\frac{5}{216}$	$\frac{4}{216}$	$\frac{3}{216}$	$\frac{2}{216}$	$\frac{1}{216}$
6	$\frac{1}{216}$	$\frac{2}{216}$	$\frac{3}{216}$	$\frac{4}{216}$	$\frac{5}{216}$	$\frac{6}{216}$	$\frac{5}{216}$	$\frac{4}{216}$	$\frac{3}{216}$	$\frac{2}{216}$	$\frac{1}{216}$

with the corresponding table for the value of the sum

	2	3	4	5	6	7	8	9	10	11	12
1	3	4	5	6	7	8	9	10	11	12	13
2	4	5	6	7	8	9	10	11	12	13	14
3	5	6	7	8	9	10	11	12	13	14	15
4	6	7	8	9	10	11	12	13	14	15	16
5	7	8	9	10	11	12	13	14	15	16	17
6	8	9	10	11	12	13	14	15	16	17	18

from which we can construct the univariate distribution of the sum by adding probabilities corresponding to the same value [the usual procedure]:

$X+Y$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Prob:	$\frac{1}{216}$	$\frac{3}{216}$	$\frac{6}{216}$	$\frac{10}{216}$	$\frac{15}{216}$	$\frac{21}{216}$	$\frac{25}{216}$	$\frac{27}{216}$	$\frac{27}{216}$	$\frac{25}{216}$	$\frac{21}{216}$	$\frac{15}{216}$	$\frac{10}{216}$	$\frac{6}{216}$	$\frac{3}{216}$	$\frac{1}{216}$

The point being: building the distribution of the sum of two independent RVs is far from trivial. Constructing their MGF (from the individual MGFs) *is* trivial, all it takes is multiplying the two functions [nothing can be easier]. ■

And one more **formula**: $M_{aX+c}(t) = \mathbb{E}\left[e^{(aX+c)\,t}\right]$ [by definition] $= e^{ct} \cdot \mathbb{E}\left[e^{atX}\right]$ [since, with respect to $\mathbb{E}$-averaging, e^{ct} is constant] =

$$e^{ct} \cdot M_X(at)$$

i.e. to build the **MGF** of a **linear** transformation of a single RV one takes the original MGF (of X) (i) replaces t by at (throughout) and (ii) multiplies the result by e^{ct}.

EXAMPLE: If X is the number of rolls till the first head, the MGF of $3X-4$ is $e^{-4t} \cdot \frac{e^{3t}}{2-e^{3t}} = \frac{e^{-t}}{2-e^{3t}}$. ■

Note: Even though a full information about the corresponding distribution is 'encoded' into a MGF, its '**decoding**' (converting **MGF** back into a table of probabilities) is somehow more involved and we will be not discussed here. Instead, we will just build a *'dictionary'* of MGFs of all distributions we encounter, so that eventually we can *recognize* a distribution by its MGF.

Optional: **Inverting MGF** is relatively easy in one important case: When a RV has only *non-negative integers* for its values, the corresponding MGF is actually a function of $z = e^t$ (i.e. t appears only in the e^t combination). Seen as a function of z, $M_X(z)$ is called PROBABILITY GENERATING FUNCTION which, when expanded in z, yields the probabilities of $X = 0$, $X = 1$, $X = 2, ...$ as coefficients of z^0, z^1, $z^2, ...$ (respectively), thus: $M(z) = p_0 + p_1 z + p_2 z^2 + p_3 z^3 + ...$

EXAMPLE: $\frac{e^t}{2-e^t} \equiv \frac{z}{2-z} = \frac{z}{2} \cdot \frac{1}{1-\frac{z}{2}} = \frac{z}{2} + \frac{z^2}{4} + \frac{z^3}{8} + \frac{z^4}{16} + ...$ We have thus recovered probabilities of the $f(i) = \frac{1}{2^i}$, $i = 1, 2, 3, ...$ distribution. ■

Conditional expected value

is, simply put, an expected value computed (via the same 'multiply values by probabilities, then add the results' rule) using the corresponding *conditional* (rather than ordinary) distribution, e.g.

$$\mathbb{E}(X|Y=1)=\sum_i i\times f_{X|Y=1}(i)$$

etc.

EXAMPLE: Using one of our old bivariate distributions

		$X=$ 1	2	3	
$Y=$	0	0.1	0	0.3	0.4
	1	0.3	0.1	0.2	0.6
		0.4	0.1	0.5	

$\mathbb{E}(X|Y=1)$ will be constructed based on the corresponding conditional distribution

$X\|Y=1$	1	2	3
Prob:	$\frac{3}{6}$	$\frac{1}{6}$	$\frac{2}{6}$

by the usual process: $1\times\frac{3}{6}+2\times\frac{1}{6}+3\times\frac{2}{6}=1.83\bar{3}$ (note that this is different from $\mathbb{E}(X)=2.1$ calculated previously).

Similarly $\mathbb{E}(X^2|Y=1)=1^2\times\frac{3}{6}+2^2\times\frac{1}{6}+3^2\times\frac{2}{6}=4.16\bar{6}$.

Based on these two, one can define $Var(X|Y=1)=4.16\bar{6}-1.83\bar{3}^2=0.8056$ [conditional variance].

$M_{X|Y=1}(t)=\frac{3e^t+e^{2t}+2e^{3t}}{6}$ [conditional MGF].

$\mathbb{E}(\frac{1}{X}|Y=1)=\frac{1}{1}\times\frac{3}{6}+\frac{1}{2}\times\frac{1}{6}+\frac{1}{3}\times\frac{2}{6}=0.694\bar{4}$ ■

Infinite expected value

Final **beware:** Not all RVs need to have a (finite) expected value.

EXAMPLE: Consider the following simple game: You bet \$1 on a flip of a coin (say you bet on heads). It you win, you collect your \$2 (\$1 net) and stop. If you lose your continue, doubling your bet. And so on, until you win. We want to compute the expected value of our net win.

Solution: The experiment is the same as flipping a coin until a head appears, with $(\frac{1}{2})^i$ being the probability of needing exactly i flips. The RVs we need are X: how much we bet on the i^{th} flip, Y: how much money have we have betted in total at that point, and Z: how much money we collect when the head appears (in the i^{th} flip), thus:

Simple event:	Prob:	X	Y	Z
H	$\frac{1}{2}$	1	1	2
TH	$\frac{1}{2^2}$	2	3	4
TTH	$\frac{1}{2^3}$	4	7	8
TTTH	$\frac{1}{2^4}$	8	15	16
⋮	⋮	⋮	⋮	⋮
i flips	$\frac{1}{2^i}$	2^{i-1}	2^i-1	2^i
⋮	⋮	⋮	⋮	⋮

Note that our net win is $Z - Y \equiv 1$, i.e. we *always* win \$1 in the end! Is this game fair (equitable)? Of course not, our probability of winning is 100% and the expected win is \$1 (a fair game must have the expected net win equal to 0). The catch is that you can play this game only if you have unlimited resources (nobody does) because the expected value of Y (the money you need to invest in the game before winning your \$1) is *infinite*: $\mathbb{E}(Y) = \sum_{i=1}^{\infty}(2^i - 1) \times (\frac{1}{2})^i = \sum_{i=1}^{\infty} 1 - \sum_{i=1}^{\infty}(\frac{1}{2})^i = \infty - 1 = \infty$. As soon as you put a limit on how much money you can spend (redoing our table accordingly – try it), the game becomes fair. ■

Remember: Some (unusual) RVs have infinite (or indefinite: $\infty - \infty$) expected value (in either case they say that the expected value *does not exist*). Other RVs may have a finite expected value, but their *variance* is infinite. These RVs behave differently from the 'usual' ones, as we will see in later chapters.

Chapter 5 SPECIAL DISCRETE DISTRIBUTIONS

We will now discuss, one by one, those discrete distributions which are most frequently encountered in applications. For each of them, we derive the probability function $f(i)$, the distribution function $F(i)$ [whenever possible], the mean and standard deviation, and the moment generating function $M(t)$. For multivariate distributions, we also like to know the covariance between any two of its RVs.

Univariate distributions

▶Bernoulli◀

Consider an experiment with only two possible outcomes (we call them SUCCESS and FAILURE) which happen with the probability of p and $q \equiv 1 - p$ respectively [examples: flipping a coin, flipping a tack, rolling a die and being concerned only with obtaining a six versus any other number, a team winning or losing a game, drawing a marble from a box with red and blue marbles, shooting against a target to either hit or miss, etc.]

We define a random variable X as the number of successes one gets in one round, or TRIAL, of this experiment. Its distribution is obviously

$X =$	0	1
Prob:	q	p

implying: $\mathbb{E}(X) = p$, $Var(X) = p - p^2 = pq$, $M(t) = q + pe^t$.

▶Binomial◀

Same as before, except now the experiment consists of n *independent* rounds (*trials*) of the Bernoulli type [independence means that the team is not improving as they play more games, the n marbles are selected *with replacement*, etc.]. The sample space consists of all n-letter words build of letters S and F, e.g. $SSFSFSSFFF$ [if n is 10]. We know that there are 2^n of these. They are *not* equally likely, the probability of each is $p^i q^{n-i}$ where i is the number of S's and $n - i$ is the number of F's (due to independence, we just multiply the individual probabilities). We also know that $\binom{n}{i}$ of these words have *exactly* i S's, luckily they all have the same probability $p^i q^{n-i}$. Thus, the probability that our random variable X [the **total number of successes**] will have the value of i is

$$\binom{n}{i} p^i q^{n-i} \qquad (f)$$

This of course is the **probability function** $f(i)$, with $i = 0, 1, 2, \ldots, n-1, n$. Let us verify that these probabilities add up to 1 (as a check): $\sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} = (p+q)^n$ [the binomial expansion 'in reverse'] $= 1^n = 1$ (✓).

The name of the formula used for such a verification usually gives name to the distribution itself [that is why several of our distributions acquire rather puzzling names]. $\mathcal{B}(n,p)$ will be our 'shorthand' for this distribution, n and p are its PARAMETERS.

There are three ways of deriving the **expected value** of X:

1. Using the basic expected-value formula: $\sum_{i=0}^{n} i \times \binom{n}{i} p^i q^{n-i}$. Evaluating this is actually quite tricky (see your textbook if interested) we will not even try it.

2. Note that X can be defined as a sum of n *independent* random variables of the Bernoulli type (number of successes in Trial 1, Trial 2,.... Trial n), say $X_1 + X_2 + ... + X_n$ [$X_1,\ X_2, ..., X_n$ are said to be INDEPENDENT, IDENTICALLY DISTRIBUTED]. Then, as we know, $\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + ... + \mathbb{E}(X_n) = p + p + ... + p =$

$$np \qquad \text{(mean)}$$

3. Using the corresponding MGF which, for the same reason (X being an independent sum of X_1, X_2, ..., X_n) must equal to $(q+pe^t) \times (q+pe^t) \times ..... \times (q+pe^t) = (q+pe^t)^n$. One simple differentiation yields: $n(q+pe^t)^{n-1}pe^t \underset{t=0}{\longrightarrow} np$ [check].

Similarly, the **variance** can be computed either from the basic definition (a rather difficult summation which we choose to bypass), or from $Var(X) = Var(X_1) + Var(X_2) + ... + Var(X_n)$ [valid for *independent* X_is, which ours are] $= pq + pq + ... + pq =$

$$npq \qquad \text{(variance)}$$

or from the moment generating function, thus: $M''_X(t) = n(n-1)(q+pe^t)^{n-2}(pe^t)^2 + n(q+pe^t)^{n-1}pe^t \underset{t=0}{\longrightarrow} n(n-1)p^2 + np = n^2p^2 - np^2 + np$, yielding the value of $\mathbb{E}(X^2)$. Subtracting $\mathbb{E}(X)^2 = n^2p^2$ gives: $Var(X) = np - np^2 = npq$ [check].

Remark: When deriving the *central* moments, it is often more convenient to use $M_{X-\mu}(t) = e^{-\mu t} \cdot M_X(t)$ [the general formula] $= (qe^{-pt} + pe^{qt})^n$ [in this particular case]. Then, obviously, $M''_{X-\mu}(0) = Var(X)$, $M'''_{X-\mu}(0) = \mathbb{E}[(X-\mu)^3]$, etc. Verify that this also results in $Var(X) = npq$. ■

The **moment generation function** was already derived, based on the $X = X_1 + X_2 + ... + X_n$ argument. Let us re-derive it directly from the basic definition of $M_X(t) = \mathbb{E}(e^{Xt}) = \sum_{i=0}^{n} e^{it} \times \binom{n}{i} p^i q^{n-i} = \sum_{i=0}^{n} \binom{n}{i} (pe^t)^i q^{n-i} =$

$$(q+pe^t)^n \qquad \text{(MGF)}$$

[based on the same binomial formula] ✓. Surprisingly [yet typically] it is easier to find the expected value of e^{Xt} than the expected value of X itself (this is one of the big advantages of MGFs).

There is **no** formula for the **distribution function** $F(i)$, which means that the probability of any *range* of values can be computed only by adding the individual

probabilities. For example, if $n = 20$ and $p = \frac{1}{6}$ [rolling a die 20 times, counting the sixes] the probability of getting *at least* 10 [different from *more* than 10, be careful about this] equals $\binom{20}{10}(\frac{1}{6})^{10}(\frac{5}{6})^{10} + \binom{20}{11}(\frac{1}{6})^{11}(\frac{5}{6})^{9} + \binom{20}{12}(\frac{1}{6})^{12}(\frac{5}{6})^{8} + ...+ \binom{20}{0}(\frac{1}{6})^{20}(\frac{5}{6})^{0}$ [must evaluate, one by one, and add] $= 0.05985\%$.

Your **main task** will be first to *recognize* a binomial RV when you see one, and be able to correctly apply the formulas of this section to specific questions.

▶Geometric◀

distribution is based on the same kind of experiment, where the independent Bernoulli-type trials are performed, repeatedly, until the first success appears. This time the random variable (we may as well call it X again, otherwise we would run out of letters much too soon) is the total **number of trials** needed. We already know that the simple events are S, FS, FFS, $FFFS$, ... with the probabilities of p, qp, q^2p, q^3p, ... and the corresponding values of X equal to 1, 2, 3, 4, ... respectively.

The general formula for $\Pr(X = i) \equiv f(i)$ is thus

$$pq^{i-1} \qquad (f)$$

where $i = 1, 2, 3, ...$ To check that these probabilities add up to 1, we proceed as follows: $\sum_{i=1}^{\infty} pq^{i-1} = p(1 + q + q^2 + q^3 + ...) = \frac{p}{1-q} = \frac{p}{p} = 1$. The summation was performed using the *geometric* formula (thus the name of the distribution). The distribution has a single parameter p, and will be referred to as $\mathcal{G}(p)$.

To evaluate $\mathbb{E}(X)$ directly, i.e. by means of $\sum_{i=1}^{\infty} i \times pq^{i-1}$ would be quite difficult [try it if you like, but you must know how to add $1 + 2q + 3q^2 + 4q^3 + ...$ first], so we will use the MGF technique instead. To derive $M(t)$, we need: $\sum_{i=1}^{\infty} e^{it} \times pq^{i-1} = \sum_{i=1}^{\infty} e^t p(e^t q)^{i-1} = pe^t[1 + e^t q + (e^t q)^2 + (e^t q)^3 + ...]$ which is again quite simple to deal with (a geometric series). The answer is $\dfrac{pe^t}{1 - qe^t}$. Differentiating it with respect to t, we get $\frac{pe^t(1-qe^t)-pe^t qe^t}{(1-qe^t)^2} = \frac{pe^t}{(1-qe^t)^2} \xrightarrow[t=0]{} \frac{p}{p^2} =$

$$\frac{1}{p} \qquad \text{(mean)}$$

This is then the expected value of X [on the average it should take $\frac{1}{\frac{1}{6}} = 6$ rolls to get a six; that seems to check, since, in an infinite sequence of trials, one sixth of all outcomes would yield a 6].

Similarly, one more differentiation of $M(t)$ results in $\frac{pe^t(1-qe^t)^2 - 2(1-qe^t)(-qe^t)pe^t}{(1-qe^t)^4} \xrightarrow[t=0]{} \frac{p^3+2p^2q}{p^4} = \frac{1}{p} + 2\frac{1-p}{p^2} = \frac{2}{p^2} - \frac{1}{p}$, which yields $\mathbb{E}(X^2)$. This implies that $Var(X) = \frac{2}{p^2} - \frac{1}{p} - (\frac{1}{p})^2 =$

$$\frac{1}{p}\left(\frac{1}{p} - 1\right) \qquad \text{(variance)}$$

e.g. the **standard deviation** of the number of trials to get the first 6 is $\sqrt{6 \times 5} = 5.477$ [almost as big as the mean itself, implying large variation].

To find the **distribution function** $F(j)$ we first compute $\Pr(X > j) = \sum_{i=j+1}^{\infty} \Pr(X = i) = pq^j + pq^{j+1} + pq^{j+2} + ... = pq^j(1 + q + q^2 + ...) = \frac{pq^j}{1-q} = q^j$ for any $j = 0, 1, 2, 3, ...$ From this

$$F(j) = \Pr(X \leq j) = 1 - q^j$$

easily follows. Thus, for example, the probability that it will take *at least* 10 rolls to get the first 6 [same as *more than* 9] is $\Pr(X > 9) = (\frac{5}{6})^9 = 19.38\%$. The probability that it will take more than 18 rolls is $(\frac{5}{6})^{18} = 3.756\%$ [implying that the geometric distribution has a long TAIL].

▶Negative Binomial◀

distribution is [in spite of its name] a simple extension of the *geometric* (not binomial) distribution. This time the random variable X is the **number of trials** until (and including) the k^{th} success. It can be expressed a sum of k *independent* [a die cannot remember] random variables of the previous, *geometric* type, thus: $X = X_1 + X_2 + ... + X_k$ where X_1 is the number of trials to get the first success, X_2 is the number of trials to get the second success (from that point on), etc. This simplifies getting the **mean and variance** of X to mere multiplication of the 'geometric' answers by k, resulting in $\frac{k}{p}$ and $\frac{k}{p}(\frac{1}{p} - 1)$ respectively. Similarly, the new **MGF** is obtained by raising the old MGF to the power of k: $(\frac{pe^t}{1-qe^t})^k$. The distribution's parameters are k and p, its symbolic name will be $\mathcal{NB}(k, p)$.

To get the individual **probabilities** of the $\Pr(X = i)$ type we must proceed differently: we break the experiment into two uneven parts, namely: (i) the first $i - 1$ rolls, and (ii) the last roll. To get the k^{th} success in this last roll we must first get the first $k - 1$ successes *anywhere* within the first $i - 1$ rolls, followed by a success in the last, i^{th} roll. The former event has a probability of the *binomial* type: $\binom{i-1}{k-1} p^{k-1} q^{i-k}$, the latter one's probability is simply p. To get the overall answer we multiply these two (due to independence), to get

$$\binom{i-1}{k-1} p^k q^{i-k} \equiv \binom{i-1}{i-k} p^k q^{i-k} \qquad (f)$$

where $i = k, k + 1, k + 2, ...$ It helps to display these in an explicit table:

$X =$	k	$k+1$	$k+2$	$k+3$	....
Prob:	p^k	kp^kq	$\binom{k+1}{2}p^kq^2$	$\binom{k+2}{3}p^kq^3$	....

To verify that these probabilities add up to 1 we proceed as follows: $1 \equiv p^k(1 - q)^{-k} = p^k \left[1 - \binom{-k}{1}q + \binom{-k}{2}q^2 - \binom{-k}{3}q^3 + ...\right] = p^k \left[1 + kq + \binom{k+1}{2}q^2 + \binom{k+2}{3}q^3 + ...\right]$. The main part of this proof was the generalized binomial expansion of $(1-q)^{-k}$ [an expression with a *negative* exponent], which explains the name of the distribution. We can now easily deal with questions like: what is the probability that it will take

exactly 5 flips of a coin to get the third head [answer: $\binom{4}{2}(\frac{1}{2})^3(\frac{1}{2})^{5-3} = 18.75\%$] and: what is the probability of requiring exactly 10 rolls to get the second 6 [answer: $\binom{9}{1}(\frac{1}{6})^2(\frac{5}{6})^8 = 5.814\%$].

To be able to answer questions like: 'what is the probability that we will need *more than* 10 rolls to get a second 6', the **distribution function** $F(j)$ would come handy [otherwise we would have to add the individual probabilities of this event, or its complement − neither of which is very practical]. To answer the general question of 'requiring more than j trials to get the k^{th} success' we realize that this is *identical* to 'getting *fewer* than k successes in the first j trials'. And the problem is solved, as we know how to deal with the second question: we just need to add the corresponding *binomial* probabilities [of 0, 1, 2, ...$k-1$ successes in j trials]: $\sum_{i=0}^{k-1}\binom{j}{i}p^iq^{j-i}$. This implies that

$$\Pr(X \le j) \equiv F(j) = 1 - \sum_{i=0}^{k-1}\binom{j}{i}p^iq^{j-i}$$

EXAMPLE: More than 10 rolls of a die will be needed to get the second 6 with the probability of $(\frac{5}{6})^{10} + \binom{10}{1}(\frac{5}{6})^9(\frac{1}{6})^1 = 48.45\%$ [≡ *fewer* than 2 successes in 10 rolls]. ■

►Hypergeometric◄

distribution relates to the following experiment: Suppose there are N objects, K of which have some *special* property, such as being red (marbles), being spades, aces (cards), defective (items of some kind), women (people), etc. [let's call the remaining $N-K$ objects '*ordinary*']. Of these N objects [in total], n [distinct] are *randomly* selected [SAMPLING WITHOUT REPLACEMENT]. Let X be the number of 'special' objects found in the sample. The sample space consists of a list of all possible ways of selecting n objects out of N (order irrelevant). We know that the *total* number of these is $\binom{N}{n}$ and that they are [when the selection is perfectly random] equally likely. We also know (having solved many questions of this type) that $\binom{K}{i} \times \binom{N-K}{n-i}$ of these simple events contains exactly i 'special' objects. Thus

$$\Pr(X = i) \equiv f(i) = \frac{\binom{K}{i} \times \binom{N-K}{n-i}}{\binom{N}{n}}$$

with $\max(0, n-N+K) \le i \le \min(n, K)$ [i cannot be any bigger than either n or K, and it cannot be any smaller than either 0 or $n-(N-K)$; the last restriction corresponds to the situation when the SAMPLE SIZE n is bigger than the total number of 'ordinary' objects]. It is sufficient to remember that, whenever the evaluation of the above formula would lead to *negative* factorials, we are out of range! The formula which verifies that these probabilities add up to 1 is called hypergeometric. Since we never studied this formula, we must skip the corresponding proof [we will simply trust that our derivation of $f(i)$ was correct]. The only bad news is: if we don't know how to do this (hypergeometric) summation, we can derive the corresponding MGF either. Also, there is no 'shortcut' formula for

the cumulative probabilities $F(j)$, which implies that the probability of a *range* of values must be computed by a tedious addition of the individual probabilities.

Note that if we change the experiment slightly to make it a sampling *with replacement* [we select a marble, observe its color, put it back in the box, shake and sample again], X will have the *exact binomial* distribution (with parameters n and $p = \frac{K}{N}$) instead. Naturally, when both N and K are *large* (hundreds or more), it makes little difference whether we sample with or without replacement, and the hypergeometric distribution can be *approximated* (quite accurately) by the somehow simpler *binomial* distribution, using $\frac{K}{N}$ for p.

The **expected value** of X is a bit of a challenge: we don't know how to deal with the direct summation, we could not derive the MGF, could we possibly try the $X = X_1 + X_2 + ... + X_n$ approach [which requires selecting the objects, *without replacement*, one by one, and defining X_1. X_2, ... as the number of 'special' objects obtained in the first, second,.... draw]? The individual X_i's are of the Bernoulli type, with $p = \frac{K}{N}$, but this time they are *not independent* [trouble!?]. Well, let us proceed, anyhow: The expected value of X is easy: $\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + ... + \mathbb{E}(X_n) = \frac{K}{N} + \frac{K}{N} + ... + \frac{K}{N} =$

$$n\frac{K}{N} \qquad \text{(mean)}$$

[an exact analog of the binomial np formula] since we don't have to worry about independence and the X_is are *identically* distributed with the same mean of $\frac{K}{N}$.

To understand the **marginal distribution** of each of the X_i's it helps to visualize the experiment done as follows: as the n sample objects (say, marbles) are drawn, they are placed under individual cups labelled 1, 2, 3, ..., n *without* observing their color! Then X_i is the number of *red* marbles under Cup i, *regardless* of what the other cups contain. This prevents us from confusing the marginal [unconditional] distributions (which we need) from conditional distributions (which would be incorrect).

To establish the **variance** of X, we use: $Var(X) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j) = n \cdot Var(X_1) + 2\binom{n}{2} \cdot Cov(X_1, X_2)$ as all the n variances, and all the $\binom{n}{2}$ covariances, must have the same value due to the SYMMETRY of the experiment (the third cup must have the same probability of containing a red marble as the first cup, etc.). We know from Bernoulli distribution that $Var(X_i) \equiv \frac{K}{N} \cdot \frac{N-K}{N}$ [the pq formula]. To find $Cov(X_1, X_2)$ we first build the joint distribution of X_1 and X_2:

$X_1=$ / $X_2=$	0	1
0	$\frac{N-K}{N} \cdot \frac{N-K-1}{N-1}$	$\frac{K}{N} \cdot \frac{N-K}{N-1}$
1	$\frac{N-K}{N} \cdot \frac{K}{N-1}$	$\frac{K}{N} \cdot \frac{K-1}{N-1}$

from which it easily follows that $\mathbb{E}(X_1 \cdot X_2) = \frac{K(K-1)}{N(N-1)} \Rightarrow Cov(X_1, X_2) = \frac{K(K-1)}{N(N-1)} - \left(\frac{K}{N}\right)^2 = \frac{K^2N - KN - K^2N + K^2}{N^2(N-1)} = -\frac{K(N-K)}{N^2(N-1)}$. This enables us to complete the previous computation: $Var(X) = n\frac{K(N-K)}{N^2} - n(n-1)\frac{K(N-K)}{N^2(N-1)} = n\frac{K(N-K)}{N^2}\left[1 - \frac{n-1}{N-1}\right] =$

$$n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1} \qquad \text{(variance)}$$

which is our final variance formula. Note that it is an analog of the binomial npq, further multiplied by an extra 'correction factor' of $\frac{N-n}{N-1}$ [sampling *without* replacement makes the variance a bit smaller]. When $n = 1$ the two formulas naturally agree, when $n = N$ the variance becomes 0 (as X has then the so called DEGENERATE DISTRIBUTION with only one possible value of X, namely K). These two extreme cases help us remember what the correction factor should be.

EXAMPLE: There are 30 red and 70 blue marbles in a box. If 10 marbles are randomly drawn (*without* replacement), what is the probability that exactly 4 of these are red?

Answer: $\frac{\binom{30}{4}\cdot\binom{70}{6}}{\binom{100}{10}} = 20.76\%$.

If done *with* replacement, the *binomial* formula would have been used instead: $\binom{10}{4} \times 0.3^4 \times 0.7^6 = 20.01\%$, the two **answers** are already quite **similar** [starting with 300 red and 700 blue marbles, and the two answers would be 20.08% and 20.01%, respectively, i.e. practically identical]. ■

►Poisson◄

distribution relates to the following type of experiment: Suppose a man is fishing at a large lake knowing, from past experience, that he will catch a fish *on the average* every 50 minutes (the reciprocal, $\varphi = \frac{1}{50}/\min. \equiv \frac{60}{50} = 1.2$ / hour is the RATE at which the fishes are caught). Let X be the **number of fishes** he catches during the next hour (T minutes in general). We would like to build a good model for the distribution of X.

We may start by assuming that the probability of catching a fish during any 5 minute interval is 0.1 ($= 5\,\min./50\,\min.$) and then use the corresponding binomial distribution: $\binom{12}{i} \times 0.1^i \times 0.9^{12-i}$, $i = 0, 1, 2, ..., 12$. Or we may break one hour into 1 min. intervals, take the probability of catching a fish during any of these subintervals to be $\frac{1}{50}$, and use $\binom{50}{i}(\frac{1}{50})^i(\frac{49}{50})^{50-i}$, $i = 0, 1,...,50$. In general we may divide T into n subintervals, take the probability of catching a fish during any of these to be $\frac{T\varphi}{n}$ and use $\binom{n}{i}(\frac{T\varphi}{n})^i(1-\frac{T\varphi}{n})^{n-i}$, $i = 0, 1, ...n$. Each of these binomial distributions is just an *approximation* to what the correct distribution should look like, since none of them can prevent the possibility of catching *two* or more fishes during a *single* subinterval and thus violating the basic assumption of the binomial distribution (a trial can result in either zero or one success *only*).

The correct description can be reached in the limit of $n \to \infty$ (infinite number of subintervals): $\lim_{n\to\infty}\binom{n}{i}(\frac{T\varphi}{n})^i(1-\frac{T\varphi}{n})^{n-i} = \lim_{n\to\infty}\frac{1}{i!}\frac{n}{n}\frac{n-1}{n}\frac{n-2}{n}....\frac{n-i+1}{n}(T\varphi)^i(1-\frac{T\varphi}{n})^{n-i} = \frac{(T\varphi)^i}{i!}e^{-T\varphi}$ where $i = 0, 1, 2, ...$ Introducing a single parameter λ for $T\varphi$, the formula simplifies to

$$\frac{\lambda^i}{i!}e^{-\lambda} \qquad (f)$$

where $i = 0, 1, 2, ...$ (all non-negative integers). This is the **probability function** $f(i)$ of our new (one-parameter) Poisson distribution [$\mathcal{P}(\lambda)$ for short]. We can easily verify that these probabilities add up to 1 as $\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = 1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + ...$ is the expansion of e^λ.

There is no 'shortcut' formula for $F(j)$, we again must add the individual probabilities to deal with a range of values.

The Poisson distribution provides a good description of: the number of arrivals to a store, library, gas station, etc., the number of accidents at an intersection, the number of phone calls received by on office, during some fixed period of time.

The corresponding **MGF** is $e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}e^{i\,t} = e^{-\lambda}\sum_{i=0}^{\infty}\frac{(\lambda e^t)^i}{i!} = e^{-\lambda}\cdot e^{\lambda e^t} = e^{\lambda(e^t-1)}$. This time it is more convenient to work with the natural logarithm of MGF, let us call it $R_X(t) = \ln\left(M_X(t)\right) =$

$$\lambda(e^t - 1) \qquad (R)$$

One can easily derive that in general $R'(0) = \frac{M'(0)}{M(0)} = \mu_X$ [getting the corresponding expected value] and $R''(0) = \frac{M''(0)M(0)-M'(0)M'(0)}{M(0)^2} = \mathbb{E}(X^2)-\mu_x^2 = Var(X)$ [getting the variance more directly then through $M(t)$]. In our case this results in $R'(t) = \lambda e^t \underset{t=0}{\longrightarrow} \lambda$ [the **mean** of the Poisson distribution] and $R''(t) = \lambda e^t \underset{t=0}{\longrightarrow} \lambda$ [the corresponding **variance**].

One can also show that a **sum** of two **independent** RVs of the Poisson type [with λ_1 and λ_2 as the individual means], has a MGF equal to $e^{\lambda_1(e^t-1)}\cdot e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$, implying that the corresponding distribution is also Poisson, with the mean of $\lambda_1+\lambda_2$. This is a rather *special* property of the Poisson distribution [note that by adding two binomial-type RVs with arbitrary parameters one does *not* get a binomial distribution as a result, similarly adding two geometric-type RVs does not result in any simple distribution unless $p_1 \equiv p_2$, etc.].

Optional: One can extend the previous results concerning $R(t)$ to: $R'''(0) = \mathbb{E}\left[(X-\mu)^3\right]$ [the third central moment] and $R'^v(0) = \mathbb{E}\left[(X-\mu)^4\right] - 3Var(X)^2$ [now it gets a bit more complicated: this is the so called *forth* CUMULANT of the distribution; $R^{(n)}(0)$ yields the n^{th} cumulant]. One can show that a cumulant of a sum of *independent* RVs equals the sum of corresponding cumulants of the individual contributions [we knew this was so for the mean and variance, now we can extend it to the third central moment and higher cumulants]. ■

EXAMPLE: Customers arrive at an *average* rate of 3.7/hour.

1. What is the probability of exactly 1 arrival during the next 15 min.

   Solution: $\lambda = T\theta = \frac{1}{4}\times 3.7 = 0.925$ [make sure to use the *same units* of time for both T and θ].

   Answer: $e^{-0.925}\times\frac{0.925}{1!} = 36.68\%$.

2. What is the probability of *at least* 4 arrival during the next *30 min.*

   Answer [through the complement, λ will of course double]: $1-\Pr(X=0)-\Pr(X=1)-\Pr(X=2)-\Pr(X=3) = 1-(1+1.85+\frac{1.85^2}{2}+\frac{1.85^3}{6})\,e^{-1.85} = 11.69\%$.

3. If the store opens at 8:00 what is the probability that their *second* customer arrives between 8:20 and 8:45 ?

   Solution: Define A: at least two arrivals by 8:45, and B: at least two arrivals by 8:20. We need $\Pr(A \cap \bar{B}) = \Pr(A) - \Pr(B)$ since $B \subset A$.

   Answer: $\left[1 - e^{-\lambda_1}(1+\lambda_1)\right] - \left[1 - e^{-\lambda_2}(1+\lambda_2)\right]$ where $\lambda_1 = \frac{3}{4} \times 3.7 = 2.775$ and $\lambda_2 = \frac{1}{3} \times 3.7 = 1.23333$, i.e. $e^{-1.23333} \times 2.23333 - e^{-2.775} \times 3.775 = 41.52\%$. ■

Multivariate distributions

►Multinomial◄

distribution is an extension of the binomial distribution, in which each *trial* can result in 3 (or more) possible outcomes (not just S and F). The trials are still repeated, independently, n times; this time we need three RVs X, Y and Z, which count the total number of outcomes of the first, second and third type, respectively [we will develop our formulas assuming that there are 3 possibilities, the extension to 4 or more possible outcomes is then quite obvious].

Examples: A team playing a series of n games, each of which they either win, lose or tie (3 possibilities). A die is rolled n times, keeping track of the ones, twos, ..., sixes (6 possibilities).

The sample space (in the case of three possible outcomes) consists of 3^n simple events [all possible n-letter words build out of thee letters, w, ℓ and t say]. If X counts the wins, Y the losses, and Z the ties, $\Pr(X = i \cap Y = j \cap Z = k)$ is the sum of probabilities of all simple events consisting of exactly i w's, j ℓ's and k t's, we know that there are $\binom{n}{i,j,k}$ of these, each of them [luckily] having the same probability of $p_X^i \, p_Y^j \, p_Z^k$ where p_X, p_Y and p_Z are the probabilities of a win, a loss and a tie, respectively, in a single trial (game) [obviously $p_X + p_Y + p_Z \equiv 1$]. The **joint probability function** is thus

$$\Pr(X = i \cap Y = j \cap Z = k) = \binom{n}{i,j,k} p_X^i \, p_Y^j \, p_Z^k$$

for any non-negative integer values of i, j, k which add up to n. This formula can be easily extended to the case of 4 or more possible outcomes.

The **marginal distribution** of X is obviously *binomial* (with n and $p \equiv p_X$ being the two parameters), and similarly for Y and Z [this can be verified algebraically by summing the above formula over j and k, try it]. This yields the individual **mean**s and **variance**s [$\mathbb{E}(X) = np_X$ and $Var(X) = np_X \cdot (1 - p_X)$, etc.].

EXAMPLES:

1. A team plays a series of 10 games. The probability of winning a game is 0.40, losing a game: 0.55, and tying a game: 0.05. What is the probability of finishing with 5 wins, 4 losses and 1 tie?

   Answer: $\binom{10}{5,4,1} \times 0.4^5 \times 0.55^4 \times 0.05 = 5.90\%$.

Suplimentary: What is the probability that they win the series (more wins than losses)?

Answer: $\Pr(X > 5)$ [*binomial* probabilities] + $\Pr(X = 5) - \Pr(X = 5 \cap Y = 5) + \Pr(X = 4) - \Pr(X = 4 \cap Y \geq 4) + \Pr(X = 3 \cap Y < 3) + \Pr(X = 2 \cap Y < 2) + \Pr(X = 1 \cap Y = 0) = 23.94\%$.

2. Roll a die 18 times, what is the probability of getting 3 ones, twos, ..., sixes [this should be the most likely outcome]?

   Answer: $\binom{18}{3,3,3,3,3,3}(\frac{1}{6})^{18} = 0.135\%$ [why is it so small?]. ■

An important issue now is to find the **covariance** between X and Y (or any other two RVs of a multinomial distribution). First we break each X and Y into their individual 'components': $X = X_1 + X_2 + ... + X_n$ and $Y = Y_1 + Y_2 + ... + Y_n$ where X_1, X_2, ... is the number of wins in Game 1, Game 2, ... [each obviously of Bernoulli type], and similarly Y_1, $Y_2, ...$ is the number of losses. We know that $Cov(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, Y_j) = \sum_{i=1}^{n} Cov(X_i, Y_i) + \sum_{i \neq j}^{n} Cov(X_i, Y_j)$ [the first term corresponds to X_i, Y_i from the *same* game, the second term corresponds to X_i, Y_j from two different games]. We are assuming the games to be played *independently* of each other, thus X_i and Y_j $(i \neq j)$ are independent RVs, and their covariance must equal to zero. Furthermore, due to the symmetry of the experiment, $Cov(X_1, Y_1) = Cov(X_2, Y_2) = ... = Cov(X_n, Y_n)$. Thus we need to know $Cov(X_1, Y_1)$ only, to complete the exercise. First we build a table of the required bivariate probabilities:

$Y_1 = \backslash X_1 =$	0	1
0	p_Z	p_X
1	p_Y	0

$\Rightarrow Cov(X_1, Y_1) = \mathbb{E}(X_1 \cdot Y_1) - \mathbb{E}(X_1) \cdot \mathbb{E}(Y_1) = 0 - p_X \cdot p_Y = -p_X p_Y \Rightarrow$

$$Cov(X, Y) = -np_X\, p_Y$$

Similarly, $Cov(X, Z) = -np_X\, p_Z$ etc. Note that our derivation was fully general and did not assume the case of three possible outcomes (with the understanding that $p_Z \equiv 1 - p_X - p_Y$).

EXAMPLES:

1. Referring to the previous Example 1, what is the covariance between the number of wins and the number of losses (in a 10 game series)?

   Answer: $-10 \times 0.40 \times 0.55 = -2.2$.

2. Rolling a die 18 times, what is the covariance between the number of 3's and the number of 6's obtained?

   Answer: $-18 \times \frac{1}{6} \times \frac{1}{6} = -0.5$.

3. 10 dice are rolled and we are paid \$5 for each six, but have to pay \$6 for each one. What is the expected value and the standard deviation of our net win?

   Solution: introduce X for the (total) number of 6's and Y for the number of 1's, our net win is $5X - 6Y$. Its expected value is $5 \times 10 \times \frac{1}{6} - 6 \times 10 \times \frac{1}{6} = -1.\bar{6}$

[both X and Y are binomial, we use the np formula to get their means], its variance equals: $5^2 \times 10 \times \frac{1}{6} \times \frac{5}{6} + (-6)^2 \times 10 \times \frac{1}{6} \times \frac{5}{6} + 2 \times 5 \times (-6) \times (-10) \times \frac{1}{6} \times \frac{1}{6} = 101.3\bar{8}$ [using our old formula for $Var(aX + bY)$, the npq formula for the two binomial variances, and the latest $-np_Xp_Y$ formula for the covariance].

Answer: -1.667 ± 10.07 [$=\sqrt{101.3\bar{8}}$] dollars (we normally express the mean and the corresponding standard deviation as $\mu \pm \sigma$; note that a negative mean implies that, *on the average*, we *lose* \$1.667; the big standard deviation implies that in a single game we may occasionally win \$20 or more, if we are lucky).

4. A die is rolled 18 times, U is the number of 'small' outcomes (meaning ≤ 3), V is the number of even outcomes (2, 4 and 6). Find $Cov(U, V)$.

   Solution: We have to realize that, because of the '**overlap**' between U and V (2 will contribute to both U and V), our basic covariance formula no longer applies. We can easily extend it to cover the 'overlap' case as follows: We define T as the RV which counts the overlap outcomes, $U_0 = U - T$ and $V_0 = V - T$ [note that U_0, V_0 and T are now of the multinomial, non-overlapping type]. Then $Cov(U, V) = Cov(U_0 + T, V_0 + T) = Cov(U_0, V_0) + Cov(U_0, T) + Cov(T, V_0) + Var(T) = -n(p_U - p_T)(p_V - p_T) - n(p_U - p_T)p_T - np_T(p_V - p_T) + np_T(1 - p_T) = ... =$

$$-n(p_Up_V - p_T)$$

   (a) Answer: $-18 \times (\frac{1}{2} \times \frac{1}{2} - \frac{1}{6}) = -1.5$ ■

►Multivariate Hypergeometric◄

distribution is again a simple extension of the hypergeometric distribution to the case of having thee (or more) types of objects (rather than just 'special' and 'ordinary'), e.g. red, blue and green marbles or hearts, diamonds, spades and clubs, etc. We now assume that the total number of objects of each type is K_1, K_2 and K_3 (we again develop the formulas for three types only) where $K_1 + K_2 + K_3 = N$. The sample space will still consist of $\binom{N}{n}$ possible selections of n of these [unordered, without duplication] which are all equally likely. We also know that $\binom{K_1}{i} \times \binom{K_2}{j} \times \binom{K_3}{k}$ of these will contain exactly i objects of Type 1, j objects of Type 2 and k objects of Type 3. Thus, the **joint probability function** is

$$\Pr(X = i \cap Y = j \cap Z = k) = \frac{\binom{K_1}{i}\binom{K_2}{j}\binom{K_3}{k}}{\binom{N}{n}}$$

where X, Y and Z count the number of objects of Type 1, 2 and 3, respectively, in the sample. Naturally, $i + j + k = n$. Otherwise, i, j and k can be any non-negative integers for which the above expression is meaningful (i.e. no negative factorials).

The **marginal distribution** of X (and Y, and Z) is *univariate* hypergeometric (of the old kind) with obvious parameters. Thus, the individual means and variances follow from our old formulas.

The only joint quantity we usually need is the **covariance** between X and Y. We again break X into its individual components $X_1 + X_2 + ... + X_n$ and similarly $Y = Y_1 + Y_2 + ... + Y_n$ [visualize placing the drawn marbles under cups, labelled 1, 2, ...n, with X_1 and Y_1 being the number of red and blue marbles, respectively, under Cup 1, etc.]. This implies: $Cov(X,Y) = \sum_{i=1}^{n} Cov(X_i, Y_i) + \sum_{i \neq j}^{n} Cov(X_i, Y_j) = n \times Cov(X_1, Y_1) + n(n-1) \times Cov(X_1, Y_2)$ since all covariances of the $Cov(X_i, Y_i)$ type (same cup) must equal to each other, similarly, all $Cov(X_i, Y_j)$ with $i \neq j$ (different cups) are identical. Now we need the joint distribution of X_1 and Y_1:

$Y_1 =$ \ $X_1 =$	0	1
0	$\frac{N-K_1-K_2}{N}$	$\frac{K_1}{N}$
1	$\frac{K_2}{N}$	0

$\Rightarrow Cov(X_1, Y_1) = 0 - \frac{K_1}{N} \cdot \frac{K_2}{N}$, and of

$Y_2 =$ \ $X_1 =$	0	1
0	rest	$\frac{K_1}{N} \cdot \frac{N-1-K_2}{N-1}$
1	$\frac{K_2}{N} \cdot \frac{N-1-K_1}{N-1}$	$\frac{K_1}{N} \cdot \frac{K_2}{N-1}$

$\Rightarrow Cov(X_1, Y_2) = \frac{K_1 K_2}{N(N-1)} - \frac{K_1}{N} \cdot \frac{K_2}{N} = \frac{K_1 K_2}{N^2 (N-1)}$. Putting it together yields $Cov(X,Y) = -n\frac{K_1 K_2}{N^2} + n(n-1)\frac{K_1 K_2}{N^2 (N-1)} =$

$$-n \frac{K_1}{N} \frac{K_2}{N} \frac{N-n}{N-1}$$

Note that this is an analog of the multinomial covariance $[-np_X p_Y]$, further multiplied by the correction factor of the old $Var(X)$ formula.

One can easily **extend** the covariance formula to the case of 'overlapping' RVs U and V [e.g. counting spades and aces, respectively]:

$$Cov(U,V) = -n \left(\frac{K_1}{N} \frac{K_2}{N} - \frac{K_{12}}{N} \right) \frac{N-n}{N-1}$$

where K_1 and K_2 is the total (including the overlap) number of objects contributing to U and V respectively [spades and aces] K_{12} is the number of objects which contribute to *both* U and V [the 'overlap': ace of spades]. The proof would pretty much duplicate the multinomial case (Example 4).

EXAMPLES:

1. Pay \$15 to play the following game: 5 cards are dealt from the ordinary deck of 52 and you get paid \$20 for each ace, \$10 for each king and \$5 for each queen. Find the expected value of 'net win' (loss if negative) and its standard deviation.

   Solution: we introduce X, Y and Z for the number of aces, kings and queens, respectively, found in the dealt five-card hand. The net win is the following RV: $W = 20X + 10Y + 5Z - 15$. Its expected value is $20\mathbb{E}(X) + 10\mathbb{E}(Y) + 5\mathbb{E}(Z) - 15 = 20 \times 5 \times \frac{4}{52} + 10 \times 5 \times \frac{4}{52} + 5 \times 5 \times \frac{4}{52} - 15 = -1.538$ [with the help of the $n\frac{K}{N}$ formula], its variance equals $20^2 Var(X) + 10^2 Var(Y) + 5^2 Var(Z) + 2 \times 20 \times 10 \times Cov(X,Y) + 2 \times 20 \times 5 \times Cov(X,Z) + 2 \times 10 \times 5 \times Cov(Y,Z) = [5 \times (400 + 100 + 25) \times \frac{1}{13} \times \frac{12}{13} - 5 \times (400 + 200 + 100) \times \frac{1}{13} \times \frac{1}{13}] \times \frac{47}{51} = 152.69$ [with the help of the $n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$ and $-n\frac{K_1}{N}\frac{K_2}{N}\frac{N-n}{N-1}$ formulas].

   Answer: -1.538 ± 12.357 dollars.

2. Five cards are dealt; we get \$1 for each spade and \$2 for each diamond, but we have to pay \$10 for each ace. Find the expected value and standard deviation of the net win.

   Solution: introduce X, Y and U for the number of spades, diamonds and aces. $W = X + 2Y - 10U \Rightarrow \mathbb{E}(W) = 5 \times \frac{13}{52} + 2 \times 5 \times \frac{13}{52} - 10 \times 5 \times \frac{4}{52} = -0.09614$ and $Var(W) = 5 \times [\frac{1}{4} \times \frac{3}{4} + 2^2 \times \frac{1}{4} \times \frac{3}{4} + (-10)^2 \times \frac{1}{13} \times \frac{12}{13} - 2 \times 2 \times \frac{1}{4} \times \frac{1}{4}] \times \frac{47}{51} = 35.886$ [note that $Cov(X, U) = Cov(Y, U) = 0$ as the 'overlap' formula has a $\frac{1}{4} \cdot \frac{1}{13} - \frac{1}{52} \equiv 0$ factor].

   Answer: -0.0961 ± 5.9905. ■

Comprehensive examples

1. A die is rolled 5 times and we are paid \$2 for each dot obtained; then a coin is flipped 10 times and we have to pay \$7 for each head shown. Find the expected value and standard deviation of the game's net win (overall).

   Solution: Let X_1, X_2, ...X_5 represent the number of dots obtained in each roll, and Y be the total number of heads shown (a binomial RV with $n = 10$ and $p = \frac{1}{2}$). Then $W = 2(X_1 + X_2 + ... + X_5) - 7Y \Rightarrow \mathbb{E}(W) = 2(3.5 + 3.5 + ... + 3.5) + 7 \times 10 \times \frac{1}{2} = 0$ [a FAIR game] and $Var(W) = 2^2 \times (\frac{35}{12} + \frac{35}{12} + ... + \frac{35}{12}) + 7^2 \times 10 \times \frac{1}{2} \times \frac{1}{2} = 180.8\bar{3}$ [note that all the X's and Y are *independent* of each other $\Rightarrow$ zero covariance].

   Answer: 0 ± 13.45 dollars.

2. Pay \$35, then roll a die until the 3^{rd} six is obtained and be paid \$2 for each roll. Find μ_W and σ_W, where W is your net win.

   Solution: This time we introduce only X: the number of rolls to get the 3^{rd} six [it has the negative binomial distribution with $k = 3$ and $p = \frac{1}{6}$]. Obviously $W = 2X - 35 \Rightarrow \mathbb{E}(W) = 2 \times \frac{3}{\frac{1}{6}} - 35 = 1$ and $Var(W) = 2^2 \times 3 \times 6 \times 5$.

   Answer: 1 ± 18.97 dollars.

   Supplementary: What is the expected value and standard deviation of the *total* net win after 15 rounds of this game?

   Solution: The games are obviously played independently of each other, therefore $\mathbb{E}(W_1 + W_2 + ... + W_{15}) = 15\mu_W$ and $Var(W_1 + W_2 + ... + W_{15}) = 15\sigma_W^2 \Leftrightarrow \sigma_{W_1+W_2+...+W_{15}} = \sqrt{15}\sigma_W$.

   Answer: 15 ± 73.485 dollars. ■

Sampling from a distribution – Central Limit Theorem

(In your textbook this topic is discussed in Sections 8.1 and 8.2). By now we know what a random variable is, how to define one (based on a specific random experiment), how to give it a name (X say) and find its distribution. Performing the actual experiment will give us a single *random* value of X, repeating this independently n times will give us the so called

►Random Independent Sample◄

(we shorten this to RIS) of SIZE n. The individual values are called X_1, X_2, ...X_n, they are INDEPENDENT, IDENTICALLY DISTRIBUTED (IID) random variables (visualize them as the *would-be values*, before the experiment is actually performed].

The **sample mean** is (unlike the old 'means' which were constant parameters) a *random variable,* defined by:

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

Its **expected value** equals: $\mathbb{E}(\bar{X}) = \frac{1}{n}\mathbb{E}(X_1) + \frac{1}{n}\mathbb{E}(X_2) + ... + \frac{1}{n}\mathbb{E}(X_n) = \frac{1}{n}\mu + \frac{1}{n}\mu + ... + \frac{1}{n}\mu = \mu$, where μ is the expected value of the distribution (in this sampling context sometimes also called 'POPULATION') from which the sample is taken [not surprisingly, $\bar{X}$ is often used as an ESTIMATOR of μ when its value is uncertain].

Similarly $Var(\bar{X}) = (\frac{1}{n})^2 Var(X_1) + (\frac{1}{n})^2 Var(X_2) + ... + (\frac{1}{n})^2 Var(X_n) = (\frac{1}{n})^2\sigma^2 + (\frac{1}{n})^2\sigma^2 + ... + (\frac{1}{n})^2\sigma^2 = \frac{\sigma^2}{n}$ where σ is the standard deviation of the original distribution. This implies that

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

[the standard deviation of $\bar{X}$ is $\sqrt{n}$ smaller than σ; sometimes it is also called the **standard error** of $\bar{X}$]. Note that the standard error tends to zero as the sample size n increases.

So now we know how the mean and standard deviation of $\bar{X}$ relate to the mean and standard deviation of the distribution from which we are sampling. How about the **shape** of the $\bar{X}$-distribution, how does it relate to the shape of the sampling distribution? The surprising answer is: It doesn't (for n more than a handful), instead, the distribution of $\bar{X}$ has always *the same* regular shape (yet to be discovered by us), common to all distributions, from which we may sample! More about this later.

We already mentioned that $\bar{X}$ is frequently used to estimate the value of μ [the mean of the sampling distribution/population]. Sometimes we also need to estimate the value of the distribution's variance σ^2; this is done by the so called **sample variance** (also a random variable!) defined by:

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

[taking its square root, one gets s, the so called SAMPLE STANDARD DEVIATION]. Note that the numerator is the sum of squares of individual deviations from the *sample* mean; the definition intentionally avoids using the distribution mean μ, as its value is usually unknown.

To find the **expected value** of s^2, we simplify its numerator first: $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2 = \sum_{i=1}^{n}(X_i - \mu)^2 + n \cdot (\bar{X} - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu)$ [note that $\bar{X} - \mu$, being free of i, is considered constant by the summation] $\Rightarrow \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \sum_{i=1}^{n} Var(X_i) + n \cdot Var(\bar{X}) - 2\sum_{i=1}^{n} Cov(\bar{X}, X_i) = n\sigma^2 + n \cdot \frac{\sigma^2}{n} - 2n \cdot \frac{\sigma^2}{n} = \sigma^2(n-1)$, since $Cov(\bar{X}, X_1) = \frac{1}{n}\sum_{i=1}^{n} Cov(X_i, X_1) =$

$\frac{1}{n}Cov(X_1, X_1) + 0 = \frac{1}{n}Var(X_1) = \frac{\sigma^2}{n}$, and $Cov(\bar{X}, X_2)$, $Cov(\bar{X}, X_3)$, ... must have the same value. This implies that

$$\mathbb{E}(s^2) = \frac{\sigma^2 (n-1)}{n-1} = \sigma^2$$

Thus, s^2 is a so called UNBIASED ESTIMATOR of the distribution's variance σ^2 [meaning it has the correct expected value].

Does this imply that $s \equiv \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$ (the **sample standard deviation**) has the expected value of σ? The answer is 'no', s is a (slightly) biased estimator of the population's standard deviation [the exact value of the bias depends on the shape of the corresponding distribution].

This s is useful when **estimating** the value of the population mean μ. We know that $\bar{X}$ is the unbiased estimator of μ, having the standard deviation (error) of $\frac{\sigma}{\sqrt{n}}$. We would like to express this as $\mu \approx \bar{X} \pm \frac{\sigma}{\sqrt{n}}$ [the so called CONFIDENCE INTERVAL for estimating μ] but since we ordinarily don't know the exact value of σ either, we have to substitute the next best thing, namely its estimator s, thus: $\mu \approx \bar{X} \pm \frac{s}{\sqrt{n}}$. In later chapters we investigate these issues in more detail.

Proving Central Limit Theorem

Let us now return to the main topic of this section, which is investigating the distribution of $\bar{X}$. We already know the mean and standard deviation of this distribution are μ and $\frac{\sigma}{\sqrt{n}}$ respectively, now we would like to establish its ASYMPTOTIC (i.e. large-n) **shape**. This is, in a sense, trivial: since $\frac{\sigma}{\sqrt{n}} \underset{n\to\infty}{\longrightarrow} 0$, we get in the $n \to \infty$ limit a DEGENERATE (single-valued, with zero variance) distribution, with all probability concentrated at μ.

We can prevent this distribution from shrinking to a zero width by **standardizing** $\bar{X}$ first, i.e. defining a new RV

$$Z \equiv \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

and investigating its asymptotic ($n \to \infty$) distribution instead [the new random variable has the mean of 0 and the standard deviation of 1, thus its shape cannot 'disappear' on us].

We do this by constructing the **MGF** of Z and finding its $n \to \infty$ limit. Since $Z = \frac{\frac{\sum_{i=1}^{n}(X_i - \mu)}{n}}{\frac{\sigma}{\sqrt{n}}} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma\sqrt{n}}\right)$ [a sum of independent, identically distributed RVs], its MGF is the MGF of $\frac{X_1 - \mu}{\sigma\sqrt{n}} \equiv Y$, raised to the power of n.

We know that $M_Y(t) = 1 + \mathbb{E}(Y) \cdot t + \mathbb{E}(Y^2) \cdot \frac{t^2}{2} + \mathbb{E}(Y^3) \cdot \frac{t^3}{3!} + ... = 1 + \frac{t^2}{2n} + \frac{ct^3}{6n^{3/2}} + \frac{dt^4}{24n^2} + ...$ where c, d,... is the skewness, kurtosis, ... of the original distribution. Raising $M_Y(t)$ to the power of n and taking the $n \to \infty$ limit results in $e^{\frac{t^2}{2}}$ *regardless* of the values of c, d, ... (since they are divided by higher-than-one power of n). Thus, we get a rather unexpected result: the distribution of Z has (for large n) the same symmetric shape (described by the above MGF limit), not in

the least affected by the shape of the original distribution (form which the sample is taken).

What exactly is the common shape of the distribution of all these Z's, regardless of what sampling distribution they came from? Well, so far we have not yet seen a random variable whose MGF would equal to $e^{\frac{t^2}{2}}$, we can only hope that we will 'discover' it soon, as its distribution will supply the answer (some of you may already know that it is the so called **normal** or Gaussian distribution which we study in the next chapter). $\square$

Part II

CONTINUOUS DISTRIBUTIONS

Chapter 6 CONTINUOUS RANDOM VARIABLES

We are now returning to Section 3.3 of your textbook. First we introduce the concept of

Univariate probability density function

As we already know, for continuous-type RVs $\Pr(X = x)$ [where x is some specific number – we used to call it i in the discrete case] is always equal to zero and thus of no help to us. Yet the probability of X falling into any interval, such as $\Pr(x \le X < x + \varepsilon)$ [where $\varepsilon > 0$] is nonzero. We can now take the *ratio* of this *probability* to the *length* of the interval, making this length go down to zero:

$$\lim_{\varepsilon \to 0} \frac{\Pr(x \le X < x + \varepsilon)}{\varepsilon} \equiv f(x)$$

getting a *non-negative* function of x which is called **probability density function** (pdf) of the random variable X (the same $f(x)$ notation was used earlier for the *probability function* of the discrete case – luckily, these two can never be confused). Its argument will normally be a small letter corresponding to the variable's name, due to this, there will be less need for my old notational extension of the $f_X(x)$ type.

Based on such an $f(x)$, the **probability of** X falling into *any* **interval** (a, b) say can be computed by integrating $f(x)$ from a to b:

$$\Pr(a < X < b) = \int_a^b f(x)\, dx$$

This immediately implies that $\int_{-\infty}^{\infty} f(x)\, dx = 1$ [the *total* probability of X having *any* value].

Note that:

- *Probability* is now directly *represented by* the corresponding *area* [between the x-axis and $f(x)$].
- $\Pr(a \le X \le b) = \Pr(a \le X < b) = \Pr(a < X \le b) = \Pr(a < X < b)$ as, for *continuous* RVs one extra value (included or excluded) does not make any difference.
- The individual values of $f(x)$ do *not* directly represent a probability of any event, thus some of these values *may*, quite legitimately, *exceed* 1. ■

An important part of a definition of $f(x)$ is its **range** (the interval of permissible values). As we will see in our examples, some RVs allow any real value as an outcome, some are limited to a finite range, say (L, H) beyond which $f(x)$ drops down to 0.

usually called $F(x)$ returns

$$\Pr(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

which has the following, rather obvious properties:

(i) it is flatly equal to zero until L (the lowest possible value) is reached

(ii) it increases, between the x-values of L and H, until 1 is reached (at H)

(iii) stays equal to 1 for all x beyond H [when $H = \infty$, $F(x)$ reaches 1 in the $x \to \infty$ limit]. ■

The distribution function is very convenient for computing probabilities (of an interval), as $\Pr(a < X < b) = F(b) - F(a)$ [we no longer need to integrate].

Sometimes we have to deduce $f(x)$ from $F(x)$, which is of course quite trivial: $f(x) = \frac{dF(x)}{dx}$.

EXAMPLES:

1. Consider our old example of a spinning wheel with a pointer where X is the final value of the resulting angle (measured from some fixed direction), in radians. We know that all angles $[0, 2\pi)$ must be equally likely, thus $f(x)$ must be, in this range, a constant which, when integrated between 0 and 2π, yields 1 [and is equal to zero otherwise]:

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2\pi} & 0 < x \leq 2\pi \\ 0 & x > 2\pi \end{cases}$$

We will normally simplify this to: $f(x) = \frac{1}{2\pi}$ for $0 < x \leq 2\pi$, with the understanding of 'zero *otherwise*'. Note that this $f(x)$ is *not* continuous. The corresponding distribution function is:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{2\pi} & 0 < x \leq 2\pi \\ 1 & x > 2\pi \end{cases}$$

[note that each $F(x)$ must be continuous $\Rightarrow$ adjacent 'pieces' of the function must agree at the common boundary e.g., in this case, $\frac{x}{2\pi} \underset{x=0}{\longrightarrow} 0$ and $\frac{x}{2\pi} \underset{x=2\pi}{\longrightarrow} 1$]. Again, we will shorten this to $F(x) = \frac{1}{2\pi}$ for $0 < x \leq 2\pi$ with the *understanding* that $F(x) \equiv 0$ for $x < 0$ (the smallest possible value), and $F(x) \equiv 1$ for $x > 2\pi$ (the largest possible value).

2. Consider the following pdf: $f(x) = e^{-x}$ for $x > 0$ [zero otherwise – this goes without saying]. Since $f(x) \geq 0$ and $\int_0^\infty e^{-x}\,dx = 1$, this pdf is certainly 'legitimate' [later on we will call the distribution of this type 'exponential']. The corresponding distribution function is $F(x) = \int_0^x e^{-u}\,du = [-e^{-u}]_{u=0}^x = 1 - e^{-x}$ for $x > 0$ (note that this approaches 1 as $x \to \infty$).

3. The non-zero part of $f(x)$ is sometimes itself defined in a 'piece-wise' manner, e.g.:
$f(x) = \begin{cases} \frac{2}{3}(1+x) & -1 \leq x < 0 \\ \frac{4}{3}(1-x) & 0 \leq x < 1 \end{cases}$ [zero otherwise – the last reminder]. Integration of this function (to verify that its total area – probability – equals to 1) must be done a 'piece-wise' manner as well: $\int_{-1}^1 f(x)\,dx = \int_{-1}^0 \frac{2}{3}(1+x)\,dx + \int_0^1 \frac{4}{3}(1-x)\,dx = \frac{2}{3}[x - \frac{x^2}{2}]_{x=-1}^0 + \frac{4}{3}[x - \frac{x^2}{2}]_{x=0}^1 = \frac{1}{3} + \frac{2}{3} = 1$ (check).

   Building the distribution function must be done even more carefully [we display it over the full range of real values, not using our simplifying convention in this case]: $F(x) = \begin{cases} 0 & x \leq -1 \\ \frac{1}{3}(1+x)^2 & -1 < x \leq 0 \\ 1 - \frac{2}{3}(1-x)^2 & 0 < x \leq 1 \\ 1 & x > 1 \end{cases}$ [note that each of the individual results is constructed as the indefinite integral of the corresponding expression for $f(x)$, *plus* a constant designed to match the function's values at each boundary – usually to the *previous* 'piece']. Also note that our $F(x)$, in spite of consisting of several 'pieces' (expressions), is a *single* function [that goes for $f(x)$, too].

   Based on this $F(x)$ we can easily answer questions of the type: $\Pr(-\frac{1}{2} < X < \frac{1}{2}) = F(\frac{1}{2}) - F(-\frac{1}{2}) = 1 - \frac{2}{3}(\frac{1}{2})^2 - \frac{1}{3}(\frac{1}{2})^2 = 75\%$ [just make sure to select the appropriate expression for each evaluation]. ■

Bivariate (multivariate) pdf

Now we extend the concept of 'probability density function' to two and more RVs of the continuous type.

►Joint Probability Density Function◄

of X and Y will be, by definition,

$$f(x,y) = \lim_{\substack{\varepsilon \to 0 \\ \delta \to 0}} \frac{\Pr(x \leq X < x + \varepsilon \cap y \leq Y < y + \delta)}{\varepsilon \cdot \delta}$$

i.e. the *probability* of the X, Y pair 'falling' (i.e. having their values) inside a rectangle with the point (x, y) at one of its corners, *divided* by the *area* of this rectangle, in the limit of this rectangle shrinking to the point itself [instead of a

rectangle, we could have used any 2-D region, with (x, y) located within the region − not necessarily on its boundary].

Automatically, $f(x, y) \geq 0$. The bivariate pdf must also 'add up' (integrate) to 1, i.e.

$$\iint_{\mathcal{R}} f(x, y)\, dxdy \equiv 1$$

where $\mathcal{R}$ is the two-dimensional **region of definition** [sometimes we visualize it as a 'target'] of the (X, Y)-distribution, i.e. the set of all *possible* (x, y)-values [those for which $f(x, y) \neq 0$].

The **probability** that the (X, Y)-values will fall inside **any** other two-dimensional set (**region**) $\mathcal{A}$ is computed by

$$\iint_{\mathcal{A}} f(x, y)\, dxdy$$

The main challenge here is to understand 2-D regions, their mathematical description, and the corresponding double integration. Note that now, the *probability* of a region is the corresponding *volume* between the x-y plane and $f(x, y)$.

Similarly, a **three-dimensional pdf** $f(x, y, z)$ is defined as the probability of X, Y and Z having their values fall (occur) in some specific [small] 3-D region containing (x, y, z) [usually a cube with (x, y, z) at one of its corners], divided by the *volume* of this region, in the limit of this region shrinking to the (x, y, z) point itself. The probability of X, Y and Z having their values fall in a subset $\mathcal{A}$ of the 3-D space equals

$$\iiint_{\mathcal{A}} f(x, y, z)\, dx\, dy\, dz$$

The extension to four or more RVs is quite obvious.

EXAMPLE:

Consider the following bivariate pdf: $f(x, y) = x + y$ for $\begin{cases} 0 \leq x \leq 1 \\ 0 \leq y \leq 1 \end{cases}$, zero otherwise (eventually, this too will go without saying). Thus, the resulting 2-D point can occur only in the unit square with corners at (0,0) and $(1, 1)$, but *not* with the same probability [points near $(1, 1)$ are the most likely, points near $(0, 0)$ are very unlikely to happen].

Obviously, in its region of definition, $f(x, y) \geq 0$. Let us verify the second condition of 'legitimacy', namely that $\iint_{\mathcal{R}} f(x, y)\, dxdy = 1$. The double integration can be converted to two consecutive single-variable integrals, thus: $\int_{y=0}^{1} \left(\int_{x=0}^{1} (x + y)\, dx \right) dy$ which we usually write as $\int_0^1 \int_0^1 (x + y)\, dxdy$ [it is understood that dx goes with the *inner* set of limits and dy corresponds to the *outer* set, similar to *nesting* parentheses]. The inner (dx) integration

must be performed first: $\int_0^1 \left[\frac{x^2}{2}+xy\right]_{x=0}^1 dy = \int_0^1 (\frac{1}{2}+y)dy$ followed by the dy-integration: $\left[\frac{y}{2}+\frac{y^2}{2}\right]_{y=0}^1 = 1$ (check).

Let us now compute $\Pr(X<\frac{1}{2} \cap Y<\frac{1}{2})$. First we must be able to visualize the corresponding region, then its *overlap* with the *region of definition* [getting the $(0,0)$-$(\frac{1}{2},\frac{1}{2})$ square]. The pdf is then integrated over this overlap: $\int_0^{\frac{1}{2}}\int_0^{\frac{1}{2}}(x+y)\,dxdy = \int_0^{\frac{1}{2}}\left[\frac{x^2}{2}+xy\right]_{x=0}^{\frac{1}{2}} dy = \int_0^{\frac{1}{2}}(\frac{1}{8}+\frac{y}{2})\,dy = [\frac{y}{8}+\frac{y^2}{4}]_{y=0}^{\frac{1}{2}} = \frac{1}{8}$. Note that the area of the $(0,0)$-$(\frac{1}{2},\frac{1}{2})$ square is $\frac{1}{4}$ of the area of the region of definition (the unit square), but its probability is only $\frac{1}{8}$ of the total probability, as it covers the least likely part of the 'target'.

Moral: For bivariate distributions, probability is, in general, *not* proportional to the corresponding area [this is possible only in a very special case of a *constant* pdf, i.e. all points of the target are equally likely to be 'hit'].

And the final question: Compute $\Pr(X+Y<\frac{1}{2})$.

First we realize that the condition represents a half-plane *below* the $x+y=\frac{1}{2}$ straight line, then take its overlap with the unit-square 'target', getting a triangle [with corners at $(0,0)$, $(0,\frac{1}{2})$ and $(\frac{1}{2},0)$]. To get the answer, we integrate the pdf over this triangle: $\int_0^{\frac{1}{2}}\int_0^{\frac{1}{2}-y}(x+y)\,dxdy = \int_0^{\frac{1}{2}}\left[\frac{x^2}{2}+xy\right]_{x=0}^{\frac{1}{2}-y} dy = \int_0^{\frac{1}{2}}(\frac{1}{8}-\frac{y^2}{2})\,dy = [\frac{y}{8}-\frac{y^3}{6}]_{y=0}^{\frac{1}{2}} = \frac{1}{24}$.

Note how the triangle is sliced in the x-direction: the x-limits follow from the equations of the two boundaries (where x enters and leaves the triangle), and the y-limits are established from the *projection* of the triangle into the y-axis [its y-'shadow'] and are [must always be] free of x. Also note that the answer is *not* one half of the previous answer (even though the corresponding *areas* are in that ratio). ■

▶Marginal Distributions◀

To get the marginal pdf of Y from a joint pdf of X and Y one has to *integrate* $f(x,y)$ over the **conditional range** of x-values given y. The answer (a function of y) is valid (i.e. non-zero) over the **marginal range** of y [the idea of piercing the region of definition by x-parallel lines, and observing its y-axis shadow].

Similarly, to get the X-marginal, one does

$$f(x) = \int_{All\ y|x} f(x,y)\,dy$$

[the *integration* is over the *conditional* range of y given x], valid over the *marginal* range of the x values.

EXAMPLES:

1. Using the joint pdf of the previous example, we get $f(x) = \int_0^1 (x+y)\,dy = xy + \frac{y^2}{2}\Big|_{y=0}^1 = x + \frac{1}{2}$ for $0 < x < 1$ [zero otherwise]. Note that spelling out the marginal range of x is an important part of the answer! The answer must be a regular, univariate pdf: $\int_0^1 (x+\frac{1}{2})\,dx = \frac{x^2}{2} + \frac{x}{2}\Big|_{x=0}^1 = 1$ (check). Since the original bivariate pdf and its region of definition are symmetric with respect to the $x \leftrightarrow y$ interchange, it's obvious that $f(y) = y + \frac{1}{2}$ for $0 < y < 1$.

2. Consider $f(x,y) = \frac{1}{y}$ for $\begin{cases} 0 < x < y \\ 0 < y < 1 \end{cases}$ [zero otherwise]. It is essential to be able to visualize the region of definition [a triangle with corners at (0,0), $(0,1)$ and $(1,1)$]. Note that this region has been described by 'horizontal-lines', i.e. specifying the *conditional* range of x-values first, followed by the *marginal* range of y. One can of course reverse this: $\begin{cases} x < y < 1 \\ 0 < x < 1 \end{cases}$ covers exactly the *same* triangle. This is the crucial thing to understand, as we often need to switch from one description to the other.

   To get the Y marginal, we do $\int_0^y \frac{dx}{y} = \frac{1}{y} \cdot [x]_{x=0}^y = 1$ for $0 < y < 1$ [here we needed the horizontal-line description], to get the X marginal: $\int_x^1 \frac{dy}{y} = \ln y|_{y=x}^1 = -\ln x$ for $0 < x < 1$ [vertical lines]. Let us check that $f(x)$ is a legitimate pdf: $-\ln x$ is positive (since $0 < x < 1$), it integrates to $-\int_0^1 \ln x\,dx = x(1-\ln x)|_{x=0}^1 = 1$ (check) [recall that $\lim_{x\to 0} x \ln x = 0$].

3. Consider $f(x,y) = \frac{1}{\pi}$ for $x^2 + y^2 < 1$ (zero otherwise) [i.e. the pdf has a constant value of $\frac{1}{\pi}$ inside the unit disk centered on the origin]. Recall that $\frac{1}{\pi} \iint_{x^2+y^2<1} dx\,dy = \frac{1}{\pi} \cdot Area(x^2+y^2<1) = 1$ [check].

   To get the Y marginal: $\frac{1}{\pi} \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} dx = \frac{2}{\pi}\sqrt{1-y^2}$ for $-1 < y < 1$ [draw a picture of the disk and the x-parallel segments 'shading' it]. Obviously $f(x) = \frac{2}{\pi}\sqrt{1-x^2}$ for $-1 < x < 1$ based on the $x \leftrightarrow y$ symmetry.

   Remember: $\iint_{\mathcal{A}} c\,dx\,dy = c \cdot Area(\mathcal{A})$, i.e. integrating a constant can be done 'geometrically', *bypassing* the integration!

4. Consider $f(x,y) = 2x(x-y)$ for $\begin{cases} 0 < x < 1 \\ -x < y < x \end{cases}$ [this is a triangle with corners at $(0,0)$, $(1,1)$ and $(1,-1)$; also note that we are not always consistently starting with the conditional range, here we first quote the marginal range of x]. Since we want to build both marginals, the most difficult part will be

expressing the same region in the 'horizontal-line' representation. Here we have one extra difficulty: describing the triangle by horizontal lines is impossible in 'one sweep' as we have to switch between formulas to spell out the lower conditional limit of x [these things are short of impossible to visualize *unless* you draw a *picture*]. Thus, the horizontal-line description is a union of two 'sub-regions': $\left\{ \left[\begin{array}{c} -1 < y < 0 \\ -y < x < 1 \end{array} \right] \cup \left[\begin{array}{c} 0 < y < 1 \\ y < x < 1 \end{array} \right] \right.$. Now we can easily find the marginal pdf's:

$$f(x) = 2x \int_{-x}^{x} (x - y)\, dy = 2x[xy - \tfrac{y^2}{2}]_{y=-x}^{x} = 4x^3 \text{ for } 0 < x < 1.$$

The Y marginal has to be build in two parts: When $-1 < y < 0$, the expression for $f(y)$ is $2 \int_{-y}^{1} (x^2 - xy)\, dx = 2[\frac{x^3}{3} - \frac{x^2}{2}y]_{x=-y}^{1} = \frac{2}{3} - y + \frac{5}{3}y^3$. When $0 < y < 1$, we get $f(y) = 2 \int_{y}^{1} (x^2 - xy)\, dx = 2[\frac{x^3}{3} - \frac{x^2}{2}y]_{x=y}^{1} = \frac{2}{3} - y + \frac{y^3}{3}$. To summarize: $f(y) = \left\{ \begin{array}{ll} \frac{2}{3} - y + \frac{5}{3}y^3 & -1 < y < 0 \\ \frac{2}{3} - y + \frac{y^3}{3} & 0 < y < 1 \end{array} \right.$.

Let us also compute some probabilities:

$\Pr(X + Y < 1)$. The joint pdf needs to be integrated over the intersection of the original triangle and the half-plane below $x + y = 1$. This can be done in several distinct ways:

(a) Starting with the y-integration (vertical lines) and then integrating x over its marginal (projection) range, but only when broken into two sub-regions: $\int_{0}^{\frac{1}{2}} \int_{-x}^{x} f(x, y)\, dy\, dx + \int_{\frac{1}{2}}^{1} \int_{-x}^{1-x} f(x, y)\, dy\, dx$

(b) Starting with the (conditional) x-integration [still two sub-regions]: $\int_{0}^{\frac{1}{2}} \int_{y}^{1-y} f(x, y)\, dx\, dy + \int_{-1}^{0} \int_{-y}^{1} f(x, y)\, dx\, dy$

(c) Using $\Pr(A) = 1 - \Pr(\bar{A})$: $1 - \int_{0}^{\frac{1}{2}} \int_{1-y}^{1} f(x, y)\, dx\, dy - \int_{\frac{1}{2}}^{1} \int_{y}^{1} f(x, y)\, dx\, dy$

[horizontal segments] $\equiv 1 - \int_{\frac{1}{2}}^{1} \int_{1-x}^{x} f(x, y)\, dy\, dx$ [vertically].

One can verify that all four distinct ways of computing the probability yield the same answer, but the last approach is obviously the most economical, resulting in $1 - 2 \int_{\frac{1}{2}}^{1} x \int_{1-x}^{x} (x - y)\, dy\, dx = 1 - 2 \int_{\frac{1}{2}}^{1} x[xy - \frac{y^2}{2}]_{y=1-x}^{x}\, dx = 1 - 2 \int_{\frac{1}{2}}^{1} x(2x^2 - 2x + \frac{1}{2})\, dx = 1 - 2[2\frac{x^4}{4} - 2\frac{x^3}{3} + \frac{x^2}{4}]_{x=\frac{1}{2}}^{1} = \frac{41}{48} = 85.42\%$.

$\Pr(Y > 0)$ [the upper half of the original triangle]. Depending on whether we want to 'sweep' the region vertically or horizontally, we have: $2\int_0^1 x \int_0^x (x - y)\,dy\,dx = 2\int_0^1 x[xy - \frac{y^2}{2}]_{y=0}^{x}\,dx = 2\int_0^1 x(\frac{x^2}{2})\,dx = \left.\frac{x^4}{4}\right|_{x=0}^{1} = \frac{1}{4}$ or $2\int_0^1 \int_y^1 x(x - y)\,dx\,dy = 2\int_0^1 [\frac{x^3}{3} - \frac{x^2}{2}y]_{x=y}^{1}\,dy = 2\int_0^1 (\frac{1}{3} - \frac{y}{2} + \frac{y^3}{6})\,dy = 2[\frac{y}{3} - \frac{y^2}{4} + \frac{y^4}{24}]_{y=0}^{1} = \frac{1}{4}$ (the same answer).

5. Let's try a three-dimensional problem. Consider the following tri-variate pdf: $f(x, y, z) = c(x + y + z)$ where $\begin{cases} x > 0 \\ y > 0 \\ z > 0 \\ x + y + z < 1 \end{cases}$ [c being an appropriate constant]. First, we want to find the value of c. The tricky part is that the region of definition is *not* described in a manner which would readily translate into the corresponding limits of consecutive integration. What we need is first the *conditional* range of x given both y and z [think of piercing the structure by a straight line parallel to x to establish where it enters and where it leaves the volume, getting, in this case, $0 < x < 1 - y - z$], followed by partly marginal (as x is already out and we are looking at the y-z projection of the original tetragonal region) partly conditional (given z) range of y [$0 < y < 1 - z$], and finally the fully marginal (the y-z 'shadow' is itself projected into z only) range of the z-values [$0 < z < 1$]. Depending on the order in which we do this (x, y, z, or y, x, z, ...) we end up with altogether 3! distinct ways of performing the integration [they all have to result in the same answer, but some of them may be a lot more difficult − or outright impossible − than others]. In this case both the region and $f(x, y, z)$ are symmetrical, and all 6 choices are effectively identical.

We now integrate our pdf in the indicated manner: $\int_0^1 \int_0^{1-z} \int_0^{1-y-z} c(x + y + z)\,dx\,dy\,dz = c\int_0^1 \int_0^{1-z} \int_0^{1-y-z} [\frac{x^2}{2} + (y+z)x]_{x=0}^{1-y-z}\,dy\,dz = c\int_0^1 \int_0^{1-z} \left(\frac{1}{2} - \frac{(y+z)^2}{2}\right) dy\,dz = c\int_0^1 \left[\frac{y}{2} - \frac{(y+z)^3}{6}\right]_{y=0}^{1-z} dz = \frac{c}{2}\int_0^1 (\frac{2}{3} - z + \frac{z^3}{3})\,dz = \frac{c}{2}[\frac{2}{3}z - \frac{z^2}{2} + \frac{z^4}{12}]_{z=0}^{1} = \frac{c}{8}$ which must be equal to 1 (total probability). This implies that $c = 8$.

6. Finally, we do an example of a 2-D problem which can be easily solved only with the help of POLAR COORDINATES. Suppose we want to establish the value of c of the following bivariate pdf:

$$f(x, y) = \frac{c}{\sqrt{x^2 + y^2}} \quad \text{when} \quad \begin{cases} x^2 + y^2 < 1 \\ y > 0 \end{cases}$$

The region of definition is clearly the upper half of the unit disk centered on the origin.

To perform the implied integration, we introduce two new variables r and φ, replacing the old x and y pair according to the following formulas

$$\begin{aligned} x &= r\cos\varphi \\ y &= r\sin\varphi \end{aligned}$$

By substitution these into $f(x,y)$, we make it a function of r and φ, to be integrated over the same half disk (which, in the new coordinates, becomes the $\begin{cases} 0<r<1 \\ 0<\varphi<\pi \end{cases}$ *rectangle*. The only question is: What becomes of the infinitesimal area $dx\,dy$ of the original double integral? The answer is: $J\,dr\,d\varphi$, where J is the so called JACOBIAN of the transformation, computed as follows

$$J = \left|\begin{array}{cc} \frac{dx}{dr} & \frac{dx}{d\varphi} \\ \frac{dy}{dr} & \frac{dy}{d\varphi} \end{array}\right| = \left|\begin{array}{cc} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{array}\right| = r\cos^2\varphi + r\sin^2\varphi = r$$

(the bar around the 2×2 matrix implies taking the absolute value of its determinant). Geometrically, $J\,dr\,d\varphi$ represents the area of the near rectangle created by increasing φ by $d\varphi$ and increasing r by dr.

We can thus evaluate the original integral by

$$\iint\limits_{\substack{x^2+y^2<1 \\ y>0}} \frac{dx\,dy}{\sqrt{x^2+y^2}} = \iint\limits_{\substack{0<r<1 \\ 0<\varphi<\pi}} \frac{r\,dr\,d\varphi}{r} = \int\limits_0^\pi d\varphi \times \int\limits_0^1 dr = \pi$$

The value of c is thus equal to $\frac{1}{\pi}$. ■

▶Conditional Distribution◀

is the distribution of X *given* that Y has been observed to result in a specific value $\underline{y}$. The corresponding **conditional pdf** $f(x|Y=\underline{y})$ should be defined, according to the general definition, via $\lim\limits_{\varepsilon\to 0} \dfrac{\Pr(x\le X<x+\varepsilon|Y=\underline{y})}{\varepsilon}$ which unfortunately leads to $\frac{0}{0}$ [as $\Pr(Y=\underline{y})\equiv 0$]. We can make this expression meaningful by first allowing Y to be in a $[\underline{y},\underline{y}+\delta)$ interval and then taking $\delta\to 0$.

This leads to $f(x|Y=\underline{y}) = \lim\limits_{\delta\to 0}\lim\limits_{\varepsilon\to 0} \dfrac{\Pr(x\le X<x+\varepsilon|\underline{y}\le Y<\underline{y}+\delta)}{\varepsilon} =$

$\lim\limits_{\delta\to 0}\lim\limits_{\varepsilon\to 0} \dfrac{\Pr(x\le X<x+\varepsilon\cap \underline{y}\le Y<\underline{y}+\delta)}{\varepsilon\Pr(\underline{y}\le Y<\underline{y}+\delta)}$ [by the general definition of conditional probability] $= \lim\limits_{\delta\to 0}\lim\limits_{\varepsilon\to 0} \dfrac{\Pr(x\le X<x+\varepsilon\cap \underline{y}\le Y<\underline{y}+\delta)}{\varepsilon\cdot\delta}\cdot\dfrac{\delta}{\Pr(\underline{y}\le Y<\underline{y}+\delta)} =$

$$\frac{f(x,\underline{y})}{f(\underline{y})}$$

by our previous definition of joint (the numerator) and univariate (the denominator) pdf, where x varies over its conditional range, given $\underline{y}$. Note that the

denominator, representing the marginal pdf of Y at $\underline{y}$, is normally computed by $\int\limits_{All\ x|\underline{y}} f(x,\underline{y})\,dx$.

This definition can be **extended** to define a conditional pdf of X given that Y and Z have been already observed to result in a specific value of $\underline{y}$ and $\underline{z}$, respectively:

$$f(x|Y=\underline{y}\cap Z=\underline{z})=\frac{f(x,\underline{y},\underline{z})}{f(\underline{y},\underline{z})}$$

where x varies over the corresponding conditional range (found by piercing the original region of definition by a straight line parallel to x), and the denominator is computed from $\int\limits_{All\ x|\underline{y}|\underline{z}} f(x,\underline{y},\underline{z})\,dx$.

Similarly, the **joint conditional pdf** of X and Y given an (observed) value of Z is:

$$f(x,y|Z=\underline{z})=\frac{f(x,y,\underline{z})}{f(\underline{z})}$$

with x and y varying over the corresponding conditional two-dimensional range (the $Z=\underline{z}$ 'slice' of the original three-dimensional region of definition), where the denominator is $\iint\limits_{All\ (x,y)|\underline{z}} f(x,y,\underline{z})\,dx\,dy$.

Note that the most difficult part of these formulas is always the denominator containing a *marginal* pdf which needs to be constructed first by the corresponding multivariate integration.

EXAMPLES:

1. Let us go back to the bivariate pdf of the previous Example 4. We want to construct $f(x|Y=\frac{1}{2})$. Luckily, we already have the formula for not only $f(x,y)=2x(x-y)$ but also for $f_Y(y)=\frac{2}{3}-y+\frac{y^3}{3}$ (when $0<y<1$). Thus $f(x|Y=\frac{1}{2})=\frac{2x(x-\frac{1}{2})}{\frac{2}{3}-\frac{1}{2}+\frac{1}{24}}=\frac{48}{5}x(x-\frac{1}{2})$, where $\frac{1}{2}<x<1$ [cut the original triangle by a horizontal line at $y=\frac{1}{2}$ and observe the resulting conditional range of x]. To verify the answer: $\frac{48}{5}\int\limits_{x=\frac{1}{2}}^{1} x(x-\frac{1}{2})\,dx=\frac{48}{5}[\frac{x^3}{3}-\frac{x^2}{2}]_{x=\frac{1}{2}}^{1}=1$ (check $-$ a conditional pdf must also 'add up' to 1).

   Secondly, let us find $f(y|X=\frac{1}{3})=\frac{2\times\frac{1}{3}\times(\frac{1}{3}-y)}{4\times(\frac{1}{3})^3}=\frac{9}{2}(\frac{1}{3}-y)$, where $-\frac{1}{3}<y<\frac{1}{3}$ [zero otherwise] (cut the original triangle along $x=\frac{1}{3}$ to understand the conditional range of y). To verify: $\frac{9}{2}\int\limits_{-\frac{1}{3}}^{\frac{1}{3}}(\frac{1}{3}-y)\,dy=\frac{9}{2}[\frac{y}{3}-\frac{y^2}{2}]_{y=-\frac{1}{3}}^{\frac{1}{3}}=1$ (check).

2. Using the tri-variate distribution of the previous section, we find the conditional distribution of X given that $Y=\frac{1}{2}$ and $Z=\frac{1}{4}$ by

$$\begin{aligned} f(x\,|\,Y &= \frac{1}{2}\cap Z=\frac{1}{4})=\frac{x+\frac{1}{2}+\frac{1}{4}}{\int_0^{1/4}(x+\frac{3}{4})\,dx}=\frac{x+\frac{3}{4}}{\frac{7}{32}}= \\ &= \frac{32}{7}\,x+\frac{24}{7} \quad \text{for} \quad 0<x<\frac{1}{4} \end{aligned}$$

Similarly, the conditional (still bivariate) pdf of X and Y given that $Z = \frac{1}{2}$ is constructed as follows:

$$f(x,y\,|\,Z = \frac{1}{2}) = \frac{x+y+\frac{1}{2}}{\iint\limits_{\substack{x+y<\frac{1}{2}\\ x,y>0}} (x+y+\frac{1}{2})\,dA} = \frac{x+y+\frac{1}{2}}{\int\limits_0^{\frac{1}{2}}\int\limits_0^{\frac{1}{2}-y}(x+y+\frac{1}{2})dx\,dy} =$$

$$= \frac{x+y+\frac{1}{2}}{\int\limits_0^{\frac{1}{2}}(\frac{3}{8}-\frac{y}{2}-\frac{y^2}{2})\,dy} = \frac{x+y+\frac{1}{2}}{\frac{5}{48}} =$$

$$= \frac{48(x+y)+24}{5} \qquad \text{where} \qquad \begin{cases} x+y<\frac{1}{2} \\ x>0 \\ y>0 \end{cases}$$

►Mutual Independence◄

of two or more RVs can be easily established just by *inspecting* the corresponding *joint* pdf.

In the bivariate case, X and Y are independent iff

(i) $f(x,y)$ can be written as a *product* of a constant, times a function of x, times a function of y,

and

(ii) the *region of definition* can be expressed as $\begin{cases} a<x<b \\ c<y<d \end{cases}$ where a, b, c, d are constant numbers (i.e. a and b are free of y, and c and d are free of x). Note that in some cases a and/or c can have the value of $-\infty$, and b and/or d can equal to ∞. I like to call any such region (with fixed but potentially infinite limits) a 'generalized rectangle'. ■

The **extension** of these independence criteria to the case of *several* RVs should be obvious.

As in the discrete case, mutual independence of RVs **implies** that any *conditional* distribution is identically *equal* to the corresponding [ignoring the condition(s)] *marginal* distribution.

Expected value

of a continuous RV is computed via

$$\mathbb{E}(X) = \int\limits_{All\ x} x \cdot f(x)\,dx$$

[as always, the formula is fully analogous to the discrete case, only summation is replaced by integration].

Similarly:

$$\mathbb{E}[g(X)] = \int\limits_{All\ x} g(x) \cdot f(x)\,dx$$

where $g(x)$ is an arbitrary function of x, and

$$\mathbb{E}[g(X,Y)] = \iint\limits_{\mathcal{R}} g(x,y)\cdot f(x,y)\,dx\,dy$$

for any bivariate function $g(x,y)$.

Simple *moments*, central moments, variance, covariance, etc. are defined in exactly *same* manner as in the discrete case. Also, all previous formulas for dealing with *linear combinations* of RVs (expected value, variance, covariance) still hold, without change.

EXAMPLES:

1. $f(x) = \frac{1}{2\pi}$ for $0 < x < 2\pi$ [the spinning wheel]. $\mathbb{E}(X) = \frac{1}{2\pi}\int\limits_0^{2\pi} x\,dx = \frac{1}{2\pi}[\frac{x^2}{2}]_{x=0}^{2\pi} = \pi$ [not surprisingly, the mean is at the distribution's 'center'].

   In general, if the pdf is **symmetric** [i.e. $f(\alpha+x) = f(\alpha-x)$ for some α -a graph will reveal the symmetry better than any formula] then the *center* of symmetry α must be the expected value of X, *provided* that the expected value *exists*!

   Symmetry does not help though when we want to evaluate $\mathbb{E}(X^2) = \frac{1}{2\pi}\int\limits_0^{2\pi} x^2 dx = \frac{1}{2\pi}[\frac{x^3}{3}]_{x=0}^{2\pi} = \frac{4\pi^2}{3}$. This implies the variance of $\frac{4\pi^2}{3} - \pi^2 = \frac{\pi^2}{3} \Rightarrow \sigma = \frac{\pi}{\sqrt{3}}$. The $\mu \pm \sigma$ range thus contains $\frac{1}{\sqrt{3}} = 57.74\%$ of all probability, confirming our 'bulk' rule.

   Similarly, $\mathbb{E}(\frac{1}{1+X^2}) = \frac{1}{2\pi}\int\limits_0^{2\pi} \frac{dx}{1+x^2} = \frac{1}{2\pi}[\arctan x]_{x=0}^{2\pi} = 0.22488$.

2. $f(x) = e^{-x}$ for $x > 0$. $\mathbb{E}(X) = \int\limits_0^{\infty} xe^{-x}\,dx = \int\limits_0^{\infty} x(-e^{-x})'\,dx = -xe^{-x}|_{x=0}^{\infty} + \int\limits_0^{\infty} e^{-x}\,dx = [-e^{-x}]_{x=0}^{\infty} = 1$.

   To find $Var(X)$ we first need $\mathbb{E}(X^2) = \int\limits_0^{\infty} x^2 e^{-x}\,dx$. The general formula for dealing with these type of integrals is: $\int\limits_0^{\infty} x^n e^{-x}\,dx = \int\limits_0^{\infty} x^n(-e^{-x})'\,dx = 0 + n\int\limits_0^{\infty} x^{n-1}e^{-x}\,dx = n(n-1)\int\limits_0^{\infty} x^{n-2}e^{-x}\,dx = \ldots\ldots = n(n-1)(n-1)\ldots.2\times 1\int\limits_0^{\infty} e^{-x}\,dx = n!$ [where n is a non-negative integer]. Thus $\mathbb{E}(X^2) = 2!$ and $Var(X) = 2 - 1^2 = 1$. The probability of falling inside the $\mu \pm \sigma$ range is thus $1 - e^{-2} = 86.47\%$.

3. $f(x) = \begin{cases} \frac{2}{3}(1+x) & -1 < x < 0 \\ \frac{4}{3}(1-x) & 0 < x < 1 \end{cases}$ The integration needs to be done in two parts to be *added* together [a common mistake is to display them as

two separate expected values, one for each of the two ranges – this is totally nonsensical, a piece-wise $f(x)$ still corresponds to only *one* X and therefore *one* expected value]: $\mathbb{E}(X) = \frac{2}{3}\int_{-1}^{0} x(1+x)\,dx + \frac{4}{3}\int_{0}^{1} x(1-x)\,dx = \frac{2}{3}[\frac{x^2}{2} + \frac{x^3}{3}]_{x=-1}^{0} + \frac{4}{3}[\frac{x^2}{2} - \frac{x^3}{3}]_{x=0}^{1} = \frac{1}{9}$.

4. Consider the bivariate pdf of Example 4 from the 'Marginal Distributions' section. Find:

a) $\mathbb{E}(\frac{Y}{X})$

Solution: $2\int_0^1\int_{-x}^{x} \frac{y}{x}\cdot x(x-y)\,dy\,dx = 2\int_0^1 [\frac{y^2}{2}x - \frac{y^3}{3}]_{y=-x}^{x}\,dx = 2\int_0^1 (-\frac{2x^3}{3})dx = -\frac{x^4}{3}\Big|_{x=0}^{1} = -\frac{1}{3}$.

b) $Cov(X,Y)$

Solution: First we compute $\mathbb{E}(X\cdot Y) = 2\int_0^1\int_{-x}^{x} xy\cdot x(x-y)\,dy\,dx = 2\int_0^1 x^2[\frac{y^2}{2}x - \frac{y^3}{3}]_{y=-x}^{x}dx = 2\int_0^1(-\frac{2x^5}{3})dx = -\frac{2x^6}{9}\Big|_{x=0}^{1} = -\frac{2}{9}$

$\mathbb{E}(X) = 2\int_0^1\int_{-x}^{x} x\cdot x(x-y)\,dy\,dx = 2\int_0^1 x^2[xy - \frac{y^2}{2}]_{y=-x}^{x}dx = 2\int_0^1 2x^4\,dx = \frac{4x^5}{5}\Big|_{x=0}^{1} = \frac{4}{5}$

and $\mathbb{E}(Y) = 2\int_0^1\int_{-x}^{x} y\cdot x(x-y)\,dy\,dx = 2\int_0^1 x[\frac{y^2}{2}x - \frac{y^3}{3}]_{y=-x}^{x}dx = 2\int_0^1(-\frac{2}{3}x^4)\,dx = -\frac{4x^5}{15}\Big|_{x=0}^{1} = -\frac{4}{15} \Rightarrow$

Answer: $Cov(X,Y) = -\frac{2}{9} + \frac{4}{5}\times\frac{4}{15} = -0.00\bar{8}$.

c) $Var(X)$, $Var(Y)$ and ρ_{XY}.

Solution: The extra moments we need are $\mathbb{E}(X^2) = 2\int_0^1\int_{-x}^{x} x^2\cdot x(x-y)\,dy\,dx = 2\int_0^1 x^3[xy - \frac{y^2}{2}]_{y=-x}^{x}dx = 2\int_0^1 2x^5\,dx = \frac{2x^6}{3}\Big|_{x=0}^{1} = \frac{2}{3}$

and $\mathbb{E}(Y^2) = 2\int_0^1\int_{-x}^{x} y^2\cdot x(x-y)\,dy\,dx = 2\int_0^1 x[\frac{y^3}{3}x - \frac{y^4}{4}]_{y=-x}^{x}dx = 2\int_0^1(\frac{2}{3}x^5)\,dx = \frac{2x^6}{9}\Big|_{x=0}^{1} = \frac{2}{9} \Rightarrow$

$Var(X) = \frac{2}{3} - (\frac{4}{5})^2 = 0.02\bar{6}$, $Var(Y) = \frac{2}{9} - (\frac{4}{15})^2 = 0.15\bar{1}$ and $\rho_{XY} = \frac{-0.00\bar{8}}{\sqrt{0.02\bar{6}\times 0.15\bar{1}}} = -0.1400$. ■

is defined via:

$$M_X(t) \equiv \mathbb{E}(e^{tX}) = \int\limits_{All\ x} e^{tx} \cdot f(x)\, dx$$

[again, the old definition with summation replaced by integration].

The

$$M_{aX+b}(t) = e^{bt} M_X(at)$$

formula still holds true [its proof would be practically identical to the discrete case], and independence of X_1 and X_2 still implies

$$M_{X_1+X_2}(t) = M_{X_1}(t) \cdot M_{X_2}(t)$$

(which can be extended to any number of mutually independent RVs).

EXAMPLE:

$f(x) = e^{-x}$ when $x > 0 \ \Rightarrow \ M(t) = \int\limits_0^\infty e^{tx} \cdot e^{-x}\, dx = \left[\frac{e^{-x(1-t)}}{-(1-t)}\right]_{x=0}^{\infty} = \frac{1}{1-t}$.

Since this MGF can be readily expanded, thus: $\frac{1}{1-t} = 1 + t + t^2 + t^3 + \ldots$, we instantly know all the simple moments of X: $\mathbb{E}(X) = 1$, $\mathbb{E}(X^2) = 2$, $\mathbb{E}(X^3) = 3!$, $\mathbb{E}(X^4) = 4!$, ..., etc.

To find the MGF of $2X - 3$ is equally trivial, resulting in $\frac{e^{-3t}}{1-2t}$. ■

▶Conditional Expected Value◀

is again (as in the discrete case) an expected value computed using the corresponding *conditional* pdf:

$$\mathbb{E}(g(X)|Y = \underline{y}) = \int\limits_{All\ x|\underline{y}} g(x) \cdot f(x|Y = \underline{y})\, dx = \int\limits_{All\ x|\underline{y}} g(x) \frac{f(x, \underline{y})}{f(\underline{y})} dx$$

EXAMPLES:

1. Consider, again, the bivariate pdf of Example 4 from the 'Marginal Distributions' section. Find $\mathbb{E}(\frac{1}{X}|Y = \frac{1}{2})$.

   Solution: This equals $\frac{48}{5} \int\limits_{1/2}^{1} \frac{1}{x} \cdot x(x - \frac{1}{2})\, dx = \frac{48}{5}[\frac{x^2}{2} - \frac{x}{2}]_{x=\frac{1}{2}}^{1} = \frac{6}{5}$ [we utilized $f(x|Y = \frac{1}{2})$ constructed earlier].

2. Consider $f(x, y) = 2e^{-x-y}$ when $\left\{ \begin{array}{l} x > 0 \\ y > x \end{array} \right.$. Find $\mathbb{E}(e^{\frac{Y}{2}}|X = 1)$.

   Solution: First, we must construct the corresponding $f(x)$-marginal, by: $2e^{-x} \int\limits_x^\infty e^{-y} dy = 2e^{-x} \cdot [-e^{-y}]_{y=x}^{\infty} = 2e^{-2x}$ when $x > 0$. Then, we build $f(y|X = 1)$ by: $\frac{2e^{-1-y}}{2e^{-2}} = e^{1-y}$ when $y > 1$ [the ranges can be easily read off the appropriate graph]. Finally: $\mathbb{E}(e^{\frac{Y}{2}}|X = 1) = \int\limits_1^\infty e^{\frac{y}{2}} \cdot e^{1-y}\, dy = \int\limits_1^\infty e^{1-\frac{y}{2}}\, dy = e \cdot [-2e^{-\frac{y}{2}}]_{y=1}^{\infty} = 2\sqrt{e} = 3.2974$. ■

Chapter 7 SPECIAL CONTINUOUS DISTRIBUTIONS

Univariate (single RV) case

►Uniform◄

$f(x) = \frac{1}{b-a}$ [constant] when $a < x < b$. $F(x) = \frac{x-a}{b-a}$ when $a < x < b$, $\mathbb{E}(X) = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_{x=a}^{b} = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$ [center of symmetry], $\mathbb{E}(X^2) = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_{x=a}^{b} = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3} \Rightarrow$ $Var(X) = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{4(a^2+ab+b^2)-3(a^2+2ab+b^2)}{12} = \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12} \Rightarrow$

$$\sigma_X = \frac{b-a}{2\sqrt{3}}$$

This implies that $\mu_X \pm \sigma_X$, being an interval of length $\frac{b-a}{\sqrt{3}}$, contains $\frac{1}{\sqrt{3}} = 57.74\%$ of the total probability (consistent with our rough rule that this should be between 50 and 90%). Note that $\mu_X \pm 2\sigma_X$ already covers the whole possible range, and thus contains 100% of the total probability.

For this distribution, we use the following **symbolic notation**: $\mathcal{U}(a,b)$. Thus, for example, $X \in \mathcal{U}(0,1)$ implies that the distribution of X is uniform, going from 0 to 1.

►'Standardized'(i.e. $\mu = 0$ and $\sigma = 1$) Normal◄

distribution (the corresponding RV will be called Z – a 'reserved' name from now on).

In the last chapter we discovered that, when sampling from 'almost' any population, the distribution of $\frac{X_1+X_2+....+X_n-n\mu}{\sqrt{n}\sigma}$ approaches, for large n, a unique distribution whose MGF is $e^{\frac{t^2}{2}}$.

We will now show that this MGF corresponds to the following **pdf**: $f(z) = c \cdot e^{-\frac{z^2}{2}}$ for $-\infty < z < \infty$ [any z], where c is an appropriate constant. To find its value, we need $I \equiv \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,dz$. Unfortunately, $e^{-\frac{z^2}{2}}$ is one of the functions we *don't know* how to analytically integrate; to evaluate I we need to use the following trick (going to two dimensions): $I^2 = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx \times \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy = \iint_{x-y\ plane} e^{-\frac{x^2+y^2}{2}}\,dx\,dy =$ [in polar coordinates] $\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} r\,dr\,d\varphi = 2\pi\int_0^{\infty} e^{-u}\,du = 2\pi$ (we will review the change-of-variables technique in more dimensions soon, let me just briefly mention that $\begin{cases} x = r\cos\varphi \\ y = r\sin\varphi \end{cases}$ is the basic relationship between rectangular and polar coordinates,

and $\left|\begin{array}{cc} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} \end{array}\right| = \left|\begin{array}{cc} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{array}\right| = r$ is the Jacobian of the transformation). This implies that $I = \sqrt{2\pi}$, which makes it clear that we need $c = \frac{1}{\sqrt{2\pi}}$:

$$f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}} \qquad -\infty < z < \infty$$

(a symmetric 'bell-shaped' curve).

First we have to verify that this is the pdf whose **MGF** equals $e^{\frac{t^2}{2}}$: $M_Z(t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \cdot e^{zt}\, dz = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{-\frac{z^2}{2}+zt}\, dz = \frac{1}{\sqrt{2\pi}}\, e^{\frac{t^2}{2}} \int\limits_{-\infty}^{\infty} e^{-\frac{(z-t)^2}{2}}\, dz = \frac{1}{\sqrt{2\pi}}\, e^{\frac{t^2}{2}} \int\limits_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\, du =$ $e^{\frac{t^2}{2}}$ [check].

From the expansion $M(t) = 1 + \frac{t^2}{2} + \frac{\left(\frac{t^2}{2}\right)^2}{2} + ...$ we can immediately establish that $\mu_Z = 0$ [all odd **moments** equal to 0], $\sigma_Z = 1$, and the kurtosis of Z is equal to 3.

Since we don't know how to integrate $e^{-\frac{z^2}{2}}$, we cannot find an expression for the corresponding **distribution function** $F(z)$. But without $F(z)$, how do we compute probabilities related to Z (a task so crucial to us)? Well, instead of evaluating $F(z)$ based on some formula, we must now *look up* its values in Table III of our textbook (p.581). Note that this table contains $\Pr(0 < Z < z) \equiv F(z) - \frac{1}{2}$ for values of z from 0 to 3, in steps of 0.01. Now, we need to learn how to use it (some of you may have the corresponding function available in your calculator, feel free to use it, bypassing the 'look up' technique).

EXAMPLES:

1. $\Pr(0 < Z < 1.24) = 0.3925$ [just look it up in the 1.2-row and 0.04-column].

2. $\Pr(-1.3 < Z < 0.5) = \Pr(-1.3 < Z < 0) + \Pr(0 < Z < 0.5) = 0.4032 + 0.1915 = 0.5947$ [since $\Pr(-1.3 < Z < 0) \equiv \Pr(0 < Z < 1.3)$; the distribution is symmetric].

3. $\Pr(0.12 < Z < 0.263) = \Pr(0 < Z < 0.263) - \Pr(0 < Z < 0.12)$.

   We can just look up the second probability (= 0.0478), to find the first one (accurately enough) we have to resort to a so called **linear interpolation:** $\begin{cases} \Pr(0 < Z < 0.26) = 0.1026 \\ \Pr(0 < Z < 0.27) = 0.1064 \end{cases}$. Going from $z = 0.26$ to 0.27 the corresponding probability increased by 0.0038. Assuming this increase ('INCREMENT') follows a *straight line* (a sufficiently accurate approximation over this interval), to find $\Pr(0 < Z < 0.263)$ we have to add, to the first probability of 0.1026, 30% [i.e. $\frac{0.263-0.26}{0.01}$] of the increment, thus: $0.1026 + 0.3 \times 0.0038 = 0.1037$ [quoting only 4 digits after the decimal point, the result cannot get it any more accurate].

   Answer: $\Pr(0.12 < Z < 0.263) = 0.1037 - 0.0478 = 0.0559$.

4. $\Pr(0 < Z < 1.386429)$.

   Solution: $\begin{cases} \Pr(0 < Z < 1.38) = 0.4162 \\ \Pr(0 < Z < 1.39) = 0.4177 \end{cases}$ [increment of 0.0015] $\Rightarrow$

   Answer: $0.4162 + 0.6429 \times 0.0015 = 0.4172$ [note that 0.6429 was obtained from 1.386429 by moving the decimal point two digits to the right and dropping 1.38]. ■

►'General' Normal◄

RV results from a *linear transformation* of Z, thus:

$$X = \sigma Z + \mu$$

where $\sigma > 0$ and μ are two constants. From what we know about linear transformations, we can deduce immediately that $\mathbb{E}(X) = \mu$, $Var(X) = \sigma^2$ [our notation anticipated that], and $M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} \cdot e^{\frac{\sigma^2 t^2}{2}} =$

$$e^{\mu t + \frac{\sigma^2 t^2}{2}} \qquad \text{MGF}$$

[i.e. when the exponent of e is a *quadratic* polynomial in t, the distribution is (general) *normal*; furthermore, based on the polynomial's linear and quadratic coefficients, one can establish the distribution's mean and variance].

EXAMPLE: Based on $M(t) = e^{-2t+t^2}$, identify the distribution:

Answer: $\mathcal{N}(-2, \sqrt{2})$ [$\mathcal{N}(\mu, \sigma)$ will be our **symbolic notation** for 'Normal, with mean μ and standard deviation σ' − note that your textbook uses the *variance* as the second 'argument']. ■

Note that any further **linear transformation** of $X \in \mathcal{N}(\mu, \sigma)$, such as $Y = aX + b$, keeps the result Normal [this follows from $M_Y(t)$]. Furthermore, the mean and standard deviation of Y are $a\mu + b$ and $|a|\sigma$, respectively.

Also: When X_1 and X_2 are **independent** Normal RVs with any (mismatched) parameters [i.e. $X_1 \in \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \in \mathcal{N}(\mu_2, \sigma_2)$], then their **sum** $X_1 + X_2$ is also **Normal** [follows from $M_{X_1+X_2}(t) = e^{(\mu_1+\mu_2)t + \frac{\sigma_1^2+\sigma_2^2}{2}t^2}$], with the mean and standard deviation of $\mu_1 + \mu_2$ and $\sqrt{\sigma_1^2 + \sigma_2^2}$, respectively. (This is a very *unique* property of the Normal distribution, note that Uniform plus Uniform is *not* Uniform, etc.).

To get the **pdf** of a general Normal RV X ($\equiv \sigma Z + \mu$), we correspondingly transform the pdf of Z [we will learn how to do this in the next chapter, suffice to say that $z \to \frac{x-\mu}{\sigma}$ and $dz \to \frac{dx}{\sigma}$]. $\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} (dz)$ thus becomes

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} (dx) \qquad f$$

with $-\infty < x < \infty$ [we have quoted both pdfs with the corresponding infinitesimal, in parentheses, to emphasize the logic of the conversion]. An alternate derivation

of $f(x)$ would differentiate $F_X(x) = \Pr(X < x) = \Pr(\sigma Z + \mu < x) = \Pr(Z < \frac{x-\mu}{\sigma}) = F_Z(\frac{x-\mu}{\sigma})$ with respect to x, try it.

The new pdf (of X) has the same *bell-shaped curve* (as that of Z), but it is centered on μ (instead of the old 0) and it is σ times wider (horizontally) and σ times 'shorter' (vertically), keeping the total area equal to 1. Since we don't know how to directly integrate it, it is of not much use to us, with one exception: we can now **identify** a **Normal** distribution based on its pdf, from which we can also establish the value of μ and σ.

EXAMPLE: Based on $f(x) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{x^2+4x+4}{18}}$ (any x), identify the distribution.

Answer: $\mathcal{N}(-2, 3)$ [there are two ways of establishing the value of σ, they must check]. ■

Finally, the main question is: How do we **compute probabilities** related to the general Normal distribution?

Answer: By converting X back to Z (standardized), because only those tables are readily available.

EXAMPLE: If $X \in \mathcal{N}(17, 3)$, find $\Pr(10 < X < 20)$.

Solution: The last probability equals $\Pr(\frac{10-17}{3} < \frac{X-17}{3} < \frac{20-17}{3}) = \Pr(-2.\bar{3} < Z < 1)$ where Z is the RV of our Normal tables. Since $\Pr(0 < Z < 2.\bar{3}) = 0.4901 + 0.\bar{3} \times 0.0003 = 0.4902$ and $\Pr(0 < Z < 1) = 0.3413$, the final answer is $0.4902 + 0.3413 = 0.8315$ ■

For any Normally distributed RV the $\mu \pm \sigma$ interval contains 68.26% of the total probability, $\mu \pm 2\sigma$ contains 95.44%, and $\mu \pm 3\sigma$ raises this to a 'near certain' 99.74%. Thus, even though *theoretically* the outcome of such RV can yield *any real value*, its *practical range* is limited to a *finite* interval (none of us will ever see a value outside $\mu \pm 4\sigma$).

Applications related to **Central Limit Theorem:**

We learned in the previous chapter that, when sampling, independently, from (almost) any population, the distribution of the *standardized* sample mean is, for large n, approximately standardized Normal (our Z):

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tilde{\in} \mathcal{N}(0, 1)$$

This is of course the same $\bar{X} \tilde{\in} \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ or, equivalently: $X_1 + X_2 + ... + X_n \tilde{\in} \mathcal{N}(n\mu, \sqrt{n}\sigma)$. Since some of our old distributions were built in the $X_1 + X_2 + ... + X_n$ manner, they too will have to be *approximately Normal* when n is large. Let us make a list of them:

1. **Binomial** (a sum of n RVs of Bernoulli type).

   It will acquire the Normal shape when the mean np is not too close to either the smallest possible value of 0 or the largest possible value of n. Ordinarily, it is sufficient to require that $np \geq 5$ and $nq \geq 5$ [note that, when $p = 0.5$, this is met by n as small as 10].

2. **Negative Binomial** (a sum of k Geometric-type RVs).

   Here the mean $\frac{k}{p}$ must be substantially bigger than the lowest possible value of k, they usually require $\frac{kq}{p} \geq 30$ (since the original distribution is not symmetric, we have to make the distance higher than in the Binomial case).

3. **Poisson** (can be seen as a sum of n Poisson-type RVs, each with the mean of $\frac{\lambda}{n}$).

   Here we require $\lambda \geq 30$ [clear of the smallest possible value of 0].

4. **Hypergeometric** becomes approximately Binomial when K and $N-K$ are large (hundreds). This in turn becomes Normal, when it meets the $np \geq 5$ and $nq \geq 5$ conditions.

   [Thus, practically all of our common discrete distributions become Normal in special circumstances].

5. And of course **any other** distribution (typically a game of some sort) which is sampled, independently, many times.

EXAMPLES:

1. Roll a die 100 times, what is the probability of getting more than 20 sixes?

   The total number of sixes obtained in this experiment is a RV (let us call it X) having the binomial distribution with $n = 100$ and $p = \frac{1}{6}$. Since $np = 16.\bar{6}$ and $nq = 83.\bar{3}$, our two criteria are amply met, and the distribution of X will closely follow the corresponding Normal curve.

   One has to realize that the exact distribution will always consist of individual rectangles [centered on positive integers] which can match the Normal (continuous) curve only at their mid points. Also, the probability of a specific value [e.g. $\Pr(X = 21)$] is represented by the *whole* area of the corresponding *rectangle* [which extends from 20.5 to 21.5]; the corresponding approximation [to $\Pr(X = 21)$] will be computed as the area under the Normal curve over the *same* interval [i.e. from 20.5 to 21.5]. This means that the following Binomial probability: $\Pr(X > 20) = \Pr(X = 21) + \Pr(X = 22) + \Pr(X + 23) + \ ...$ will be approximated by $\Pr(20.5 < X < 21.5) + \Pr(21.5 < X < 22.5) + \Pr(22.5 < X < 23.5) + \ ... = \Pr(20.5 < X)$, where $X \in \mathcal{N}(\frac{100}{6}, \sqrt{\frac{500}{36}})$ [Normal distribution with the *matching* mean of np and the standard deviation of $\sqrt{npq}$]. This adjustment of the integer value to the corresponding half-integer is called **continuity correction** [yet another one of those illogical names − the correction relates to being *discrete*].

   Solution: $\Pr(20 < X_{Binomial}) \approx \Pr(20.5 < X_{Normal}) = \Pr\left(\frac{20.5 - \frac{100}{6}}{\sqrt{\frac{500}{36}}} < \frac{X - \frac{100}{6}}{\sqrt{\frac{500}{36}}}\right) = \Pr(1.0285913 < Z) = 0.5000 - \Pr(0 < Z < 1.0285913) = 0.5000 - (0.3461 + 0.85913 \times 0.0024) = 15.18\%$ [coincidentally, the exact answer is 15.19% − in this case, one would expect to be within 0.5% of the exact answer].

2. If X has the Poisson distribution with $\lambda = 35.14$, approximate $\Pr(X \leq 30)$.

   Assuming that $X \tilde{\in} \mathcal{N}(35.14, \sqrt{35.14})$ [the same mean and σ] and using the 'continuity' correction, our probability $\approx$ $\Pr(X < 30.5)$ [30, 29, 28, ... are to be *included*] $= \Pr\left(\frac{X-35.14}{\sqrt{35.14}} < \frac{30.5-35.14}{\sqrt{35.14}}\right) = \Pr(Z < -0.782739) = \Pr(Z > 0.782739) = 0.5000 - (0.2823 + 0.2739 \times 0.0029) = 21.69\%$ [the exact answer is 22.00% − one has to add 31 individual Poisson probabilities to get it].

3. Consider rolling a die repeatedly until obtaining 100 sixes. What is the probability that this will happen in fewer than 700 rolls?

   Solution: The RV which counts the number of rolls (X say) has the Negative Binomial distribution with $p = \frac{1}{6}$ and $k = 100$. Since $\frac{qk}{p} = 500$, the Normal approximation is fully justified: $\Pr(X < 700)$ [i.e. 699, 698, 697,...0] $= \Pr(X_{Normal} < 699.5)$ [to include 699 but exclude 700] $= \Pr\left(\frac{X_N-600}{\sqrt{3000}} < \frac{699.5-600}{\sqrt{3000}}\right)$ [using $\frac{k}{p}$ and $\sqrt{\frac{k}{p}(\frac{1}{p}-1)}$ for the mean and standard deviation] $= \Pr(Z < 1.816613) = 0.5000 + (0.4649 + 0.6613 \times 0.007) = 96.54\%$ [the exact answer is 96.00%, but it takes a computer to evaluate it].

4. If 5 cards are deal [from a standard deck of 52] repeatedly and independently 100 times, what is the probability of dealing at least 50 aces in total?

   Solution: We need $\Pr(X_1 + X_2 + ... + X_{100} \geq 50)$, where the X_i's are independent, hypergeometric (with $N = 52$, $K = 4$ and $n = 5 \Rightarrow \mu = \frac{5}{13}$ and $\sigma = \sqrt{\frac{5}{13} \times \frac{12}{13} \times \frac{47}{51}}$ each). Adding, independently 100 of them justifies the Normal approximation [let us introduce $Y \equiv \sum_{i=1}^{100} X_i$ where $\mu_Y = \frac{500}{13} = 38.462$ and $\sigma_Y = \sqrt{\frac{500}{13} \times \frac{12}{13} \times \frac{47}{51}} = 5.7200$], thus: $\Pr(Y \geq 50)$ [i.e. 50, 51, 52,...] $= \Pr\left(\frac{Y-38.462}{5.7200} > \frac{49.5-38.462}{5.7200}\right) = \Pr(Z > 1.9297) = 0.5000 - (0.4726 + 0.97 \times 0.0006) = 2.68\%$ [the exact answer is 3.00%].

5. Consider a random independent sample of size 200 form the uniform distribution $\mathcal{U}(0,1)$. Find $\Pr(0.49 \leq \bar{X} \leq 0.51)$.

   Solution: We know that $\bar{X} \tilde{\in} \mathcal{N}\left(0.5, \sqrt{\frac{1}{12 \times 200}}\right) \Rightarrow \Pr\left(\frac{0.49-0.5}{\sqrt{\frac{1}{2400}}} < \frac{\bar{X}-0.5}{\sqrt{\frac{1}{2400}}} < \frac{0.51-0.5}{\sqrt{\frac{1}{2400}}}\right) = \Pr(-0.4899 < Z < 0.4899) = 2 \times 0.1879 = 37.58\%$. Since the uniform distribution is continuous, no 'continuity' correction was required. [The exact answer is 37.56%; the approximation is now a lot more accurate for two reasons: the uniform distribution is continuous and *symmetric*].

6. Consider a random independent sample of size 100 from

$X =$	-1	0	1	2
Prob:	$\frac{3}{6}$	$\frac{2}{6}$	0	$\frac{1}{6}$

   What is the probability that the sample *total* will be negative (losing money, if this represents a game)?

Solution: First we compute a single-X mean and variance: $\mathbb{E}(X) = \frac{-3+2}{6} = -\frac{1}{6}$ and $Var(X) = \frac{3+4}{6} - \frac{1}{36} = \frac{41}{36}$, then we introduce $Y = \sum_{i=1}^{100} X_i$ and find $\mu_Y = -\frac{100}{6}$ and $\sigma_Y = \sqrt{\frac{4100}{36}}$. Note that Y can have [within certain limits] any integer value.

Answer: $\Pr(Y < 0)$ [i.e. -1, -2, -3, ...] $\approx \Pr\left(\frac{Y_{Normal} + \frac{100}{6}}{\sqrt{\frac{4100}{36}}} < \frac{-0.5+\frac{100}{6}}{\sqrt{\frac{4100}{36}}}\right)$ [continuity correction applied] $= \Pr(Z < 1.5149) = 0.5000 + 0.4345 + 0.49 \times 0.0012 = 93.51\%$ [probability of losing money after 100 rounds of the game; in each round the probability of losing money is only 50%]. The exact computation would yield 93.21%.

7. Pay \$10 to play the following game: 5 cards are dealt from a standard deck, and you receive \$10 for each ace and \$5 for each king, queen and jack. Find:

   (a) The expected value and standard deviation of your net win.
   Solution: $W = 10X + 5Y - 10$ (where X is the number of aces dealt; Y correspondingly counts the total of kings, queens and jacks). $\mathbb{E}(W) = 10 \times \frac{5}{13} + 5 \times \frac{5\times 3}{13} - 10 = -\frac{5}{13}$ dollars (≈ -38 cents) and $Var(W) = 5(10^2 \times \frac{1}{13} \times \frac{12}{13} + 5^2 \times \frac{3}{13} \times \frac{10}{13} - 2 \times 10 \times 5 \times \frac{1}{13} \times \frac{3}{13})\frac{47}{51} = 44.988 \Rightarrow \sigma_W = \sqrt{44.988} = \$6.7073$.

   (b) The probability of losing more that \$50 after 170 rounds of this game.
   Solution: $\Pr(W_1 + W_2 + ... + W_{170} < -50)$ amply justifies using the Normal approximation. Defining $T \equiv \sum_{i=1}^{170} W_i$ with $\mu_T = -\frac{170\times 5}{13} = -65.385$ and $\sigma_T = 6.7073 \times \sqrt{170} = 87.452$ we get: $\Pr(T < -50)$ [i.e. -55, -60, -65, ...] $\approx \Pr(T_{Normal} < -52.5)$ [note the unusual 'continuity' correction] $= \Pr(\frac{T_N + 65.385}{87.452} < \frac{-52.5+65.385}{87.452}) = \Pr(Z < 0.14734) = 0.5000 + 0.0557 + 0.734 \times 0.0042 = 55.88\%$. The exact answer is 56.12% (we are a quarter percent off, which is not bad, considering that the individual probabilities of a specific loss − say \$50 − are well over 2%). ■

▶Exponential◀

distribution relates to the 'fishing' experiment, where φ (say 1.2/hour) is the rate at which the fishes are caught and $\beta = \frac{1}{\varphi}$ (50 min.) is the average time to catch a fish. This time the random variable (as always, we call it X) is the length of time from the beginning of the experiment until the *first* catch.

To find its pdf, we start by subdividing each hour into N subintervals (of, say, 60 minutes), assuming that the probability of a catch during any one of these is $p = \frac{\varphi}{N}$ (2%) and using the *geometric* distribution to approximate the answer (note that p has been adjusted to correspond to the average catch of φ fishes per hour). At the same time, this necessitates introducing a new RV Y which (unlike X) measures time till the first catch in number of subintervals.

This model has the shortcoming of not being able to prevent catching more than one fish during any such subinterval (which invalidates our assumption of

dealing with Bernoulli-type trials − this was discussed when introducing Poisson distribution two chapters ago). One thus obtains the correct answer only in the $N \to \infty$ limit.

To be able to take this limit properly, we first construct $\Pr(X < x)$, where x must be measured in proper time units (hours), and not in the [ultimately meaningless] subinterval count. Based on the geometric distribution, we know that $\Pr(Y \leq i) = 1 - q^i \equiv 1 - (1 - \frac{\varphi}{N})^i$, where $i \geq 0$. Since the real time x equals to $i \cdot \frac{1}{N}$, we replace i of the previous formula by xN to obtain: $\Pr(Y \leq xN) = \Pr(X \leq x) = 1 - (1 - \frac{\varphi}{N})^{Nx} \underset{n\to\infty}{\longrightarrow} 1 - e^{-\varphi x}$ where $x > 0$.

The last expression represents the **distribution function** of X; we prefer to use β as the basic parameter, thus:

$$F(x) \equiv \Pr(X \leq x) = 1 - e^{-\frac{x}{\beta}}$$

when $x > 0$. Note that $\Pr(X > x) = e^{-\frac{x}{\beta}} \Rightarrow$ [with $\beta = 50$ min.], $\Pr(X > 1$ hr.$) = e^{-\frac{60}{50}}$ [same units must be used] $= 30.12\%$, $\Pr(X > 3$ hr.$) = e^{-\frac{180}{50}} = 2.73\%$, etc. The $\mu \pm \sigma$ interval now contains $1 - e^{-2} \equiv \Pr(X < 2\beta) = 86.47\%$ of the whole distribution.

Based on the distribution function we can easily derive the corresponding **pdf**:

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{\beta} e^{-\frac{x}{\beta}}$$

for $x > 0$ and **MGF**: $M(t) = \frac{1}{\beta} \int_0^\infty e^{-\frac{x}{\beta}} \cdot e^{tx}\, dx = \frac{1}{\beta} \int_0^\infty e^{-x(\frac{1}{\beta} - t)}\, dx = \frac{1}{\beta\,(\frac{1}{\beta} - t)} = \frac{1}{1 - \beta t}$. Expanding this results in $1 + \beta t + \beta^2 t^2 + \beta^3 t^3 + \beta^4 t^4 + ...$, which immediately yields all simple **moments** of the distribution. Based on these, we construct: $\mathbb{E}(X) = \beta$ [expected, i.e. average time to catch one fish], $Var(X) = 2\beta^2 - \beta^2 = \beta^2$ ($\Rightarrow \sigma_X = \beta$, don't confuse with the *Poisson* result of $\sqrt{\lambda}$), skewness $= \frac{6\beta^3 - 3\times 2\beta^3 + 2\beta^3}{\beta^3} = 2$, and kurtosis $= \frac{24\beta^4 - 4\times 6\beta^4 + 6\times 2\beta^4 - 3\beta^4}{\beta^4} = 9$.

The exponential distribution shares, with the geometric distribution (from which it was derived), the '**memory-less**' property of $\Pr(X - a > x | X > a) = \Pr(X > x)$, i.e. given we have been fishing, *unsuccessfully*, for time a, the probability that the first catch will take longer than x (from *now*) is the same for all values of a [including 0, i.e. we are no closer to it than someone who has just started − this is similar to rolling a die to get the first six].

Proof: $\Pr(X - a > x | X > a) = \frac{\Pr(X > x + a \cap X > a)}{\Pr(X > a)} = \frac{\Pr(X > x + a)}{\Pr(X > a)} = \frac{e^{-\frac{x+a}{\beta}}}{e^{-\frac{a}{\beta}}} = e^{-\frac{x}{\beta}} = \Pr(X > x)$. □

The exponential distribution is a good description of how long it takes to catch the first fish, but also, *from then on,* the second fish, the third fish, etc. [furthermore, these random time intervals are independent of each other; this again is inherited from the geometric distribution].

The potential **applications** of this distribution include: time intervals between consecutive phone calls, accidents, customer-arrivals, etc. (all those discussed in connection with the Poisson distribution).

$\mathcal{E}(\beta)$ will be our symbolic **notation** for the exponential distribution with the mean of β.

The Median

of a continuous distribution is the number which, when a sample is drawn, will be *exceeded* with a *50% probability;* consequently, a smaller result is obtained with the complementary probability of 50%, i.e. the median divides the distribution in two *equally probable halves*. Mathematically, we can find it is a solution to $F(\tilde{\mu}) = \frac{1}{2}$, or (equivalently) to $1 - F(\tilde{\mu}) = \frac{1}{2}$ [$\tilde{\mu}$ will be our usual **notation** for the median]. For a symmetric distribution (uniform, Normal) the mean and median must be both at the CENTER OF SYMMETRY [yet, there is an important distinction: the mean may not always exit, the median always does]. ■

The **median** of the *exponential* distribution is thus the solution to $e^{-\frac{\tilde{\mu}}{\beta}} = \frac{1}{2} \Leftrightarrow \frac{\tilde{\mu}}{\beta} = \ln 2$, yielding $\tilde{\mu} = \beta \ln 2$ $(= 0.6931\beta)$ [substantially smaller that the corresponding *mean*]. This means that if it takes, on the average, 1 hour to catch a fish, 50% of all fishes are caught in less than 41 min. and 35 sec. [you should not see this as a contradiction].

Smallest-value distribution:

To conclude our discussion of the exponential distribution we discuss a problem which will give us a head start on the topic of the next chapter, namely: taking several RVs and combining them to define (usually by some mathematical formula) a new RV whose distribution is then to be found:

Suppose there are n fishermen at some lake fishing independently of each other. What is the distribution of the time of their **first catch** (as a **group**)? We will assume that their individual times to catch a fish are exponential, with the same value of β.

Solution: If X_1, X_2,....X_n are the individual times (of the first catch), then $Y \equiv \min(X_1, X_2, ...X_n)$ is the new RV of the question. We can find $\Pr(Y > y) = \Pr(X_1 > y \cap X_2 > Y \cap .... \cap X_n > y) = \Pr(X_1 > y) \cdot \Pr(X_2 > y) \cdot .... \cdot \Pr(X_n > y) = \left(e^{-\frac{y}{\beta}}\right)^n = e^{-\frac{n\,y}{\beta}} \Rightarrow F_Y(y) = 1 - e^{-\frac{n\,y}{\beta}}$ for $y > 0$. This clearly identifies the distribution of Y as *exponential* with the mean (β_{NEW} say) equal to $\frac{\beta}{n}$ [things happen n times faster now]. This could have been done (equally easily, try it) using distinct [individual] values of β [β_1, β_2, ...β_n]; the result would have been an exponential distribution with $\beta_{NEW} = \frac{1}{\frac{1}{\beta_1} + \frac{1}{\beta_2} + ... + \frac{1}{\beta_n}}$ [simpler in terms of rates: $\varphi_{NEW} = \varphi_1 + \varphi_2 + ... + \varphi_n$].

►Gamma◄

distribution relates to the previous 'fishing' experiment. The corresponding RV X_G is **defined** as the time of the k^{th} catch (of a single fisherman), and is obviously

equal to an independent sum of k RVs $X_1 + X_2 + ... + X_k$, all of them exponential with the same mean β.

The **parameters** of this distribution are k and β, the symbolic **notation** for this distribution is $\gamma(k, \beta)$. Since X_G was defined as an independent sum of RVs of a known (exponential) type, we know immediately its **mean**: $\mathbb{E}(X_G) = k\beta$, **variance**: $Var(X_G) = k\beta^2$ [$\Rightarrow \sigma = \sqrt{k}\beta$] and **MGF**: $M(t) = \frac{1}{(1-\beta\, t)^k}$.

We would now like to derive the corresponding **pdf** and $F(x)$, which is a challenging task as we have not yet learned how to *add* RVs (in terms of their pdf). We must proceed indirectly, by going back to the Poisson distribution and recalling that we do know how to find, for a fixed time period x, the probability of achieving fewer than k catches. This of course equals to

$$e^{-\lambda}[1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + ... + \frac{\lambda^{k-1}}{(k-1)!}] \qquad (1 - F)$$

where $\lambda = \frac{x}{\beta}$. In terms of our X_G, this represents the probability that the k^{th} catch will take longer than x, namely: $\Pr(X_G > x) = 1 - F(x)$, yielding the corresponding distribution function. From this, we can easily derive $f(x)$ by a simple x-differentiation: $\frac{dF(x)}{dx} = \{e^{-\lambda}[1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + ... + \frac{\lambda^{k-1}}{(k-1)!}] - e^{-\lambda}[1 + \lambda + \frac{\lambda^2}{2} + ... + \frac{\lambda^{k-2}}{(k-2)!}]\} \cdot \frac{1}{\beta}$ [the chain-rule] $= e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} \cdot \frac{1}{\beta} =$

$$\frac{x^{k-1}e^{-\frac{x}{\beta}}}{\beta^k\,(k-1)!} \qquad (f)$$

for $x > 0$.

We can verify the correctness of this answer by computing the corresponding **MGF**: $M(t) = \frac{1}{\beta^k\,(k-1)!}\int\limits_0^\infty x^{k-1}e^{-\frac{x}{\beta}} \cdot e^{xt}\,dx = \frac{1}{\beta^k\,(k-1)!}\int\limits_0^\infty x^{k-1}e^{-x(\frac{1}{\beta}-t)}\,dx = \frac{1}{\beta^k\,(k-1)!} \cdot \frac{(k-1)!}{(\frac{1}{\beta}-t)^k}$ [using the general formula $\int\limits_0^\infty x^{n-1}e^{-\frac{x}{a}}\,dx = (n-1)!a^n$ developed earlier] $= \frac{1}{(1-\beta\, t)^k}$ (check).

You should be able to *identify* the pdf of the gamma distribution when you see one (based on the numerator alone, the denominator is just an automatic normalizing constant).

EXAMPLES:

1. If $X \in \gamma(4, 20 \text{ min.})$, find:

   (a) $\Pr(X < 30 \text{ min.}) = 1 - e^{-\frac{30}{20}}[1 + \frac{30}{20} + \frac{(\frac{30}{20})^2}{2} + \frac{(\frac{30}{20})^3}{3!}] = 6.56\%$ [catching 4 fishes in less than half an hour].

   (b) $\Pr(X > 2 \text{ hr.}) = e^{-\frac{120}{20}}[1 + 6 + \frac{6^2}{2} + \frac{6^3}{6}] = 15.12\%$ [make sure to use, consistently, either minutes or hours].

2. A fisherman whose average time for catching a fish is 35 minutes wants to bring home exactly 3 fishes. What is the probability he will need between 1 and 2 hours to catch them?

Solution: $\Pr(1 \text{ hr.} < X < 2 \text{ hrs}) = F(120 \text{ min.}) - F(60 \text{ min.}) = e^{-\frac{60}{35}}[1 + \frac{60}{35} + \frac{(\frac{60}{35})^2}{2}] - e^{-\frac{120}{35}}[1 + \frac{120}{35} + \frac{(\frac{120}{35})^2}{2}] = 41.92\%$

3. If a group of 10 fishermen goes fishing, what is the probability that the *second* catch of the *group* will take less than 5 min. Assume the value of $\beta = 20$ min. for *each* fisherman; also assume that the one who catches the first fish *continues* fishing.

   Solution: This is equivalent to having a single 'super'fisherman who catches fish at a 10 times faster rate, i.e. $\beta_{NEW} = 2$ min.

   Answer: $1 - e^{-\frac{5}{2}}[1 + \frac{5}{2}] = 71.27\%$. ■

It is possible to **extend** the definition of the gamma distribution and allow k to be a positive *non-integer* (we will call it α, to differentiate). The expression for $f(x)$ is still legitimate if we replace $(k-1)! \equiv \int_0^\infty x^{k-1}e^{-x}\,dx$ by $\int_0^\infty x^{\alpha-1}e^{-x}\,dx$ which, by definition, is called $\Gamma(\alpha)$ [THE GAMMA FUNCTION]. One can easily prove (integrate by parts) that

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

We normally need the values of $\Gamma(\alpha)$ with half-integer arguments only. With the help of $\Gamma(\frac{1}{2}) = \int_0^\infty x^{-\frac{1}{2}}e^{-x}\,dx = \int_0^\infty \sqrt{\frac{2}{z^2}}e^{-\frac{z^2}{2}}\,z\,dz$ [substitute $x = \frac{z^2}{2}$] $= \sqrt{2}\int_0^\infty e^{-\frac{z^2}{2}}\,dz$ [remember the standardized normal pdf and its normalizing constant] $= \sqrt{2} \cdot \frac{\sqrt{2\pi}}{2} = \sqrt{\pi}$, and using the previous formula, we can deal with any half-integer argument.

EXAMPLE: $\Gamma(\frac{5}{2}) = \frac{3}{2} \times \Gamma(\frac{3}{2}) = \frac{3}{2} \times \frac{1}{2} \times \Gamma(\frac{1}{2}) = \frac{3\sqrt{\pi}}{4}$ ■

Using this **generalized** pdf

$$f(x) = \frac{x^{\alpha-1}\,e^{-\frac{x}{\beta}}}{\beta^\alpha\,\Gamma(\alpha)}$$

when $x > 0$, one obtains the old (after the $k \to \alpha$ replacement) formula for the **mean** $(= \alpha\beta)$, **variance** $(= \alpha\beta^2)$ and the **MGF**: $\frac{1}{(1-\beta\,t)^\alpha}$ [the proof would be quite simple]; there is *no* simple extension of $F(x)$ [we cannot have half-integer number of terms]. Without $F(x)$, we don't know how to compute probabilities; this is a serious shortcoming to be corrected in the next chapter.

Finally, since the gamma-type RV was introduced as a sum of k exponential RVs, we know that, when k is large (say bigger than 30), the gamma distribution is also **approximately Normal** (with the mean of $k\beta$ and the standard deviation of $\sqrt{k}\beta$).

EXAMPLE: Phone calls arrive at the rate of 12.3/hour. What is the probability that the 50^{th} phone call will arrive after 1 p.m. if the office opens at 8 a.m.

1. Solution: If X is the time of the arrival of the 50^{th} phone call, in hours (setting our stop watch to 0 at 8 a.m.) we have to find $\Pr(X > 5 \text{ hr.})$. The distribution of X is $\gamma(50, \frac{1}{12.3} \text{ hr.})$ which implies that the exact answer would require using a 49-term formula (too long). We are thus forced to apply the Normal approximation: $\Pr\left(\frac{X-\frac{50}{12.3}}{\frac{\sqrt{50}}{12.3}} > \frac{5-\frac{50}{12.3}}{\frac{\sqrt{50}}{12.3}}\right) = \Pr(Z > 1.6263) = 5.19\%$.

   Alternate solution: We can also introduce Y as the number of phone calls received during the 8 a.m.-1 p.m. time interval. Its distribution is Poisson, with $\lambda = 5 \times 12.3 = 61.5$. The question can be reformulated as $\Pr(Y < 50)$ which, when evaluated exactly, requires the same 49-term formula as the previous approach. Switching to Normal approximation gives: $\Pr(\frac{Y-61.5}{\sqrt{61.5}} < \frac{49.5-61.5}{\sqrt{61.5}})$ [this time we needed the 'continuity' correction] $= \Pr(Z < -1.5302) = 6.30\%$.

   (The exact answer, obtained with a help of a computer, is 5.91%). ■

In the next chapter we introduce some more distributions of the continuous type, namely 'beta', χ^2 [read 'chi-squared'], t and F [another unfortunate name, in conflict with our $F(x)$ notation].

Bivariate (two RVs) case

This time we discuss only one (the most important) case of the **(bivariate) Normal** distribution:

Its importance follows from the following

►(Bivariate) Extension of the Central Limit Theorem◄

Consider a random independent sample of size n from a bivariate distribution of two RVs X and Y, say. We know that, *individually*, both $Z_X \equiv \frac{X_1+X_2+....+X_n-n\mu_X}{\sqrt{n}\sigma_X}$ and $Z_Y \equiv \frac{Y_1+Y_2+....+Y_n-n\mu_Y}{\sqrt{n}\sigma_Y}$ have (in the $n \to \infty$ limit) the standardized normal distribution $\mathcal{N}(0,1)$. We would like to know what happens to them *jointly*.

To investigate this, we have to introduce a new concept of **joint MGF** of X and Y, thus: $M_{X,Y}(t_1, t_2) = \mathbb{E}(e^{t_1 X + t_2 Y}) = \iint_{\mathcal{R}} e^{t_1 x + t_2 y} \cdot f(x,y)\, dx\, dy$. One can show that, when expanded in both t_1 and t_2 [generalized Taylor series], one gets:

$$M(t_1, t_2) = 1 + \mu_X t_1 + \mu_Y t_2 + \mathbb{E}(X^2)\frac{t_1^2}{2} + \mathbb{E}(Y^2)\frac{t_2^2}{2} + \mathbb{E}(X \cdot Y)\, t_1 t_2 + ...$$

[dots representing terms involving third and higher moments].

Based on such joint MGF, it is quite trivial to obtain the **marginal MGF** of X, thus: $M_X(t_1) = M_{X,Y}(t_1, t_2 = 0)$ [i.e. substitute zero for t_2], and similarly for the Y-marginal.

The **MGF** of a **linear combination** of X and Y can be derived equally easily:

$$M_{aX+bY+c}(t) = e^{ct} \cdot M_{X,Y}(t_1 = at, t_2 = bt)$$

When X and Y are independent of U and V, the MGF of the $X+U$ and $Y+V$ pair is: $M_{X+U,Y+V}(t_1, t_2) = M_{X,Y}(t_1, t_2) \cdot M_{U,V}(t_1, t_2)$. A special case of this arises

when U and V have also the same distribution as X and Y (we will rename the two pairs X_2 and Y_2, and X_1 and Y_1, respectively). Then, the MGF of the X_1+X_2 and Y_1+Y_2 pair is simply $[M_{X,Y}(t_1,t_2)]^2$. And finally, if (X_1,Y_1), (X_2,Y_2), ..., (X_n,Y_n) is a random independent sample of size n from a bivariate distribution whose MGF is $M(t_1,t_2)$, the MGF of the $X_1+X_2+...+X_n$ and $Y_1+Y_2+...+Y_n$ pair is

$$[M(t_1,t_2)]^n$$

Proof: $\mathbb{E}\left\{\exp\left[t_1(X_1+X_2+...+X_n)+t_2(Y_1+Y_2+...+Y_n)\right]\right\} =$
$\mathbb{E}\left[\exp(t_1X_1+t_2Y_1)\cdot\exp(t_1X_2+t_2Y_2)\cdot...\cdot\exp(t_1X_n+t_2Y_n)\right] =$
$\mathbb{E}\left[\exp(t_1X_1+t_2Y_1)\right]\cdot\mathbb{E}\left[\exp(t_1X_2+t_2Y_2)\right]\cdot...\cdot\mathbb{E}\left[\exp(t_1X_n+t_2Y_n)\right] =$
$M(t_1,t_2)\cdot M(t_1,t_2)\cdot...\cdot M(t_1,t_2) = [M(t_1,t_2)]^n$ $\square$

We would like to investigate now what happens to the joint MGF of $\frac{X_1+X_2+....+X_n-n\mu_X}{\sqrt{n}\sigma_X}\equiv Z_X$ and $\frac{Y_1+Y_2+....+Y_n-n\mu_Y}{\sqrt{n}\sigma_Y}\equiv Z_Y$ in the $n\to\infty$ **limit**. By the previous formula [applied to $\left(\frac{X_i-\mu_X}{\sqrt{n}\sigma_X},\frac{Y_i-\mu_Y}{\sqrt{n}\sigma_Y}\right)$, rather than the original (X_i,Y_i) pairs], we know that

$$M_{Z_X,Z_Y}(t_1,t_2) = \left[M_{\frac{X-\mu_X}{\sqrt{n}\sigma_X},\frac{Y-\mu_Y}{\sqrt{n}\sigma_Y}}(t_1,t_2)\right]^n$$

Since $M_{\frac{X-\mu_X}{\sqrt{n}\sigma_X},\frac{Y-\mu_Y}{\sqrt{n}\sigma_Y}}(t_1,t_2)$ can be expanded to yield $1+\frac{t_1^2}{2n}+\frac{t_2^2}{2n}+\frac{\rho}{n}t_1t_2+...$[terms with $n^{\frac{3}{2}}$ and higher powers of n in the denominator], we need to take the limit of $\left(1+\frac{t_1^2}{2n}+\frac{t_2^2}{2n}+\frac{\rho}{n}t_1t_2+...\right)^n$ as $n\to\infty$. This of course equals to

$$\exp\left(\frac{t_1^2+2\rho t_1t_2+t_2^2}{2}\right)$$

where ρ is the correlation coefficient between X and Y of the original (sampling) distribution.

Having thus found the MGF limit, we would like to find the corresponding **bivariate pdf** [both of its marginals will have to be *standardized Normal*]. One can show that

$$f(z_1,z_2)=\frac{1}{2\pi\sqrt{1-\rho^2}}\cdot\exp\left[-\frac{z_1^2-2\rho z_1z_2+z_2^2}{2(1-\rho^2)}\right]$$

where both z_1 and z_2 are arbitrary real numbers, is the desired answer.

Proof: To verify that this is a legitimate pdf we evaluate: $\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\exp\left[-\frac{z_1^2-2\rho z_1z_2+z_2^2}{2(1-\rho^2)}\right]dz_1dz_2 =$
$\int\limits_{-\infty}^{\infty}\exp\left(-\frac{z_2^2}{2}\right)\int\limits_{-\infty}^{\infty}\exp\left[-\frac{(z_1-\rho z_2)^2}{2(1-\rho^2)}\right]dz_1dz_2 = \sqrt{2\pi}\cdot\sqrt{2\pi(1-\rho^2)} = 2\pi\sqrt{1-\rho^2}$
(check).

To prove that this pdf results in the correct MGF, we need: $\frac{1}{2\pi\sqrt{1-\rho^2}}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\exp(t_1z_1+t_2z_2)\cdot\exp\left[-\frac{z_1^2-2\rho z_1z_2+z_2^2}{2(1-\rho^2)}\right]dz_1dz_2 = \frac{1}{2\pi\sqrt{1-\rho^2}}\cdot\exp\left(\frac{t_1^2+2\rho t_1t_2+t_2^2}{2}\right)\int\limits_{-\infty}^{\infty}\exp\left(-\frac{[z_2-(t_2+\rho t_1)]^2}{2}\right)$

$\int_{-\infty}^{\infty} \exp\left(-\frac{\{z_1-[t_1(1-\rho^2)+\rho z_2]\}^2}{2(1-\rho^2)}\right) dz_1 dz_2 = \exp\left(\frac{t_1^2+2\rho t_1 t_2+t_2^2}{2}\right)$ (check). Confirming that the two [total] exponents (of the first step) agree follows from some simple [even though tedious] algebra. □

The **graph** of this pdf is a Normal 'hill', stretched in the $z_1 = z_2$ ($z_1 = -z_2$) direction when $\rho > 0$ ($\rho < 0$). This hill becomes a narrow 'ridge' when $|\rho|$ approaches 1, acquiring a zero thickness (and infinite height − its volume is fixed) at $|\rho| = 1$. For $\rho = 0$ the hill is perfectly round, and the corresponding Z_1 and Z_2 are independent of each other [for the bivariate Normal distribution independence implies zero correlation *and reverse*].

The related **conditional** pdf of $Z_1|Z_2 = \mathbf{z}_2$ can be derived based on the standard prescription: $\frac{f(z_1,\mathbf{z}_2)}{f_{Z_2}(\mathbf{z}_2)} = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{z_1^2-2\rho z_1\mathbf{z}_2+\mathbf{z}_2^2}{2(1-\rho^2)}\right] \div \frac{\exp\left(-\frac{\mathbf{z}_2^2}{2}\right)}{\sqrt{2\pi}} =$

$$\frac{\exp\left[-\frac{(z_1-\rho\mathbf{z}_2)^2}{2(1-\rho^2)}\right]}{\sqrt{2\pi}\sqrt{1-\rho^2}}$$

where $-\infty < z_1 < \infty$. This conditional distribution can be easily identified as $\mathcal{N}\left(\rho\mathbf{z}_2, \sqrt{1-\rho^2}\right)$ [try to visualize the corresponding cross-section of the hill].

This concludes our discussion of the **standardized** (in terms of its marginals) bivariate **Normal** distribution. Now, we need to extend the results to cover the so called

▶General Bivariate Normal Distribution◀

It is the distribution of $U \equiv \rho_1 Z_1 + \mu_1$ and $V \equiv \sigma_2 Z_2 + \mu_2$, where Z_1 and Z_2 are the RVs of the previous section. This implies that the new bivariate pdf (quoted below) will have exactly the same shape and properties as the old one, only its two horizontal axes (called u and v now) will be 're-scaled'.

We already know that, individually, U and V will be (univariate) Normal, $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$, respectively. We also know that correlation coefficient of the U, V pair will be the same as that of Z_1 and Z_2 (linear transformation does not change its value).

The **joint pdf** of U and V can be derived from that of Z_1 and Z_2 by the following replacement: $z_1 \to \frac{u-\mu_1}{\sigma_1}$, $z_2 \to \frac{v-\mu_2}{\sigma_2}$, $dz_1 \to \frac{du}{\sigma_1}$ and $dz_2 \to \frac{dv}{\sigma_2}$, thus:

$$f(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2} \exp\left[-\frac{(\frac{u-\mu_1}{\sigma_1})^2 - 2\rho(\frac{u-\mu_1}{\sigma_1})(\frac{v-\mu_2}{\sigma_2}) + (\frac{v-\mu_2}{\sigma_2})^2}{2(1-\rho^2)}\right] (du\, dv)$$

for any u and v. The resulting expression is obviously a lot more complicated (and clumsier to use) than old one (for Z_1 and Z_2). This is the reason why, whenever we have to deal with the generalized Normal distribution, we usually **convert** the corresponding RVs **to** the standardized Z_1 and Z_2 (by a simple linear transformation), and answer the original question in terms of these.

Only when dealing with **conditional** probabilities, we bypass this approach (introducing Z_1 and Z_2), and proceed as follows: Assume that U and V are bivariate Normal RVs, with the mean of μ_1 and μ_2 and the standard deviation of σ_1 and σ_2, respectively, and the correlation coefficient of ρ (this is a *5-parameter distribution*). We know that we can express U as $\sigma_1 Z_1 + \mu_1$ and V as $\sigma_2 Z_2 + \mu_2$, where Z_1 and Z_2 are *standardized* Normal and have the same correlation coefficient of ρ.

We want to derive the conditional distribution of V ($\equiv \sigma_2 Z_2 + \mu_2$), given $U = \mathbf{u}$ ($\Leftrightarrow \; Z_1 = \frac{\mathbf{u}-\mu_1}{\sigma_1}$) or, symbolically: Distr$(\sigma_2 Z_2 + \mu_2 | Z_1 = \frac{\mathbf{u}-\mu_1}{\sigma_1})$.

We know that Distr$(Z_2 | Z_1 = \frac{\mathbf{u}-\mu_1}{\sigma_1})$ is $\mathcal{N}(\rho\frac{\mathbf{u}-\mu_1}{\sigma_1}, \sqrt{1-\rho^2})$ [from the previous paragraph], which implies that Distr$(\sigma_2 Z_2 + \mu_2 | Z_1 = \frac{\mathbf{u}-\mu_1}{\sigma_1})$ is $\mathcal{N}(\mu_2 + \sigma_2\rho\frac{\mathbf{u}-\mu_1}{\sigma_1}, \sigma_2\sqrt{1-\rho^2})$ [by our linear-transformation formulas].

This lead to the following conclusion:

$$\text{Distr}(V|U=\mathbf{u}) \equiv \mathcal{N}(\mu_2 + \sigma_2\rho\frac{\mathbf{u}-\mu_1}{\sigma_1}, \sigma_2\sqrt{1-\rho^2})$$

And of course, the other way around:

$$\text{Distr}(U|V=\mathbf{v}) \equiv \mathcal{N}(\mu_1 + \sigma_1\rho\frac{\mathbf{v}-\mu_2}{\sigma_2}, \sigma_1\sqrt{1-\rho^2})$$

[same proof].

The **joint MGF** of U and V is, by the linear-combination formula of the last section

$$M(t_1,t_2) = \exp(\mu_1 t_1 + \mu_2 t_2) \cdot \exp\left(\frac{\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2}\right)$$

This implies that any linear combination of U and V remains Normal.

Proof: $M_{aU+bV+c}(t) = \exp(ct)\cdot\exp(\mu_1 at + \mu_2 bt)\cdot\exp\left[(\sigma_1^2 a^2 + 2\rho\sigma_1\sigma_2 ab + \sigma_2^2 b^2)\frac{t^2}{2}\right]$, which can be identified as the MGF of $\mathcal{N}\left(a\mu_1 + b\mu_2 + c, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_1\sigma_2\rho}\right)$. □

EXAMPLES:

1. X and Y are jointly Normal with $\mu_x = 13.2$, $\sigma_x = 5.4$, $\mu_y = 136$, $\sigma_y = 13$ and $\rho = -0.27$ [the names and notation may vary from case to case – we are no longer calling them U and V]. Find:

   (a) $\mathbb{E}(X|Y=150)$ and $Var(X|Y=150)$.
   Solution: Our formulas clearly imply that $\mathbb{E}(X|Y=150) = \mu_x + \sigma_x\rho\frac{150-\mu_y}{\sigma_y} = 13.2 - 5.4 \times 0.27 \times \frac{150-136}{13} = 11.63$ and $Var(X|Y) = \sigma_x^2(1-\rho^2) = 5.4^2 \times (1-0.27^2) = 27.034$.

(b) $\Pr(X > 15|Y = 150)$.

Solution: We know that the corresponding conditional distribution is $\mathcal{N}(11.63, \sqrt{27.034}) \Rightarrow \Pr(\frac{X-11.63}{\sqrt{27.034}} > \frac{15-11.63}{\sqrt{27.034}}|Y = 150) = \Pr(Z > 0.64815) = 0.5000 - (0.2389 + 0.815 \times 0.0033) = 25.84\%$ [Note that unconditionally, i.e. not knowing the value of Y, $\Pr(X > 15) = \Pr(\frac{X-13.2}{5.4} > \frac{15-13.2}{5.4}) = \Pr(Z > 0.\bar{3}) = 36.94\%$].

2. The conditional probabilities often arise in the context of CLT: An airline knows (based on their extensive past records) that their passengers travel (in a single trip) 1275±894 miles (the average and the standard deviation) and bring 32±17.3 lb. of check-in luggage. The correlation coefficient between the two variables is 0.22. A plane has been booked by 300 passengers who will travel (for many, this is just a connecting flight) the (sample) average of 1410 miles. What is the probability that the total weight of their check-in luggage will not exceed 10,000 lb.?

Solution: Let us call the individual distances these 300 people travel X_1, X_2, ..., X_{300} (we will assume that these are independent − no couples or families), and the corresponding weight of their luggage Y_1, Y_2, ..., Y_{300}. The question is: Find $\Pr(\sum_{i=1}^{300} Y_i < 10000|\bar{X} = 1410)$.

We know that $\bar{X}$ and $\sum_{i=1}^{300} Y_i$ have, to a good approximation, the bivariate Normal distribution [CLT − note that the distribution of the individual Xs and Ys in anything but Normal] with the mean of 1275 and 300×32 (respectively), the standard deviation of $\frac{894}{\sqrt{300}}$ and $17.3 \times \sqrt{300}$ (respectively), and the correlation coefficient of 0.22. This implies that the corresponding conditional distribution is $\mathcal{N}(300 \times 32 + 17.3 \times \sqrt{300} \times 0.22 \times \frac{1410-1275}{\frac{894}{\sqrt{300}}}, 17.3 \times \sqrt{300} \times \sqrt{1 - 0.22^2}) \equiv \mathcal{N}(9772.4, 292.3)$.

Answer: $\Pr(Z < \frac{10000-9772.4}{292.3}) = \Pr(Z < 0.77865) = 78.18\%$.

Note that unconditionally [not knowing anything yet about their length of travel] $\Pr(\sum_{i=1}^{300} Y_i < 10000) = \Pr\left(\frac{\sum_{i=1}^{300} Y_i - 300 \times 32}{17.3 \times \sqrt{300}} < \frac{10000-300\times32}{17.3\times\sqrt{300}}\right) = \Pr(Z < 1.3349) = 90.90\%$. ■

►End-of-Section Examples◄

1. Suppose that (X, Y) are coordinates of an impact of a bullet in a target plane (the target itself being 20cm×20cm in size, centered on the origin). It is a good model to assume that X and Y are bivariate Normal; furthermore, in this particular case $\mu_x = 3$cm, $\mu_y = -2$cm [our rifle's sights have not been properly adjusted], $\sigma_x = 5$cm, $\sigma_y = 4$cm [the horizontal scatter is higher than the vertical one], and $\rho_{xy} = 0$. (Note that a nice graphical representation of this distribution would be created by firing many shots against the target.) We want to compute the probability of hitting the target (by a single shot).

Solution: $\Pr(-10 < X < 10 \cap -10 < Y < 10) = \Pr(-10 < X < 10) \times \Pr(-10 < Y < 10)$ [independence] $= \Pr(\frac{-10-3}{5} < \frac{X-3}{5} < \frac{10-3}{5}) \times \Pr(\frac{-10+2}{4} <$

$\frac{Y+2}{4} < \frac{10+2}{4}) = \Pr(-2.6 < Z_x < 1.4) \times \Pr(-2 < Z_y < 3)$ [Z_x and Z_y are independent standardized Normal RVs] $= (0.4953 + 0.4192) \times (0.4772 + 0.4987) = 89.25\%$.

2. Let us modify the previous example: The target is now a circle whose radius is 12cm, the rifle has been properly adjusted [$\mu_x = 0$ and $\mu_y = 0$], the vertical scatter equals the horizontal [$\sigma_x = \sigma_y = 5$cm], and X and Y remain independent.

   The probability of hitting the target is now: $\Pr(\sqrt{X^2 + Y^2} < 12) = \Pr(\sqrt{(\frac{X}{5})^2 + (\frac{Y}{5})^2} < \frac{12}{5}) = \Pr(\sqrt{Z_1^2 + Z_2^2} < 2.4)$ where again Z_1 and Z_2 are independent standardized Normal. This leads to the following two-dimensional integration: $\frac{1}{2\pi} \iint\limits_{\sqrt{z_1^2+z_2^2}<2.4} \exp\left(-\frac{z_1^2+z_2^2}{2}\right) dz_1\, dz_2$. Using polar coordinates, i.e. $\begin{cases} z_1 = r\cos\theta \\ z_2 = r\sin\theta \end{cases}$ we get, for the corresponding Jacobian: $\begin{vmatrix} \frac{\partial z_1}{\partial r} & \frac{\partial z_1}{\partial \theta} \\ \frac{\partial z_2}{\partial r} & \frac{\partial z_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$. This implies that $dz_1\, dz_2 \to r\, dr\, d\theta$ when rewriting the above integral to get the equivalent $\frac{1}{2\pi} \int\limits_0^{2\pi} \int\limits_0^{2.4} \exp\left(-\frac{r^2}{2}\right) r\, dr\, d\theta$ [note that the new limits are covering the same target region as the original integration]. The last double integral is separable, thus: $\frac{1}{2\pi} \times \int\limits_0^{2\pi} d\theta \times \int\limits_0^{2.4} \exp\left(-\frac{r^2}{2}\right) r\, dr =$

   $\frac{1}{2\pi} \times 2\pi \times \int\limits_0^{\frac{2.4^2}{2}} e^{-u}\, du = 1 - e^{-\frac{2.4^2}{2}} = 94.39\%$.

3. Consider a random independent sample of size 200 from the following distribution:

$X =$ / $Y =$	0	1
0	0.1	0.2
1	0.3	0.4

   . Find: $\Pr(\sum\limits_{i=1}^{200} Y_i \geq 150 \,|\, \sum\limits_{i=1}^{200} X_i = 108)$.

   Solution: An exact solution would require a computer, so we have to resort to the Normal approximation. First we need the five basic parameters of the given distribution, namely: $\mu_x = 0.6$, $\mu_y = 0.7$, $\sigma_x^2 = 0.6 - 0.6^2 = 0.24$, $\sigma_y^2 = 0.7 - 0.7^2 = 0.21$ and $\rho_{xy} = \frac{0.4 - 0.6 \times 0.7}{\sqrt{0.24 \times 0.21}} = -0.089087$. This implies immediately: $\mathbb{E}(\sum\limits_{i=1}^{200} X_i) = 200 \times 0.6 = 120$, $\mathbb{E}(\sum\limits_{i=1}^{200} Y_i) = 200 \times 0.7 = 140$, $Var(\sum\limits_{i=1}^{200} X_i) = 200 \times 0.24 = 48$, and $Var(\sum\limits_{i=1}^{200} Y_i) = 200 \times 0.21 = 42$. The correlation coefficient between $\sum\limits_{i=1}^{200} X_i$ and $\sum\limits_{i=1}^{200} Y_i$ will be the same as ρ_{XY}, namely -0.089087. The conditional distribution of $\sum\limits_{i=1}^{200} Y_i$ will be thus $\mathcal{N}(140 - \sqrt{42} \times 0.089087 \times \frac{108-120}{\sqrt{48}}, \sqrt{42} \times \sqrt{1 - 0.089087^2}) = \mathcal{N}(141.00, 6.455) \Rightarrow$ $\Pr(\sum\limits_{i=1}^{200} Y_i \geq 150 \,|\, \sum\limits_{i=1}^{200} X_i = 108) \cong \Pr(\frac{\sum_{i=1}^{200} Y_i - 141}{6.455} > \frac{149.5-141}{6.455} \,|\, \sum\limits_{i=1}^{200} X_i = 108)$

[note the continuity correction] $= \Pr(Z > 1.3168) = 9.39\%$. The exact answer (quite difficult to evaluate even for a computer) would be 9.14% (the exact conditional mean and variance are 141.00 and 6.423). [Unconditionally, $\Pr(\sum_{i=1}^{200} Y_i \geq 150) \cong \Pr(\frac{\sum_{i=1}^{200} Y_i - 140}{\sqrt{42}} > \frac{149.5-140}{\sqrt{42}}) = \Pr(Z > 1.4659) = 7.13\%$ – the exact answer being 6.95%] ■

Appendix

Let us, very informally, introduce the following definitions: An event is called **rare** (**extremely unlikely**, **practically impossible**) if its probability is less than $\frac{1}{2}\%$ (10^{-6}, 10^{-12}) [if every person on Earth tries the experiment once, thousands of them will still get the 'extremely unlikely' event, none are expected to get the 'practically impossible' one − on the other hand, a single individual should not expect an 'extremely unlikely' event ever happen to him].

EXAMPLES:

- Rolling a die and getting 3 sixes in a row is rare, getting 8 is extremely unlikely (it will never happen to us), getting 15 is practically impossible (no one will ever see that happen). Note that μ and σ in this experiment are 0.20 and 0.49, respectively.

- Seeing a random value from the standardized Normal distribution (Z) exceed, in absolute value, 2.8 is rare, fall beyond ± 4.9 is extremely unlikely, beyond ± 7 practically impossible.

- If we go fishing in a situation where $\beta = 5$ min. (expected time between catches), having to wait more than 26 min. for our first catch is rare, having to wait more than 1 hr. 9 min. is extremely unlikely, any longer than 2 hr. 18 min. is practically impossible. [Both μ and σ equal 5 min. here]. ■

Chapter 8 TRANSFORMING RANDOM VARIABLES

of *continuous* type *only* (the less interesting discrete case was dealt with earlier).

The main issue of this chapter is: Given the distribution of X, find the distribution of $Y \equiv \frac{1}{1+X}$ (an expression involving X). Since only one 'old' RV variable (namely X) appear in the definition of the 'new' RV, we call this a UNIVARIATE transformation. Eventually, we must also deal with the so called BIVARIATE transformations of two 'old' RVs (say X and Y), to find the distribution of a 'new' RV, say $U \equiv \frac{X}{X+Y}$ (or any other expression involving X and Y). Another simple example of this bivariate type is finding the distribution of $V \equiv X + Y$ (i.e. we will finally learn how to *add* two random variables).

Let us first deal with the

Univariate transformation

There are two basic techniques for constructing the new distribution:

►Distribution-Function (F) Technique◄

which works as follows:

When the new random variable Y is defined as $g(X)$, we find its distribution function $F_Y(y)$ by computing $\Pr(Y < y) = \Pr[g(X) < y]$. This amounts to *solving* the $g(X) < y$ *inequality* for X [usually resulting in an interval of values], and then integrating $f(x)$ over this interval [or, equivalently, substituting into $F(x)$].

EXAMPLES:

1. Consider $X \in \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$ [this corresponds to a spinning wheel with a two-directional 'pointer', say a laser beam, where X is the pointer's angle from a fixed direction when the wheel stops spinning]. We want to know the distribution of $Y = b\tan(X) + a$ [this represents the location of a dot our laser beam would leave on a screen placed b units from the wheel's center, with a scale whose origin is a units off the center]. Note that Y can have any real value.

   Solution: We start by writing down $F_X(x) =$ [in our case] $\frac{x+\frac{\pi}{2}}{\pi} \equiv \frac{x}{\pi} + \frac{1}{2}$ when $-\frac{\pi}{2} < x < \frac{\pi}{2}$. To get $F_Y(y)$ we need: $\Pr[b\tan(X) + a < y] = \Pr[X < \arctan(\frac{y-a}{b})] = F_X[\arctan(\frac{y-a}{b})] = \frac{1}{\pi}\arctan(\frac{y-a}{b}) + \frac{1}{2}$ where $-\infty < y < \infty$. Usually, we can relate better to the corresponding $f_Y(y)$ [which tells us what is likely and what is not] $= \frac{1}{\pi b} \cdot \frac{1}{1+(\frac{y-a}{b})^2} =$

$$\frac{b}{\pi} \cdot \frac{1}{b^2 + (y-a)^2} \qquad (f)$$

   [any real y]. Graphically, this function looks very similar to the Normal pdf (also a 'bell-shaped' curve), but in terms of its properties, the new distribution turns out to be totally different from Normal, [as we will see later].

The name of this new distribution is **Cauchy** [notation: $\mathcal{C}(a,b)$]. Since the $\int_{-\infty}^{\infty} y \cdot f_Y(y)\,dy$ integral leads to $\infty - \infty$, the Cauchy distribution does *not* have a mean (consequently, its variance is infinite). Yet it possesses a clear *center* (at $y = a$) and *width* ($\pm b$). These are now identified with the *median* $\tilde{\mu}_Y = a$ [verify by solving $F_Y(\tilde{\mu}) = \frac{1}{2}$] and the so called *semi-inter-quartile range* (QUARTILE DEVIATION, for short) $\frac{Q_U - Q_L}{2}$ where Q_U and Q_L are the UPPER and LOWER QUARTILES [defined by $F(Q_U) = \frac{3}{4}$ and $F(Q_L) = \frac{1}{4}$]. One can easily verify that, in this case, $Q_L = a - b$ and $Q_U = a + b$ [note that the semi-inter-quartile range contains exactly 50% of all probability], thus the quartile deviation equals to b. The most *typical* ('standardized') *case* of the Cauchy distribution is $\mathcal{C}(0,1)$, whose pdf equals

$$f(y) = \frac{1}{\pi} \cdot \frac{1}{1+y^2}$$

Its 'rare' values start at ± 70, we need to go beyond ± 3000 to reach 'extremely unlikely', and only ∓ 300 billion become 'practically impossible'. Since the mean does not exist, the central limit theorem breaks down [it is no longer true that $\bar{Y} \to \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$, there is no μ and σ is infinite]. Yet, $\bar{Y}$ must have some well defined distribution. We will discover what that distribution is in the next section.

2. Let X have its pdf defined by $f(x) = 6x(1-x)$ for $0 < x < 1$.Find the pdf of $Y = X^3$.

   Solution: First we realize that $0 < Y < 1$. Secondly, we find $F_X(x) = 6\int_0^x (x - x^2)\,dx = 6(\frac{x^2}{2} - \frac{x^3}{3}) = 3x^2 - 2x^3$. And finally: $F_Y(y) \equiv \Pr(Y < y) = \Pr(X^3 < y) = \Pr(X < y^{\frac{1}{3}}) = F_X(y^{\frac{1}{3}}) = 3y^{\frac{2}{3}} - 2y$.This easily converts to $f_Y(y) = 2y^{-\frac{1}{3}} - 2$ where $0 < y < 1$ [zero otherwise]. (Note that when $y \to 0$ this pdf becomes infinite, which is OK).

3. Let $X \in \mathcal{U}(0,1)$. Find and identify the distribution of $Y = -\ln X$ (its range is obviously $0 < y < \infty$).

   Solution: First we need $F_X(x) = x$ when $0 < x < 1$. Then: $F_Y(y) = \Pr(-\ln X < y) = \Pr(X > e^{-y})$ [note the sign reversal] $= 1 - F_X(e^{-y}) = 1 - e^{-y}$ where $y > 0$ ($\Rightarrow f_Y(y) = e^{-y}$). This can be easily identified as the exponential distribution with the mean of 1 [note that $Y = -\beta \cdot \ln X$ would result in the exponential distribution with the mean equal to β].

4. If $Z \in \mathcal{N}(0,1)$, what is the distribution of $Y = Z^2$.

   Solution: $F_Y(y) = \Pr(Z^2 < y) = \Pr(-\sqrt{y} < Z < \sqrt{y})$ [right?] $= F_Z(\sqrt{y}) - F_Z(\sqrt{y})$. Since we don't have an explicit expression for $F_Z(z)$ it would appear that we are stuck at this point, but we can get the corresponding $f_Y(y)$ by a simple differentiation: $\frac{dF_Z(\sqrt{y})}{dy} - \frac{dF_Z(-\sqrt{y})}{dy} = \frac{1}{2}y^{-\frac{1}{2}} f_Z(\sqrt{y}) + \frac{1}{2}y^{-\frac{1}{2}} f_Z(-\sqrt{y}) = \frac{y^{-\frac{1}{2}} e^{-\frac{y}{2}}}{\sqrt{2\pi}}$ where $y > 0$. This can be identified as the *gamma* distribution with

$\alpha = \frac{1}{2}$ and $\beta = 2$ [the normalizing constant is equal to $\Gamma(\frac{1}{2}) \cdot 2^{\frac{1}{2}} = \sqrt{2\pi}$, check].

Due to its importance, this distribution has yet another name, it is called the **chi-square distribution** with *one degree of freedom,* or χ^2_1 for short. It has the expected value of $(\alpha \cdot \beta =)$ 1, its variance equals $(\alpha \cdot \beta^2 =)$ 2, and the MGF is $M(t) = \frac{1}{\sqrt{1-2t}}$. To answer a probability question concerning Y, we have to convert it back to Z (and use the corresponding tables). For example: $\Pr(1 < Y < 2) = \Pr(-\sqrt{2} < Z < \sqrt{2}) - \Pr(-1 < Z < 1) = 2 \times 0.42135 - 2 \times 0.34135 = 16.00\%$. ■

General Chi-square **distribution:**

(This is an extension of the previous example). We want to investigate the RV defined by $U = Z_1^2 + Z_2^2 + Z_3^2 + ... + Z_n^2$, where Z_1, Z_2, Z_3, ...Z_n are *independent* RVs from the $\mathcal{N}(0,1)$ distribution. Its **MGF** must obviously equal to $M(t) = \frac{1}{(1-2t)^{\frac{n}{2}}}$; we can thus identify its distribution as *gamma*, with $\alpha = \frac{n}{2}$ and $\beta = 2$ ($\Rightarrow$ mean $= n$, variance $= 2n$). Due to its importance, it is also called the chi-square distribution with n (integer) degrees of freedom (χ^2_n for short).

When n is *even* we have no difficulty calculating the related **probabilities**, using $F(x)$ of the corresponding gamma distribution. Thus, for example, if $U \in \chi^2_{10}$, $\Pr(U < 18.307) = 1 - e^{-\frac{18.307}{2}} \left[1 + 9.1535 + \frac{9.1535^2}{2} + \frac{9.1535^3}{6} + \frac{9.1535^4}{24}\right] = 95.00\%$.

When n is *odd* (equal to $2k+1$), we must switch to using the following new *formula*:

$$\Pr(U < u) = \Pr(Z^2 < u) - e^{-\frac{u}{2}} \left[\frac{(\frac{u}{2})^{\frac{1}{2}}}{\Gamma(\frac{3}{2})} + \frac{(\frac{u}{2})^{\frac{3}{2}}}{\Gamma(\frac{5}{2})} + ... + \frac{(\frac{u}{2})^{k-\frac{1}{2}}}{\Gamma(k+\frac{1}{2})}\right]$$

Proof: [U_n represents a random variable having the χ^2_n-distribution] $\Pr(U_{2k+1} < u) = \frac{1}{\Gamma(k+\frac{1}{2})2^{k+\frac{1}{2}}} \int_0^u x^{k+\frac{1}{2}-1} e^{-\frac{x}{2}}\, dx = \frac{1}{\Gamma(k+\frac{1}{2})2^{k+\frac{1}{2}}} \int_0^u x^{k+\frac{1}{2}-1}(-2e^{-\frac{x}{2}})'\, dx =$

$-\frac{u^{k-\frac{1}{2}}e^{-\frac{u}{2}}}{\Gamma(k+\frac{1}{2})2^{k-\frac{1}{2}}} + \frac{1}{\Gamma(k-\frac{1}{2})2^{k-\frac{1}{2}}} \int_0^u x^{k-\frac{1}{2}-1} e^{-\frac{x}{2}}\, dx = -\frac{u^{k-\frac{1}{2}}e^{-\frac{u}{2}}}{\Gamma(k+\frac{1}{2})2^{k-\frac{1}{2}}} + \Pr(U_{2k-1} <$

$u)$. Used repeatedly, this yields: $\Pr(U_{2k+1} < u) = -\frac{u^{k-\frac{1}{2}}e^{-\frac{u}{2}}}{\Gamma(k+\frac{1}{2})2^{k-\frac{1}{2}}} - \frac{u^{k-\frac{3}{2}}e^{-\frac{u}{2}}}{\Gamma(k-\frac{1}{2})2^{k-\frac{3}{2}}} -$

$..... - \frac{u^{\frac{1}{2}}e^{-\frac{u}{2}}}{\Gamma(\frac{3}{2})2^{\frac{1}{2}}} + \Pr(U_1 < u)$. And we already know that $U_1 \equiv Z^2$. □

EXAMPLE:

$\Pr(U_9 < 3.325) = \Pr(Z^2 < 3.325) - e^{-1.6625}\left[\frac{1.6625^{\frac{1}{2}}}{\frac{1}{2}\sqrt{\pi}} + \frac{1.6625^{\frac{3}{2}}}{\frac{3}{2}\cdot\frac{1}{2}\sqrt{\pi}} + \frac{1.6625^{\frac{5}{2}}}{\frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\sqrt{\pi}} + \frac{1.6625^{\frac{7}{2}}}{\frac{7}{2}\cdot\frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\sqrt{\pi}}\right] =$
$\Pr(-1.8235 < Z < 1.8234) - 0.8818 = 5.00\%$. ■

If, in this context, we need a value of $\Pr(0 < Z < z)$ which is **outside** the Z **tables** (i.e. $z > 3$), the following approximation may come handy: $\Pr(Z > z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{u^2}{2}}\, du = \frac{1}{\sqrt{2\pi}} \int_z^\infty (-\frac{1}{u}) \cdot (e^{-\frac{u^2}{2}})'\, du =$ [by parts] $\frac{1}{\sqrt{2\pi}} \cdot \frac{e^{-\frac{z^2}{2}}}{z} - \frac{1}{\sqrt{2\pi}} \int_z^\infty \frac{e^{-\frac{u^2}{2}}}{u^2}\, du =$

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{e^{-\frac{z^2}{2}}}{z} \left(1 - \frac{1}{z^2}\right) + \frac{1}{\sqrt{2\pi}} \int_z^\infty \frac{e^{-\frac{u^2}{2}}}{u^3}\, du = \ .... =$$

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{e^{-\frac{z^2}{2}}}{z} \left(1 - \frac{1}{z^2} + \frac{3}{z^4} - \frac{5 \times 3}{z^6} + \frac{7 \times 5 \times 3}{z^8} - ...\right) \qquad (1 - F)$$

This is a so called ASYMPTOTIC EXPANSION which *diverges* for every value of z, but provides a good approximation to the above probability when truncated to the first two terms, e.g. $\Pr(Z > 3.27) \simeq \frac{e^{-\frac{3.27^2}{2}}}{\sqrt{2\pi} \times 3.27} \times (1 - \frac{1}{3.27^2}) = 0.00053$ [the exact answer being 0.00054]. The expansion becomes meaningless when z is small [use with $z > 3$ only].

►Probability-Density-Function (f) Technique◄

is a bit faster and usually somehow easier (technically) to carry out, but it works for *one-to-one* transformations *only* (e.g. it would not work in our last $Y = Z^2$ example). The procedure consists of three simple steps:

(i) Express X (the 'old' variable) in terms of y the 'new' variable [getting an expression which involves only Y].

(ii) Substitute the result [we will call it $x(y)$, switching to small letters] for the argument of $f_X(x)$, getting $f_X[x(y)]$ − a function of y!

(iii) Multiply this by $\left|\frac{dx(y)}{dy}\right|$. The result is the pdf of Y. ■

In **summary**

$$f_Y(y) = f_X[x(y)] \cdot \left|\frac{dx(y)}{dy}\right|$$

EXAMPLES (we will redo the first three examples of the previous section):

1. $X \in \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$ and $Y = b \tan(X) + a$.
   Solution: (i) $x = \arctan(\frac{y-a}{b})$, (ii) $\frac{1}{\pi}$, (iii) $\frac{1}{\pi} \cdot \frac{1}{b} \cdot \frac{1}{1+(\frac{y-a}{b})^2} = \frac{b}{\pi} \cdot \frac{1}{b^2+(y-a)^2}$ where $-\infty < y < \infty$ [check].

2. $f(x) = 6x(1-x)$ for $0 < x < 1$ and $Y = X^3$.
   Solution: (i) $x = y^{1/3}$, (ii) $6y^{1/3}(1 - y^{1/3})$, (iii) $6y^{1/3}(1 - y^{1/3}) \cdot \frac{1}{3} y^{-2/3} = 2(y^{-1/3} - 1)$ when $0 < y < 1$ [check].

3. $X \in \mathcal{U}(0, 1)$ and $Y = -\ln X$.
   Solution: (i) $x = e^{-y}$, (ii) 1, (iii) $1 \cdot e^{-y} = e^{-y}$ for $y > 0$ [check].

This does appear to be a fairly fast way of obtaining $f_Y(y)$. ■

And now we extend all this to the

Bivariate transformation

►Distribution-Function Technique◄

follows essentially the same pattern as the univariate case:

The new random variable Y is now defined in terms of two 'old' RVs, say X_1 and X_2, by $y \equiv g(X_1, X_2)$. We find $F_Y(y) = \Pr(Y < y) = \Pr[g(X_1, X_2) < y]$ by realizing that the $g(X_1, X_2) < y$ inequality (for X_1 and X_2, y is considered fixed) will now result in some *2-D region*, and then integrating $f(x_1, x_2)$ over this region.

Thus, the technique is simple in principle, but often quite involved in terms of technical details.

EXAMPLES:

1. Suppose that X_1 and X_2 are independent RVs, both from $\mathcal{E}(1)$, and $Y = \dfrac{X_2}{X_1}$.

   Solution: $F_Y(y) = \Pr\left(\dfrac{X_2}{X_1} < y\right) = \Pr(X_2 < yX_1) = \iint\limits_{0<x_2<yx_1} e^{-x_1-x_2}\,dx_1\,dx_2 = \int\limits_0^\infty e^{-x_1} \int\limits_0^{yx_1} e^{-x_2}\,dx_2\,dx_1 = \int\limits_0^\infty e^{-x_1}(1-e^{-yx_1})\,dx_1 = \int\limits_0^\infty (e^{-x_1} - e^{-x_1(1+y)})\,dx_1 = 1 - \dfrac{1}{1+y}$, where $y > 0$. This implies that $f_Y(y) = \dfrac{1}{(1+y)^2}$ when $y > 0$. (The median $\tilde{\mu}$ of this distribution equals to 1, the lower and upper quartiles are $Q_L = \dfrac{1}{3}$ and $Q_U = 3$).

2. This time Z_1 and Z_2 are independent RVs from $\mathcal{N}(0,1)$ and $Y = Z_1^2 + Z_2^2$ [here, we know the answer: χ^2_2, let us proceed anyhow].

   Solution: $F_Y(y) = \Pr(Z_1^2+Z_2^2 < y) = \frac{1}{2\pi} \iint\limits_{z_1^2+z_2^2<y} e^{-\frac{z_1^2+z_2^2}{2}}\,dz_1\,dz_2 = \frac{1}{2\pi} \int\limits_0^{2\pi} \int\limits_0^{\sqrt{y}} e^{-\frac{r^2}{2}} \cdot r\,dr\,d\theta =$[substitution: $w = \frac{r^2}{2}$] $\int\limits_0^{\frac{y}{2}} e^{-w}\,dw = 1 - e^{-\frac{y}{2}}$ where (obviously) $y > 0$.

   This is the *exponential* distribution with $\beta = 2$ [not χ^2_2 as expected, how come?]. It does not take long to realize that the two distributions are identical.

3. (**Sum of two independent RVs**): Assume that X_1 and X_2 are independent RVs from a distribution having L and H as its lowest and highest possible value, respectively. Find the distribution of $X_1 + X_2$ [finally learning how to *add* two RVs!].

   Solution: $F_Y(y) = \Pr(X_1 + X_2 < y) = \iint\limits_{\substack{x_1+x_2<y \\ L<x_1,x_2<H}} f(x_1) \cdot f(x_2)\,dx_1\,dx_2 =$

   $$\begin{cases} \int\limits_L^{y-L} \int\limits_L^{y-x_1} f(x_1) \cdot f(x_2)\,dx_2 dx_1 & \text{when } y < L+H \\ 1 - \int\limits_{y-H}^H \int\limits_{y-x_1}^H f(x_1) \cdot f(x_2)\,dx_2 dx_1 & \text{when } y > L+H \end{cases}$$

   . Differentiating this with

respect to y (for the first line, this amounts to: substituting $y - L$ for x_1 and dropping the dx_1 integration – contributing zero in this case – plus: substituting $y - x_1$ for x_2 and dropping dx_2; same for the first line, except that we have to *subtract* the second contribution) results in $f_Y(y) = \begin{cases} \int_L^{y-L} f(x_1) \cdot f(y - x_1)\, dx_1 \text{ when } y < L + H \\ \int_{y-H}^{H} f(x_1) \cdot f(y - x_1)\, dx_1 \text{ when } y > L + H \end{cases}$ or, equivalently,

$$f_Y(y) = \int_{\max(L,y-H)}^{\min(H,y-L)} f(x) \cdot f(y - x)\, dx$$

where the y-range is obviously $2L < y < 2H$. The right hand side of the last formula is sometimes referred to as the CONVOLUTION of two pdfs (in general, the two fs may be distinct).

Examples:

- In the specific case of the **uniform** $\mathcal{U}(0, 1)$ distribution, the last formula yields, for the pdf of $Y \equiv X_1 + X_2$:

$$f_Y(y) = \int_{\max(0,y-1)}^{\min(1,y)} dx = \begin{cases} \int_0^y dx = y & \text{when} \quad 0 < y < 1 \\ \int_{y-1}^{1} dx = 2 - y & \text{when} \quad 1 < y < 2 \end{cases}$$ ['triangular' distribution].

- Similarly, for the 'standardized' **Cauchy** distribution $\left[f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}\right]$, we get: $f_{X_1+X_2}(y) = \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{1}{1+x^2} \cdot \frac{1}{1+(y-x)^2}\, dx =$[partial fractions] $\frac{1}{\pi^2} \int_{-\infty}^{\infty} \left[\frac{A+Bx}{1+x^2} + \frac{C+D(x-y)}{1+(y-x)^2}\right] dx$. We find A, B, C and D from $(A + Bx)\left[1 + (x - y)^2\right] + [C + D(x - y)]\,(1 + x^2) \equiv 1$ by first substituting $x = i$ and getting $Ay^2 + 2By = 1$ (real part) and $-2Ay + By^2 = 0$ (purely imaginary) $\Rightarrow A = \frac{1}{4+y^2}$ and $B = \frac{2}{y(4+y^2)}$, then substituting $x = y + i$ and similarly getting $C = \frac{1}{4+y^2}$ and $D = \frac{-2}{y(4+y^2)}$. The B and D parts of the dx integration cancel out, the A and C contributions add up to $\frac{2}{\pi} \cdot \frac{1}{4+y^2}$ [where $-\infty < y < \infty$].

  The last result can be easily converted to the pdf of $\bar{X} = \frac{X_1+X_2}{2}$ [the *sample mean* of the two random values], yielding $f_{\bar{X}}(\bar{x}) = \frac{2}{\pi} \cdot \frac{1}{4+(2\bar{x})^2} \cdot 2 = \frac{1}{\pi} \cdot \frac{1}{1+\bar{x}^2}$. Thus, the sample mean $\bar{X}$ has the *same* Cauchy distribution as do the two individual observations (the result can be extended to *any number* of observations). We knew that the Central Limit Theorem [$\bar{X} \,\widetilde{\in}\, \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$] would not apply to this case, but the actual distribution of $\bar{X}$ still comes as a big surprise. This implies that the sample mean of even millions of values (from a Cauchy distribution) cannot estimate the center of the distribution any better than a *single* observation [one can verify this by actual simulation]. Yet, one feels that there must be a way of substantially improving the estimate (of

the location of a laser gun hidden behind a screen) when going from a single observation to a large sample. Yes, there is, if one does not use the sample *mean* but something else; later on we discover that the sample *median* will do just fine. ■

▶Pdf (Shortcut) Technique◀

works a bit faster, even though it may *appear* more complicated as it requires the following (several) steps:

1. The procedure can work only for **one-to-one** ('invertible') transformations. This implies that the new RV $Y \equiv g(X_1, X_2)$ must be accompanied by yet another *arbitrarily* chosen function of X_1 and/or X_2 [the original Y will be called Y_1, and the auxiliary one Y_2, or vice versa]. We usually choose this second (auxiliary) function in the simplest possible manner, i.e. we make it equal to X_2 (or X_1):

2. **Invert** the transformation, i.e. solve the two equations $y_1 = g(x_1, x_2)$ and $y_2 = x_2$ for x_1 and x_2 (in terms of y_1 and y_2). Getting a unique solution guarantees that the transformation is one-to-one.

3. **Substitute** this solution $x_1(y_1, y_2)$ and $x_2(y_2)$ into the joint pdf of the 'old' X_1, X_2 pair (yielding a function of y_1 and y_2).

4. Multiply this function by the transformation's **Jacobian** $\left| \begin{array}{cc} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{array} \right|$. The result is the joint pdf of Y_1 and Y_2. At the same time, establish the region of possible (Y_1, Y_2) values in the (y_1, y_2)-plane [this is often the most difficult part of the procedure].

5. Eliminate Y_2 [the 'phoney', auxiliary RV introduced to help us with the inverse] by integrating it out (finding the Y_1 **marginal**). Don't forget that you must integrate over the *conditional* range of y_2 *given* y_1.

EXAMPLES:

1. X_1, $X_2 \in \mathcal{E}(1)$, independent; $Y = \frac{X_1}{X_1+X_2}$ [the time of the first 'catch' relative to the time needed to catch two fishes].

   Solution: $Y_2 = X_2 \Rightarrow x_2 = y_2$ and $x_1 y_1 + x_2 y_1 = x_1 \Rightarrow x_1 = \frac{y_1 \cdot y_2}{1-y_1}$. Substitute into $e^{-x_1-x_2}$ getting $e^{-y_2\left(1+\frac{y_1}{1-y_1}\right)} = e^{-\frac{y_2}{1-y_1}}$, multiply by $\left| \begin{array}{cc} y_2 \frac{1-y_1+y_1}{(1-y_1)^2} & \frac{y_1}{1-y_1} \\ 0 & 1 \end{array} \right| = \frac{y_2}{(1-y_1)^2}$ getting $f(y_1, y_2) = \frac{y_2}{(1-y_1)^2} e^{-\frac{y_2}{1-y_1}}$ with $0 < y_1 < 1$ and $y_2 > 0$. Eliminate Y_2 by $\int_0^\infty \frac{y_2}{(1-y_1)^2} e^{-\frac{y_2}{1-y_1}} dy_2 = \frac{1}{(1-y_1)^2} \cdot (1-y_1)^2 \equiv 1$ when $0 < y_1 < 1$ [recall the $\int_0^\infty x^k e^{-\frac{x}{a}} dx = k! \cdot a^{k+1}$ formula]. The distribution of Y is thus $\mathcal{U}(0, 1)$. Note that if we started with X_1, $X_2 \in \mathcal{E}(\beta)$ instead of $\mathcal{E}(1)$, the result would have been the same since this new $Y = \frac{X_1}{X_1+X_2} \equiv \frac{\frac{X_1}{\beta}}{\frac{X_1}{\beta}+\frac{X_2}{\beta}}$ where $\frac{X_1}{\beta}$ and $\frac{X_2}{\beta} \in \mathcal{E}(1)$ [this can be verified by a simple MGF argument].

2. Same X_1 and X_2 as before, $Y = \frac{X_2}{X_1}$.

   Solution: This time we reverse the labels: $Y_1 \equiv X_1$ and $Y_2 = \frac{X_2}{X_1} \Rightarrow x_1 = y_1$ and $x_2 = y_1 \cdot y_2$. Substitute into $e^{-x_1-x_2}$ to get $e^{-y_1(1+y_2)}$, times $\begin{vmatrix} 1 & 0 \\ y_2 & y_1 \end{vmatrix} = y_1$ gives the joint pdf for $y_1 > 0$ and $y_2 > 0$. Eliminate y_1 by $\int_0^\infty y_1 e^{-y_1(1+y_2)}\, dy_1 = \frac{1}{(1+y_2)^2}$, where $y_2 > 0$. Thus, $f_Y(y) = \frac{1}{(1+y)^2}$ with $y > 0$ [check, we have solved this problem before].

3. In this example we introduce the so called ►Beta **distribution**◄

   Let X_1 and X_2 be independent RVs from the gamma distribution with parameters (k, β) and (m, β) respectively, and let $Y_1 = \frac{X_1}{X_1+X_2}$.

   Solution: Using the argument of Example 1 one can show that β 'cancels out', and we can assume that $\beta = 1$ without affecting the answer. The definition of Y_1 is also the same as in Example 1 $\Rightarrow x_1 = \frac{y_1 y_2}{1-y_1}$, $x_2 = y_2$, and the Jacobian $= \frac{y_2}{(1-y_1)^2}$. Substituting into $f(x_1, x_2) = \frac{x_1^{k-1} x_2^{m-1} e^{-x_1-x_2}}{\Gamma(k)\cdot\Gamma(m)}$ and multiplying by the Jacobian yields $f(y_1, y_2) = \dfrac{y_1^{k-1} y_2^{k-1} y_2^{m-1} e^{-\frac{y_2}{1-y_1}}}{\Gamma(k)\Gamma(m)(1-y_1)^{k-1}} \cdot \dfrac{y_2}{(1-y_1)^2}$ for $0 < y_1 < 1$ and $y_2 > 0$. Integrating over y_2 results in: $\dfrac{y_1^{k-1}}{\Gamma(k)\Gamma(m)(1-y_1)^{k+1}} \int_0^\infty y_2^{k+m-1} e^{-\frac{y_2}{1-y_1}}\, dy_2 =$

$$\frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \cdot y_1^{k-1}(1-y_1)^{m-1} \qquad \text{(f)}$$

   where $0 < y_1 < 1$.

   This is the **pdf** of a new two-parameters (k and m) distribution which is called beta. Note that, as a by-product, we have effectively proved the following formula: $\int_0^1 y^{k-1}(1-y)^{m-1} dy = \frac{\Gamma(k)\cdot\Gamma(m)}{\Gamma(k+m)}$ for any $k, m > 0$. This enables us to find the distribution's **mean**: $\mathbb{E}(Y) = \frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \int_0^1 y^k (1-y)^{m-1}\, dy = \frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \cdot \frac{\Gamma(k+1)\cdot\Gamma(m)}{\Gamma(k+m+1)} =$

$$\frac{k}{k+m} \qquad \text{(mean)}$$

   and similarly $\mathbb{E}(Y^2) = \frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \int_0^1 y^{k+1}(1-y)^{m-1}\, dy = \frac{\Gamma(k+m)}{\Gamma(k)\cdot\Gamma(m)} \cdot \frac{\Gamma(k+2)\cdot\Gamma(m)}{\Gamma(k+m+2)} = \frac{(k+1)\,k}{(k+m+1)\,(k+m)} \Rightarrow Var(Y) = \frac{(k+1)\,k}{(k+m+1)\,(k+m)} - (\frac{k}{k+m})^2 =$

$$\frac{k\,m}{(k+m+1)\,(k+m)^2} \qquad \text{(variance)}$$

   Note that the distribution of $1 - Y \equiv \frac{X_2}{X_1+X_2}$ is also beta (why?) with parameters m and k [reversed].

   We learn how to compute related **probabilities** in the following set of Examples:

(a) $\Pr(X_1 < \frac{X_2}{2})$ where X_1 and X_2 have the gamma distribution with parameters $(4, \beta)$ and $(3, \beta)$ respectively [this corresponds to the probability that Mr.A catches 4 fishes in less than half the time Mr.B takes to catch 3].

Solution: $\Pr(2X_1 < X_2) = \Pr(3X_1 < X_1 + X_2) = \Pr(\frac{X_1}{X_1+X_2} < \frac{1}{3}) = \frac{\Gamma(4+3)}{\Gamma(4)\cdot\Gamma(3)} \int_0^{\frac{1}{3}} y^3(1-y)^2 dy = 60 \times \left[\frac{y^4}{4} - 2\frac{y^5}{5} + \frac{y^6}{6}\right]_{y=0}^{\frac{1}{3}} = 10.01\%$.

(b) Evaluate $\Pr(Y < 0.4)$ where Y has the beta distribution with parameters $(\frac{3}{2}, 2)$ [half-integer values are not unusual, as we learn shortly].

Solution: $\frac{\Gamma(\frac{7}{2})}{\Gamma(\frac{3}{2})\cdot\Gamma(2)} \int_0^{0.4} y^{\frac{1}{2}}(1-y)\, dy = \frac{5}{2} \cdot \frac{3}{2} \cdot \left[\frac{y^{\frac{3}{2}}}{\frac{3}{2}} - \frac{y^{\frac{5}{2}}}{\frac{5}{2}}\right]_{y=0}^{0.4} = 48.07\%$.

(c) Evaluate $\Pr(Y < 0.7)$ where $Y \in \text{beta}(4, \frac{5}{2})$.

Solution: This equals [it is more convenient to have the half-integer first] $\Pr(1-Y > 0.3) = \frac{\Gamma(\frac{13}{2})}{\Gamma(\frac{5}{2})\cdot\Gamma(4)} \int_{0.3}^{1} u^{\frac{3}{2}}(1-u)^3\, du = \frac{\frac{11}{2}\cdot\frac{9}{2}\cdot\frac{7}{2}\cdot\frac{5}{2}}{3!} \left[\frac{y^{\frac{5}{2}}}{\frac{5}{2}} - 3\frac{y^{\frac{7}{2}}}{\frac{7}{2}} + 3\frac{y^{\frac{9}{2}}}{\frac{9}{2}} - \frac{y^{\frac{11}{2}}}{\frac{11}{2}}\right]_{y=0.3}^{1} = 1 - 0.3522 = 64.78\%$.

The main challenge arises when both parameters are **half-integers**. Suggested substitution is $y = \sin^2 u$, $1 - y = \cos^2 u$ and $dy = 2\sin u \cos u\, du$. Let us try it:

(d) $\Pr(Y < 0.5)$ when $Y \in \text{beta}(\frac{3}{2}, \frac{1}{2})$.

Solution: $\frac{\Gamma(2)}{\Gamma(\frac{3}{2})\cdot\Gamma(\frac{1}{2})} \int_0^{0.5} y^{\frac{1}{2}}(1-y)^{-\frac{1}{2}} dy = \frac{1}{\frac{1}{2}\cdot\sqrt{\pi}\cdot\sqrt{\pi}} \int_0^{\arcsin\sqrt{0.5}} \sin u \cdot \frac{1}{\cos u} \cdot 2\sin u \cos u\, du = \frac{4}{\pi} \int_0^{\arcsin\sqrt{0.5}} \sin^2 u\, du = \frac{2}{\pi} \int_0^{\arcsin\sqrt{0.5}} (1 - \cos 2u)\, du = \frac{2}{\pi} \left[u - \frac{\sin 2u}{2}\right]_0^{0.785398} = 18.17\%$.

The beta distribution turns out to be a convenient means of evaluating probabilities of two other important distributions (yet to be discussed).

4. In this example we introduce the so called '**Student**' or ▶**t-distribution**◀

[notation: t_n, where n is called 'degrees of freedom' − the only parameter]. We start with two independent RVs $X_1 \in \mathcal{N}(0,1)$ and $X_2 \in \chi^2_n$, and introduce a new RV by $Y_1 = \dfrac{X_1}{\sqrt{\frac{X_2}{n}}}$.

To get its **pdf** we take $Y_2 \equiv X_2$, solve for $x_2 = y_2$ and $x_1 = y_1 \cdot \sqrt{\frac{y_2}{n}}$, substitute into $f(x_1, x_2) = \dfrac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} \cdot \dfrac{x_2^{\frac{n}{2}-1} e^{-\frac{x_2}{2}}}{\Gamma(\frac{n}{2}) \cdot 2^{\frac{n}{2}}}$ and multiply by $\begin{vmatrix} \sqrt{\frac{y_2}{n}} & \frac{1}{2} \cdot \frac{y_1}{\sqrt{n y_2}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{y_2}{n}}$ to get $f(y_1, y_2) = \dfrac{e^{-\frac{y_1^2 y_2}{2n}}}{\sqrt{2\pi}} \cdot \dfrac{y_2^{\frac{n}{2}-1} e^{-\frac{y_2}{2}}}{\Gamma(\frac{n}{2}) \cdot 2^{\frac{n}{2}}} \cdot \sqrt{\frac{y_2}{n}}$ where $-\infty < y_1 < \infty$ and

$y_2 > 0$. To eliminate y_2 we integrate: $\frac{1}{\sqrt{2\pi}\Gamma(\frac{n}{2})\,2^{\frac{n}{2}}\sqrt{n}}\int_0^\infty y_2^{\frac{n-1}{2}} e^{-\frac{y_2}{2}(1+\frac{y_1^2}{n})} dy_2 =$

$$\frac{\Gamma(\frac{n+1}{2})\,2^{\frac{n+1}{2}}}{\sqrt{2\pi}\Gamma(\frac{n}{2})\,2^{\frac{n}{2}}\sqrt{n}\left(1+\frac{y_1^2}{n}\right)^{\frac{n+1}{2}}} =$$

$$\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \cdot \frac{1}{\left(1+\frac{y_1^2}{n}\right)^{\frac{n+1}{2}}} \qquad (f)$$

with $-\infty < y_1 < \infty$. Note that when $n = 1$ this gives $\frac{1}{\pi} \cdot \frac{1}{1+y_1^2}$ (Cauchy), when $n \to \infty$ the second part of the formula tends to $e^{-\frac{y_1^2}{2}}$ which is, up to the normalizing constant, the pdf of $\mathcal{N}(0,1)$ [implying that $\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \underset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}}$, why?].

Due to the symmetry of the distribution $[f(y) = f(-y)]$ its **mean** is zero (when is exists, i.e. when $n \geq 2$).

To compute its **variance** we first realize that $\int_{-\infty}^{\infty} \frac{dy}{\left(1+\frac{y^2}{n}\right)^{\frac{n+1}{2}}} = \frac{\Gamma(\frac{n}{2})\sqrt{n\pi}}{\Gamma(\frac{n+1}{2})}$ which implies (after the $\frac{y^2}{n} = \frac{x^2}{a}$ substitution) that $\int_{-\infty}^{\infty} \frac{dx}{\left(1+\frac{x^2}{a}\right)^{\frac{n+1}{2}}} = \frac{\Gamma(\frac{n}{2})\sqrt{a\pi}}{\Gamma(\frac{n+1}{2})}$ for any $a > 0$ and $n > 0$. With the help of this formula we get $Var(Y) = \mathbb{E}(Y^2) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \int_{-\infty}^{\infty} \frac{(y^2+n-n)\,dy}{\left(1+\frac{y^2}{n}\right)^{\frac{n+1}{2}}} = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left[n \cdot \frac{\Gamma(\frac{n-2}{2})\sqrt{n\pi}}{\Gamma(\frac{n-1}{2})} - n \cdot \frac{\Gamma(\frac{n}{2})\sqrt{n\pi}}{\Gamma(\frac{n+1}{2})}\right] =$

$n \cdot \frac{\frac{n-1}{2}}{\frac{n-2}{2}} - n =$

$$\frac{n}{n-2} \qquad \text{(variance)}$$

for $n \geq 3$ (for $n = 1$ and 2 the variance is infinite).

The main task again is to learn how to compute the corresponding **probabilities**. We can see immediately that $Y \in \mathsf{t}_n$ implies that

$$\frac{Y^2}{Y^2+n} \qquad (\mathsf{beta})$$

$= \frac{\frac{nZ^2}{\chi_n^2}}{\frac{nZ^2}{\chi_n^2}+n} \equiv \frac{Z^2}{Z^2+\chi_n^2} \equiv \frac{\chi_1^2}{\chi_1^2+\chi_n^2} \equiv \frac{\mathsf{gamma}(\frac{1}{2},2)}{\mathsf{gamma}(\frac{1}{2},2)+\mathsf{gamma}(\frac{n}{2},2)}$ has the beta distribution with parameters $\frac{1}{2}$ and $\frac{n}{2}$. And that distribution we know how to deal with.

Examples:

(a) Compute $\Pr(Y < 1.4)$ where $Y \in \mathsf{t}_6$.

Solution: $= \frac{1}{2} + \Pr(0 < Y < 1.4) = \frac{1}{2} + \frac{1}{2}\Pr(-1.4 < Y < 1.4) = \frac{1}{2} + \frac{1}{2}\Pr(Y^2 < 1.4^2) = \frac{1}{2} + \frac{1}{2}\Pr(\frac{Y^2}{Y^2+6} < \frac{1.4^2}{1.4^2+6}) = \frac{1}{2} + \frac{1}{2}\Pr(U < 0.24623)$ where $U \in \mathsf{beta}(\frac{1}{2}, 3)$.

Answer: $\frac{1}{2} + \frac{1}{2}\frac{\Gamma(\frac{7}{2})}{\Gamma(\frac{1}{2})\Gamma(3)} \int_0^{0.24623} u^{-\frac{1}{2}}(1-u)^2\,du = \frac{1}{2} + \frac{\frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}}{4}\left[\frac{u^{\frac{1}{2}}}{\frac{1}{2}} - 2\frac{u^{\frac{3}{2}}}{\frac{3}{2}} + \frac{u^{\frac{5}{2}}}{\frac{5}{2}}\right]_{u=0}^{0.24623} =$ 89.45%.

(b) Find $\Pr(Y < -2.1)$ where $Y \in \mathsf{t}_7$.

Solution: $= \frac{1}{2} - \frac{1}{2}\Pr(-2.1 < Y < 2.1) = \frac{1}{2} - \frac{1}{2}\Pr(\frac{Y^2}{Y^2+7} < 0.38650) = \frac{1}{2} - \frac{1}{2}\frac{\Gamma(4)}{\Gamma(\frac{1}{2})\Gamma(\frac{7}{2})} \int_0^{0.3865} u^{-\frac{1}{2}}(1-u)^{\frac{5}{2}}du = \frac{1}{2} - \frac{1}{2}\cdot\frac{6}{\frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\cdot\pi} \int_0^{\arcsin\sqrt{0.3865}} \frac{1}{\sin t}\cdot\cos^5 t\cdot 2\sin t\cos t\,dt = \frac{1}{2} - \frac{16}{5\pi}\int_0^{0.6709}\cos^6 t\,dt = \frac{1}{2} - \frac{16}{5\pi}\int_0^{0.6709}\frac{\cos 6t + 6\cos 4t + 15\cos 2t + 10}{32}\,dt =$ 3.69% [the coefficients of the $\cos^6 t$ expansion correspond to one half row of Pascal's triangle, divided by 2^{6-1}].

Note that when $n \geq 30$ the t-distribution can be closely approximated by $\mathcal{N}(0,1)$.

5. And finally, we introduce the **Fisher's ▶F-distribution◀**

(notation: $\mathsf{F}_{n,m}$ where n and m are its two parameters, also referred to as 'DEGREES OF FREEDOM'), defined by $Y_1 = \frac{\frac{X_1}{n}}{\frac{X_2}{m}}$ where X_1 and X_2 are *independent*, both having the chi-square distribution, with degrees of freedom n and m, respectively.

First we solve for $x_2 = y_2$ and $x_1 = \frac{n}{m}\,y_1\,y_2 \Rightarrow$ Jacobian equals to $\frac{n}{m}\,y_2$. Then we substitute into $\frac{x_1^{\frac{n}{2}-1}e^{-\frac{x_1}{2}}}{\Gamma(\frac{n}{2})\,2^{\frac{n}{2}}}\cdot\frac{x_2^{\frac{m}{2}-1}e^{-\frac{x_2}{2}}}{\Gamma(\frac{m}{2})\,2^{\frac{m}{2}}}$ and multiply by this Jacobian to get $\frac{(\frac{n}{m})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})\,2^{\frac{n+m}{2}}}\,y_1^{\frac{n}{2}-1}\cdot y_2^{\frac{n+m}{2}-1}e^{-\frac{y_2\,(1+\frac{n}{m}y_1)}{2}}$ with $y_1 > 0$ and $y_2 > 0$. Integrating over y_2 (from 0 to ∞) yields the following formula for the corresponding **pdf**

$$f(y_1) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})}\,(\frac{n}{m})^{\frac{n}{2}}\cdot\frac{y_1^{\frac{n}{2}-1}}{(1+\frac{n}{m}y_1)^{\frac{n+m}{2}}}$$

for $y_1 > 0$. As a by-product we get (after the $\frac{n}{m}\,y_1 \to ax$ substitution) the value of the following integral: $\int_0^\infty \frac{x^{\frac{N}{2}-1}}{(1+ax)^{\frac{N+M}{2}}}\,dx = \frac{\Gamma(\frac{N}{2})\,\Gamma(\frac{M}{2})}{\Gamma(\frac{N+M}{2})\,a^{\frac{N}{2}}}$ for *any* $N > 0$, $M > 0$ and $a > 0$.

With its help we can find $\mathbb{E}(Y) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})}\,(\frac{n}{m})^{\frac{n}{2}}\int_0^\infty \frac{y^{\frac{n}{2}}\,dy}{(1+\frac{n}{m}y)^{\frac{n+m}{2}}} =$ [take $N = n+2$, $M = m-2$ and $a = \frac{n}{m}$] $\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})}\,(\frac{n}{m})^{\frac{n}{2}}\cdot\frac{\Gamma(\frac{n+2}{2})\,\Gamma(\frac{m-2}{2})}{\Gamma(\frac{n+m}{2})\,(\frac{n}{m})^{\frac{n+2}{2}}} =$

$$\frac{\frac{n}{2}}{(\frac{m}{2}-1)\cdot\frac{n}{m}} =$$

$$\frac{m}{m-2} \qquad \text{(mean)}$$

for $m \geq 3$ (the mean is infinite for $m = 1$ and 2).

Similarly $\mathbb{E}(Y^2) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})}\,(\frac{n}{m})^{\frac{n}{2}} \int_0^\infty \frac{y^{\frac{n}{2}+1}\,dy}{(1+\frac{n}{m}y)^{\frac{n+m}{2}}} =$ [take $N = n+4$, $M = m-4$ and $a = \frac{n}{m}$] $\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})}\,(\frac{n}{m})^{\frac{n}{2}} \cdot \frac{\Gamma(\frac{n+4}{2})\,\Gamma(\frac{m-4}{2})}{\Gamma(\frac{n+m}{2})\,(\frac{n}{m})^{\frac{n+4}{2}}} = \frac{(\frac{n}{2}+1)\cdot\frac{n}{2}}{(\frac{m}{2}-1)\cdot(\frac{m}{2}-2)\cdot(\frac{n}{m})^2} = \frac{(n+2)\,m^2}{(m-2)\,(m-4)\,n} \Rightarrow Var(Y) = \frac{(n+2)\,m^2}{(m-2)\,(m-4)\,n} - \frac{m^2}{(m-2)^2} = \frac{m^2}{(m-2)^2}\cdot\left[\frac{(n+2)\,(m-2)}{(m-4)\,n} - 1\right] =$

$$\frac{2\,m^2\,(n+m-2)}{(m-2)^2\,(m-4)\,n} \qquad \text{(variance)}$$

for $m \geq 5$ [infinite for $m = 1, 2, 3$ and 4].

Note that the distribution of $\frac{1}{Y}$ is obviously $\mathsf{F}_{m,n}$ [degrees of freedom reversed], also that $\mathsf{F}_{1,m} \equiv \frac{\chi^2_1}{\frac{\chi^2_m}{m}} \equiv \frac{Z^2}{\frac{\chi^2_m}{m}} \equiv \mathsf{t}^2_m$, and finally when both n and m are large (say > 30) then Y is **approximately normal** $\mathcal{N}\left(1, \sqrt{\frac{2(n+m)}{n\cdot m}}\right)$.

The last assertion can be proven by introducing $U = \sqrt{m}\cdot(Y-1)$, getting its pdf: (i) $y = 1 + \frac{u}{\sqrt{m}}$, (ii) substituting: $\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})}\,(\frac{n}{m})^{\frac{n}{2}} \cdot \frac{(1+\frac{u}{\sqrt{m}})^{\frac{n}{2}-1}}{(1+\frac{n}{m}+\frac{n}{m}\frac{u}{\sqrt{m}})^{\frac{n+m}{2}}} \cdot \frac{1}{\sqrt{m}}$ [the Jacobian] $= \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\,\Gamma(\frac{m}{2})\,\sqrt{m}} \cdot \frac{(\frac{n}{m})^{\frac{n}{2}}}{(1+\frac{n}{m})^{\frac{n+m}{2}}} \cdot \frac{(1+\frac{u}{\sqrt{m}})^{\frac{n}{2}-1}}{(1+\frac{n}{n+m}\frac{u}{\sqrt{m}})^{\frac{n+m}{2}}}$ where $-\sqrt{m} < u < \infty$. Now, taking the limit of the last factor (since that is the only part containing u, the rest being only a normalizing constant) we get [this is actually easier with the corresponding logarithm, namely $(\frac{n}{2}-1)\ln(1+\frac{u}{\sqrt{m}}) - \frac{n+m}{2}\ln(1+\frac{n}{n+m}\frac{u}{\sqrt{m}}) = -\frac{u}{\sqrt{m}} - \left[(\frac{n}{2}-1) - \frac{n^2}{2(n+m)}\right]\cdot\frac{u^2}{2m} - \ldots. = -\frac{u}{\sqrt{m}} + \frac{u^2}{2m} - \frac{n}{n+m}\frac{u^2}{4} - \ldots. \underset{n,m\to\infty}{\longrightarrow} -\frac{1}{1+\frac{m}{n}}\,\frac{u^2}{4}$ [assuming that the $\frac{m}{n}$ ratio remains finite]. This implies that the limiting pdf is $C\cdot e^{-\frac{u^2 n}{4(n+m)}}$ where C is a normalizing constant (try to establish its value). The limiting distribution is thus, obviously, $\mathcal{N}\left(0, \sqrt{\frac{2(n+m)}{n}}\right)$. Since this is the (approximate) distribution of U, $Y = \frac{U}{\sqrt{m}} + 1$ must be also (approximately) normal with the mean of 1 and the standard deviation of $\sqrt{\frac{2(n+m)}{n\cdot m}}$. □

We must now learn to compute the corresponding **probabilities**. We use the following observation: $\frac{n}{m}Y \equiv \frac{\chi^2_n}{\chi^2_m} \Rightarrow$

$$\frac{\frac{n}{m}Y}{1+\frac{n}{m}Y} \qquad \text{(beta)}$$

$\equiv \frac{\chi_n^2}{\chi_n^2+\chi_m^2} \in \mathsf{beta}(\frac{n}{2},\frac{m}{2})$. And we know how to deal with the last distribution.

Example:

Find $\Pr(Y<4)$ where $Y \in \mathsf{F}_{9,4}$.

Solution: $= \Pr\left\{\frac{\frac{9}{4}Y}{1+\frac{9}{4}Y} < \frac{\frac{9}{4}\cdot 4}{1+\frac{9}{4}\cdot 4}\right\} = \Pr(U<0.9)$ where $U \in \mathsf{beta}(\frac{9}{2},2)$.

Answer: $\frac{\Gamma(\frac{13}{2})}{\Gamma(\frac{9}{2})\cdot\Gamma(2)} \int_0^{0.9} u^{\frac{7}{2}}(1-u)\,du = \frac{11}{2}\cdot\frac{9}{2}\cdot\left[\frac{u^{\frac{9}{2}}}{\frac{9}{2}} - \frac{u^{\frac{11}{2}}}{\frac{11}{2}}\right]_0^{0.9} = 90.25\%$. ■

We will see more examples of the F, t and χ^2 distributions in the next chapter, which discusses the importance of these distributions to Statistics, and the context in which they usually arise.