BIOSTATISTICS
TOPIC 3: INTRODUCTORY PROBABILITY

I. INTRODUCTION

Sir George Pickering, a prominent British medical researcher, once noted that "doctors want to help patients, but the extent to which they can help obviously depends on the doctor's knowledge. But knowledge is a matter of probability. Diagnosis is a matter of probability, and in judging treatment, doctors have to base their judgment on knowledge of probability". Some of you may have reservation about this comment, but it could be argued that the reality of the world is harsh and unyielding, and must be dealt with on its own terms. We work in a world of randomness, and there is no way to eliminate completely the risks of being wrong. I think our real problem is not how to eliminate them, but how to live with them intelligently. In medical research, things do not always work out the way we hypothesized or we planned. The main reasons for this are likely that (i) our hypothesis is incorrect and/or (ii) we do not have enough evidence to reject/accept the hypothesis. The former is hypothetical idea which can be re-defined, however, the latter is fact (nothing but the fact) and can not be changed but can be dealt with in probabilistic terms.

In the last topic, we have been concerned with the area of statistics generally known as descriptive analysis. We mentioned that statistical inference make guesses about a population by using information obtained from a sample taken randomly from the population. Statistical inference is largely based on probability theory, primarily because probability theory provides a means of determining the reliability of inferences. In this topic, we will introduce the basic concepts of probability theory to understand the conclusions that result from the application of statistical techniques to data analysis as well as the reasons behind the requirements for probability sampling in the collection of data.

Before introducing the operation of probability, we will survey briefly some main ideas such as set theory, events, permutation and combination.

II. SET THEORY, NOTATION AND CONCEPTS

DEFINITION: *A set is a collection of well-defined, distinct objects with common characteristics. The objects are called* ***elements***.

The phase *well-defined* indicates that we must be able to determine with certainty whether or not a given object belongs to the set under study. Thus, in the "*set of women who had fractures*" is well-defined because we do know what a fracture is. However, the set of "*dishonest salesmen in Australia*" is not well-defined because there is no universal standard by which to gauge the virtue of honesty or dishonesty.

At this stage, let us denote a set by an uppercase letter and its elements by lowercase letters. For example, a set of vowels: A = {*a, e, i, o, u, . . .*}.

(a) **Equality of sets**. Two sets A and B are equal if every element of set A is equal to every element of set B. For example, A={*a, e, i, o, u*} is equal to set B={*i, e, u, a, o*}.

(b) **Subset of set.** If every element of set A is belong to the set B, then A is called a subset of B. For example, if A={*k, l, m, n*} and B={*m, r, l, n, k*}, then A is being to B (which is written as $A \in B$).

(c) **Union of sets.** Given two sets A and B, the *union* of A and B ($A \cup B$) is the set consisting of elements that belongs to set A and/or set B. For example, if A={*a, e, i*} and B={*c, d, e, i*}, then $A \cup B$={*a, e, i, c, d*}.

(d) **Intersection of sets.** Given two sets A and B, the *intersection* of A and B ($A \cap B$) is the set consisting of elements that belongs to both set A and set B. For example, if A={*a, e, i*} and B={*a, c, d, e, i*}, then $A \cap B$={*a, e, i*}.

(e) **Complement**. The complement of a set contains all elements not in the set, but still in the universe. *A*' is denoted to be complement of *A*. So, if *U*={1, 2, 3, 4, 5, 6, 7, 8, 9} and A={1, 2, 3, 4}, then clearly *A* is a subset of the universe *U*, and the complement of A is *A*'={5, 6, 7, 8, 9}.

III. SAMPLE SPACE AND EVENTS

(A) SAMPLE SPACE

DEFINITION: *The sample space is defined as the set of all possible outcomes of an experiment*.

The term *experiment* here must be understood from a general point of view, not just in terms of biological viewpoint. Throwing a die can be regarded as an experiment, as is the study of incidence of fractures. According to this definition, when a die is thrown, the sample space is 1, 2, 3, 4, 5 and 6 (since the die has six faces). Likewise, when 2 coins are tossed, the sample space must be: head-head, head-tail, tail-head and tail-tail. Similarly in a study of fracture, the sample space is those who have fracture and those who do not have fracture.

We often denote the sample space by capital letter ie. X.

(B) EVENTS

DEFINITION: *An event is a possible outcome of an experiment*.

We will denote the possible outcome of an experiment by x_i, where i runs from 1 to n. Of course, x_i is a subset of a sample space (by the definition of set theory we learned earlier). We will now survey a number of types of events.

Mutually Exclusive Events are events that do not have common sample points. That is if A = {5, 6, 7} and B = {1, 2, 3, 4}, then $A \cap B = \varnothing$ (no intersection).

Not Mutually Exclusive Events. Two events are said to be not mutually exclusive if they contain one or more common sample points. For example, let A = {5, 6, 7} and B = {7, 8, 9, 10} then $A \cap B =$ [illegible] are mutually inclusive.

IV. COUNTING TECHNIQUE I: PERMUTATION

(A) SELECTION WITH REPLACEMENT. When selecting objects with replacement, the first object can be selected in n different ways, since there are n objects in the set from which to choose, and any of them can be chosen. Since the selected object is returned to the set before making the second selection, the second object can also be selected in n different ways. Similarly, each of the other r objects can be chosen in the same n ways. Hence, r objects together can be selected in a number of ways equal to the product:

$$\underbrace{n \times n \times n \ldots \times n}_{r \text{ factors}} = n^r$$

Example 1: How many 3-digit numbers can be formed from the digits 2, 4, 6, 7 and 9 ?

There are three positions to be filled in a 3-digit number. Each position can be filled in 5 different ways. Thus, the three positions can be filled in $5 \times 5 \times 5 = 125$ ways.

(B) SELECTION WITHOUT REPLACEMENT. If an event can happen in m different ways, and after this was occurred, another event can happen in n different ways, then the two events can happen in mn different ways.

Furthermore, if a first event can occur in n_1 ways, a second event can occur in n_2 ways, a third event can occur in n_3 ways, and so on, then the number of ways for these events to occur in succession is $n_1 \times n_2 \times n_3 \times \ldots$ ways.

Example 2: If we have available 10 rats for experimental purposes and wish to select 3 of the rats for three different experiments. In how many ways can the selection be made?

The first rat can be selected in 10 different ways, since any one of the available rats can be selected. Having selected the first rate, there remains 9 rats for selection, so the second rat can be selected in 9 different ways. Similarly, the third rat can be

selected in 8 different ways. Consequently, the number of ways in which three rats can be chosen is given by $10 \times 9 \times 8 = 720$ ways. //

When counting numbers of outcomes, the following notation is extremely useful: the product of the first n natural numbers is called **factorial *n*** and is denoted by $n!$ Thus

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1 .$$

In fact we also can write: $n! = n(n - 1)!$.

(C) SELECTION OF *r* OBJECTS FROM *n* OBJECTS WITHOUT REPLACEMENT.

We now consider the case in which the r objects are selected one-by-one without replacement from n objects. That is, the first object is not replaced before the second is selected, the first two are not replaced before the third is selected, and so on. A selection made in this way is called a **permutation (**or an **ordered selection**) of r objects from n.

As before, the first object can be selected in n ways. However, the second object can be selected in $(n - 1)$ ways. Similarly, the third can be selected in $(n - 2)$ ways, and so on. The last object can be chosen in $(n - r + 1)$ ways. Thus the number of ways in which the whole permutations can be selected is given by the product:

$$n(n - 1)(n - 2). \ldots .(n - (r - 1)).$$

This product is denoted by P_r^n and is read as "*the number of permutations of r objects from n possible objects*". This is given by:

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

Example 3: From a group of 8 persons, it is required to select individuals to participate in 5 different tests. How many ways can the selection be done?

Since the tests are different, the order in which the 5 are chosen is significant. The number of ways in which the choice can be made is therefore the number of permutations of 5 out of 8 subjects, which is actually:

$$P_5^8 = \frac{8!}{(8-5)!} = \frac{8!}{3!} = 5! = 6720 \text{ . //}$$

V. COUNTING TECHNIQUE II: COMBINATION

Suppose that we have 7 patients A, B, C, D, E, F and G. The number of *ordered* selections of 3 patients from the 7 is P_3^7. Now consider the all the subset of patients A, B, C, say. As ordered selection, there are 3! such subsets (ie. ABC, ACB, BCA, BAC, CAB, CBA). But only who in the subgroup is of interest, then these 3! permutation is just one combination (one *subgroup*), since all permutations have exactly the same patients A, B and C.

Hence, the total number of *subgroups* of 3 patients possible from 7 patients is equal to $\frac{1}{3!}$ of total ordered selections: $\frac{1}{3!} P_3^7$; that is:

$$C_3^7 = \frac{7!}{3!(7-3)!}$$

or generally

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

and is read as: "*the number of combinations of r objects from n objects*".

Example 4: Four rats are selected for experiment from a group of 6 white and 4 brown rats. In how many ways can the selection be made so that the selected group contains: (a) two brown rats; (b) at least two brown rats ?

(a) We have to select, in all, 4 rats for the experiment. Since the selection has to contain 2 brown rats, so other 2 must be white. Now the number of selections of 2 brown rats from 4 is given by: C_2^4; and the number of selections of 2 brown rats from 6 is given by: C_2^6. In total, we can select by $C_2^4 \times C_2^6 = 90$ ways.

(b) In this case, we have to select at least two brown rats, the selection can have either (i) 2 brown and 2 white, (ii) 3 brown and 1 white or (iii) 4 brown (and no

white) rats. Case (i) can have 90 different ways (see (a)). Case 2 can have $C_3^4 \times C_1^6 = 24$ ways. Case (iii) can have $C_4^4 \times C_0^6 = 1$ way. Thus, in total, we have 90+24+1 = 115 ways of selection. //

There are many applications for combination and permutation, however, we are not going to these as they are either too complicated or beyond the scope of this introductory topic.

VI. PROBABILITY

DEFINITION: *If an event A can occur in n equally likely outcomes, n_A of which have attribute A, then we can say that attribute A has a probability of n_A/n.*

It is somewhat conventional to denote the statement "the probability that A occurs" by P(A).

It follows from this definition that, probability is a number between 0 and 1 which expressed the chance that a specific event occurs under a stated condition. It is also clear that the probability of an event other than A occurs is 1-P(A). This is called **complementary** probability.

Observations of phenomena can result in many different outcomes, some of which are more likely than others. For example, if we throw a fair die and suppose that the number 7 turns up, what is the chance that, if we throw the die again, the number 7 will again turn up. There have been arguments that the number 7 is likely to turn up because it has turned up. On the other hand, there have been counter-argument that since the number 7 has turned up, therefore, the chance it turns up again is less than other numbers. A number of attempts have been made to give a precise definition for probability of an outcome. We will discuss a few of these:

Classical interpretation arose from games of chance. Typical probability statement of this type are "the probability that a flip of coin will show 'head' is 1/2 and the probability of drawing an ace is 4/52". The numerical values of these probabilities arise from the nature of the games. A coin flip has only two possible outcomes: head or tail; so the probability of a head should be 1/2. Similarly there

are 4 aces in a standard deck of 52 cards, so the probability of drawing an ace in a single draw is 4/52. In this classical concept, each possible distinct result is called an *outcome*; an *event* is identified as a collection of outcomes. The application of this interpretation depends on the assumption that all outcomes are equally likely. If this assumption does not hold, the probabilities indicated by the classical interpretation will be in error.

Relative frequency concept of probability is an empirical approach to probability. If an experiment is repeated a large number of times and event E occurs in 30% of the times, then 0.30 should be a very good approximation to the probability of event E. Symbolically, if an experiment is conducted n different times and if the event E occurs n(E) of these trials, then the probability of event E is approximately n(E)/n.

We say "approximate" because we think of the actual probability P(event E) as the relative frequency of the occurrence of the event E over a very large number of observations or repetitions of the phenomenon. The fact that we can check probabilities that have a relative frequency interpretation (by simulating many repetitions of the experiment) make this concept very attractive and practical.

The personal or subjective probability can be applied in situations in which it is difficult to imagine a repetition of an experiment. These are "one shot" experiments. For example, a doctor estimates the probability of survival after an operation on a patient A would not be thinking of a long series of repeated operations on the patient. Rather, he would use a subjective (personal) probability to make a one-shot statement of belief regarding the likelihood passage of the proposed operation. The problem with subjective probability is that they can vary from person to person and they can not be checked.

Example 5: In a district, there are m people with fractures and n people without fracture. What is the probability that 10 persons, selected randomly from this district, are fracture subjects.

Let event A = {fracture}. The number of selections of 10 fracture people from (m + n) people is: C_{10}^{m+n}. On the other hand, the number of selections of 10 fracture

subjects from *m* fracture subjects (event A) is: $n_A = C_{10}^m$. Hence, the probability that 10 subjects are chosen is: $P(A) = \frac{C_{10}^n}{C_{10}^{m+n}}$.

Example 6: Suppose that in a lottery ticket, there are 40 numbers, of which, there are 6 winning numbers in each draw. If we buy one ticket, what is the probability that:
(a) we win 4 out of 6 numbers;
(b) we win 5 out of 6 numbers;
(c) we win all 6 numbers.

The problem is equivalent to drawing 6 balls from an urn in which there are 6 white balls (winning numbers) and 40-6=37 black balls (not winning). In (a) the number of selections 6 numbers out of 40 numbers is C_6^{40}; and the number of selections of 4 winning numbers from 6 winning numbers and 2 numbers from 34 non-winning numbers is $C_4^6 \times C_2^{34}$; then the probability of winning 4 numbers out of 6 is: $\frac{C_4^6 \times C_2^{34}}{C_6^{40}}$. Similarly for (b) the probability is: $\frac{C_5^6 \times C_1^{34}}{C_6^{40}}$. And finally, the event of winning all 6 numbers is $\frac{C_6^6 \times C_0^{34}}{C_6^{40}} = \frac{1}{C_6^{40}}$ //

We now examine a number of properties and theorems probability follow from the concepts and postulates presented in the last few sections:

(A) JOINT, MARGINAL AND CONDITIONAL PROBABILITY

Example 4: Consider the following table which tabulates the number of women with and without fracture according to age group:

Table 1: Incidence of fractures among women classified by age group

Event	Age			
	70-79	80-89	90+	Total
Fracture	7	7	8	23
Non fracture	43	23	12	77
Total	50	30	20	100

(i) In statistics, we call this tabulation a **bivariate** sample space because the basic outcome (fracture versus non-fracture) can be considered a **joint outcome** of the second variable *age*. The probability of the joint outcome is referred to as a **joint probability.** For instance, P(Fracture $\cap$ 70-79) = 7/100 = 0.07 is a joint probability of fracture for a women aged between 70-79 years.

(ii) With bivariate sample space, we are also frequently interested in the probability distribution of an individual variable considered separately. For instance, for the above data, the marginal probability of fracture is:

$$\text{P(fracture)} = \frac{7}{100} + \frac{7}{100} + \frac{8}{100} = \frac{23}{100} = 0.23$$

Since the probabilities obtained by summing across either one of the classifications are shown in the margins, they are called the **marginal probabilities**.

(iii) **Conditional probability**. Often, we want to know the probability of one event occurring, given that a second event occurs. We denote the statement "the probability that A occurs given that B occurs" as P(A | B), where the "|" is

equivalent to the expression "given". The conditional probability is defined as follows:

If A and B are any two events of a sample space and P(B) is not equal to 0, the conditional probability of A given B, is: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$.

It follows from this definition that in Table 1, the probability that a woman will have fracture given that she is 90+ years of age is equal to:

$$P(\textit{fracture} \mid 90+) = \frac{P(\textit{fracture} \cap 90+)}{P(90+)} = \frac{8/100}{20/100} = 0.40.$$

Similarly: $P(\textit{fracture} \mid 70\text{-}79) = \frac{7/100}{50/100} = 0.14.$

On the other hand:

$$P(70\text{-}79 \mid \textit{fracture}) = \frac{P(70-79 \cap \textit{fracture})}{P(\textit{fracture})} = \frac{7/100}{23/100} = 0.30.$$

(B) THE ADDITION THEOREM. For any two events A and B of a sample space,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ if A and B are not mutually exclusive
$P(A \cup B) = P(A) + P(B)$ if A and B are mutually exclusive.

For example, referring to data in Table 1 again, let the events A={70-79} and B={*fracture*}. Then by definition, we have:

P(*fracture* OR 70-79) = P(A $\cup$ B)
= P(A) + P(B) - P(A$\cap$ B)
= 0.50 + 0.23 - 0.07
= 0.66.

However, in that table the events of fracture and the event of a person aged 80+ are mutually exclusive. Now P(*fracture*) = 0.23 and P(80+) = 0.20 + 0.30 = 0.50, hence, the probability that P(*fracture* $\cup$ 80+) = 0.23 + 0.50 = 0.73.

(C) THE MULTIPLICATION THEOREM

For any two events A and B of a sample space,

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

Example 5: Let A and B denote the events of osteoporosis and death, respectively. Suppose that the probability of death is 0.80. Suppose further that the conditional probability of death given having osteoporosis is 0.90. We wish to find the probability of developing osteoporosis AND death.

It is perhaps easier to write these information in symbols: P(A) = 0.15, P(B) = 0.80 and P(B | A) = 0.90. Then what we wish to find could be expressed as:

$$\begin{aligned} P(A \cap B) &= P(A).\ P(B \mid A) \\ &= 0.15 \times 0.90 \\ &= 0.135 \end{aligned}$$ //

(D) INDEPENDENT EVENTS

In statistics, when we talk about two independent events A and B when we mean:

$$P(A \mid B) = P(A) \text{ or } P(B \mid A) = P(B)$$

Or more popularly:

$$P(A \cap B) = P(A) \times P(B).$$

Example 6: The following table shows the relationship between length of stay and insurance status.

Table 2: Length of Stay (days) and Insurance Status

LOS	Insurance		
	Insured	Uninsured	Total
<5	**0.42**	0.18	**0.60**
5-10	0.21	0.09	0.30
>10	0.07	0.03	0.10
Total	**0.70**	0.30	1.00

In this Table, the event "<5 days" and the event "*insured*" are independent since P(<5 and *Insured*) = 0.60 × 0.70 = 0.42.

VII. SPECIFICITY AND SENSITIVITY

The terms *specificity* and *sensitivity* were introduced by Yerushamly back in 1947 to describe the efficiency of a diagnostic test. Since then a number of new terms have been coined and admittedly tend to confuse statistical users than to clarify the usage of these indices. In this section, we will survey these terms and introduce the use of sensitivity and specificity as in evaluating a screening test, which we have briefly touched upon.

Consider the following scenario. We have a test which can predict whether a subject will develop osteoporosis. The test gives either positive (P) or negative result (N). Suppose having administered the test and we follow a group patients for a period and observe the present (P) or absence (N) of osteoporosis. We can represent the result of this study as follows:

Result of test	Confirmed after follow-up P	N	Sum
P	a	b	a+b
N	c	d	c+d
Sum	a+c	b+d	N

where *a, b, c* and *d* are frequencies of observations and N=*a*+*b*+*c*+*d* is the total number of patients.

We could describe the table in a rather probabilistically friendly format as follows:

Result of test	Confirmed after follow-up P	N
P	true positive	false positive
N	false negative	true negative

(a) According to classical definition, the **sensitivity** (s) = (**true positive rate**) of the test is the probability of having osteoporosis given a positive test result:

$$s = \frac{a}{a+c}$$

(b) And the **specificity** (f) = (**true negative rate**) is defined as the probability of having no osteoporosis given that the test is negative:

$$f = \frac{d}{b+d}$$

(c) **Predictive value**. When we started the study with a population whose condition had already been confirmed, but the doctor who later uses the test starts with patients whose condition is not known. Yet, the purpose of the test is to predict what the patient's condition really is. So, the doctor wants to know its predictive accuracy or how well the test would perform for an unknown (or indeed any) patient. If the test is positive, is the patient actually has osteoporosis? If the test is negative, is the actual osteoporosis is likely to be absent?

To answer these question we calculate the **positive predictive value** (denoted by v)as the index of positive accuracy (v) as follows:

$$v = \frac{a}{a+b}$$

and the **negative predictive value** as the index of negative accuracy (denoted by g):

$$g = \frac{d}{c+d}$$

Occasionally, the positive predictive and negative predictive values are referred to as **posterior probability of disease** and **posterior probability of no disease,** respectively.

(d) It is obvious from the above table that the prevalence of the disease could be estimated as the of pre-test likelihood of disease (or prior probability of disease), which is:

$$\frac{a+c}{N}.$$

(e) The **likelihood ratio (LR)** is a measure of discriminant by a test result. A LR of greater than 1 raises the probability of disease and is often referred to as "positive" test result. A LR of less than 1 is usually referred to as "negative" test result.

$$LR = \frac{sensitivity}{1 - specificity}$$

VII. BAYES' THEOREM

Up to now, we know that P(D | S) is generally different to P(S | D) for any two events D and S.

We now consider the following typical medical problem. Let **D** denote the presence of a disease and **ND** the absence of the disease. Let **S** denote the symptom of the disease. It is often that we know the prevalence of the disease in the general population **P(D)**, the probability that a subject with disease will exhibit the symptom **P(S | D)**, and the probability that a subject without disease will exhibit the symptom **P(S | ND)**. We wish to find P(D | S), the probability of getting the disease given that the subject exhibits the symptom.

Notice that all subjects with S will have either D or ND, so S can be written as "S and D" or "S and ND", since the two events are mutually exclusive:

$$P(S) = P(S \cap D) + P(S \cap ND)$$

But $$P(S \mid D) = \frac{P(S \cap D)}{P(D)}$$

and $$P(S \mid ND) = \frac{P(S \cap ND)}{P(ND)}$$

It follows that, if we multiply by P(D) and P(ND), respectively, that:

$$P(S \cap D) = P(D).\ P(S \mid D)$$

and $$P(S \cap ND) = P(ND).\ P(S \mid ND).$$

But $$P(D \mid S) = \frac{P(D \cap S)}{P(S)} = \frac{P(D \cap S)}{P(S \cap D) + P(S \cap ND)}$$

hence $$P(D \mid S) = \frac{P(D) \times P(S \mid D)}{P(D) \times P(S \mid D) + P(ND) \times P(S \mid ND)}$$

This (last expression) is called Bayes' theorem. Notice that the expression on the right involves only quantities that we have assumed to be (and is usually) known.

Example 7: Suppose that a new screening test is proposed for the detection of fracture. The prevalence of fracture in the general population is known to be 10%. The test has been investigated in fracture subjects and was found to give positive result in 70% of such cases (sensitivity). When given to subjects without fracture, the test yielded a positive result of 20% i.e. specificity is 80%).
We want to know what is the proportion of subjects with positive test who, when followed up, will actually be found to have fracture?

Let us represent the prevalence of fracture by P(D) = 0.10, the probability of being positive in fracture subjects by P(S | D) = 0.70, and the probability of being positive in non-fracture subjects by P(S | ND) = 0.20. Then, the probability that a subject will have fracture given a positive result is:

$$\begin{aligned} \text{P(D | S)} &= \frac{P(D)\times P(S \mid D)}{P(D)\times P(S \mid D)+P(ND)\times P(S \mid ND)} \\ &= \frac{0.1\times 0.7}{(0.1\times 0.7)+(1-0.1)\times 0.2} \\ &= 0.07 / 0.25 \\ &= 0.28. \qquad // \end{aligned}$$

IX. EXERCISES

1. Evaluate the following:
   (a) $\frac{n!}{(n-2)!2!}$ (b) $\frac{7!-5!}{4!}$ (c) $\frac{10!+8!}{8!}$ (d) $\frac{100!}{99!}-\frac{99!}{98!}$
2. Show that $\frac{n^2}{n!}=\frac{1}{(n-1)!}+\frac{1}{(n-2)!}$
3. Solve the equation $C_1^x+C_2^x+C_3^x=\frac{7x}{2}$
4. Simplify (a) C_2^5 (b) C_0^5 (c) C_5^5
5. Simplify (a) C_{n-2}^n (b) C_{n-1}^n
6. Suppose that a dietitian has available the following foods listed by their main vitamin content:

Vitamin A:	**Vitamin B**	**Vitamin C**
lectuce	peanuts	oranges
carrots	peas	lemons
squash	lean meat	
egg yolk	egg white	
butter	liver	
	milk	
	cereal	

   How many meals are possible if each contains

   (a) one food from each vitamin group?
   (b) 3 foods from group A and none from group B or group C ?
   (c) 2 from group, 3 from group B and none from group C ?
   (d) 2 food from group A, 3 from group B and 1 from group C ?
   (e) 4 from group A, 4 from group B and 1 from group C ?.

7. In a group of 15 women, there are 7 osteoporotic women. What is the probability that if 12 women are selected, then there are:
   (a) exactly 6 osteoporotic women (b) at least 6 osteoporotic women.

8. In a group of rats, there are 6 of genetic type R, 4 of type W and 3 of type B. If three are selected randomly from the group. What is the probability that there are:

(a) all R rats (b) all W rats
(c) all the same genetic type (d) different genetic types
(e) 2 of type R and 1 of type W (f) exactly 2 W rats
(g) at least 1 W rat (h) a particular rat is included.

9. There were 10 male and 5 female rats in a cage. If two animals are drawn out from the cage in blind and randomly, what is the probability that:
(a) both animals are females (b) both animals are males
(c) at least one animal is male (d) no male.
if the sampling is with replacement (the first rat is returned before a second rat is selected).

10. Under the same condition as in question 9, but the sampling is without replacement. Find the required probabilities (a) to (d).

11. Let A and B denote two genetic characteristics and suppose that the probability is 1/2 that an individual chosen at random will have A, 3/4 that he/she will have B. Assume that these characteristics occur independently. What is the probability that an individual chosen at random will have
(a) both (b) neither (c) exactly one characteristics?

12. Five identical rabbits are in a cage. Some have inoculated against a virus. Find the probability that you select the inoculated rabbits if you select:
(a) one and only one was inoculated
(b) three and two were inoculated
(c) two and two were inoculated

13. Suppose that a certain opthalmic trait is associated with eye colour. Three hundred randomly selected subjects are studied with results as follows:

	Eye colour		
Trait	Blue	Brown	Other
Yes	70	30	20
No	20	110	50

What is:
(a) P(trait=Yes)
(b) P(blue eyes and trait=Yes)
(c) P(blue or brown eyes and trait=yes)
(d) P(brown eyes | trait=Yes).

14. Treatment Y causes a toxic reaction in 25% of persons to whom it is given. What is the probability that 0, 1, 2, 3 or 4 of four persons chosen at random will have a toxic reaction ?

15. Twenty percent of women in a community is diabetic. Of these, 75% have low bone mineral density (BMD). Of those who did not have diabetes, 20% have low BMD. What is the probability that a randomly selected low BMD woman who has diabetes?

16. Let P(A | B) = 0.2, P(not A | not B) = 0.4 and P(B) = 0.3.
(a) Use Bayes' theorem to calculate P(B | A)
(b) Construct the 2x2 probability table with column (A, not A) and rows (B, not B). Compute P(B | A) directly from the table.

17. One problem with using the angiogram to diagnose stroke is the slight risk of death associated with this test (<1%). Some studies have attempted to use the PET scanner (which measure blood flow in the brain) to detect stroke disease non-invasively as an alternative to the angiogram. A comparison was made on the same patients between these two methodologies for detecting stroke, with the results given in the following table:

	Angiogram result	
PET scan	+	-
+	21	3
-	8	32

Let us now regard the angiogram as the definitive test. Using the above data to calculate relevant statistics and comment on the new (PET scan) test.

18. Consider the following strategy for the diagnosis of pancreatic cancer using 4 tests (ERP, US, PFT and ANG).

(i) If two or more of the tests are positive, diagnose pancreatic cancer;

(ii) If more than two of the tests are negative, diagnose no pancreatic cancer.

(a) Set up a 2x2 table for this strategy, cross-classifying diagnosis with pancreatic cancer for the data shown in Table 1:

(b) Compute the sensitivity, specificity and predictive values of this strategy.

(c) How does this strategy compare to ERP alone?

Table 1: Four test results and surgical diagnosis in 42 patients who had all 4 tests performed (NEJM 1977).

					Pancreatic disease		Other diseases	
Comb	ANG	ERP	US	PFT	Cancer	Inflam.	Cancer	Inflamm.
A	+	+	+	+				12
B	-	-	-	+			2	12
C	-	-	+	-			1	1
D	-	+	-	-				
E	+	-	-	-				3
F	-	-	+	+				
G	-	+	-	+	1			
H	+	-	-	+				
I	-	+	+	-				
J	+	-	+	-			1	
K	+	+	-	-	2			
L	-	+	+	+		2		
M	+	-	+	+		2		
N	+	+	-	+	1			
O	+	+	+	-		2		
P	+	+	+	+	10			
Total					**14**	**6**	**4**	**18**

19. About 10% of young adults (aged 11-30) with sore throats have streptococcal pharyngitis, as indicated by a positive throat culture. Investigation has indicated that a new test called Gram stain of the pharyngeal exudate. The following table shows the sensitivity and specificity of the Gram stain, several signs, and history of exposure to a family member with streptococcal pharyngitis.

	Fever >38^o	Cervical adenopathy	Pharyngeal exudate	History of exposure	Positive Gram stain
Sensitivity	0.33	0.73	0.45	0.18	0.73
Specificity	0.89	0.55	0.78	0.92	0.96

(a) Compute the positive and negative likelihood ratios.
(b) A 20-year old man with a sore throat has a fever of 39^o . What is the probability of streptococcal pharyngitis.
(c) A 20-year old man with a sore throat has a sister with streptococcal pharyngitis. He has a fever of 39^o . What are the odds and probability of streptococcal pharyngitis.
(d) Which of the test do you think are independent?
(d) Which is the best test according to these data?