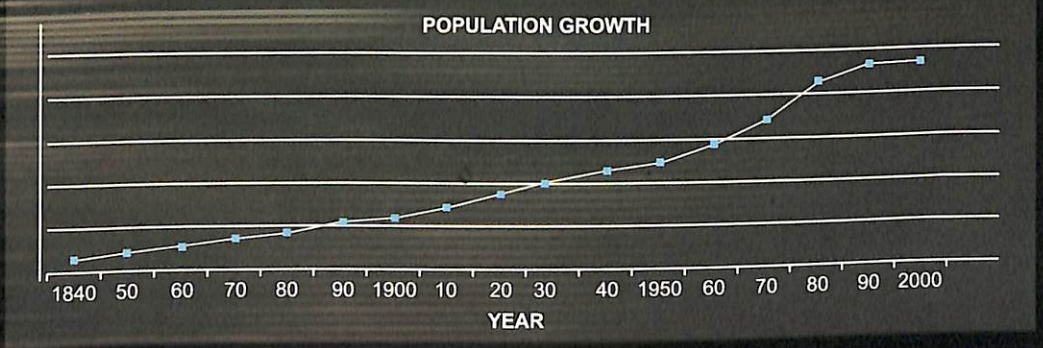
MEDICAL STATISTICS

ANIL GORE - SHARAYU PARANJPE - MADHAV KULKARNI



PUBLIC HEALTH
Healthy People In Healthy Communities

Share
Facilitate
Inform
Promote

SIPF ACADEMY

Medical Statistics

(An introductory text designed for a one semester elective course for B.Sc. Statistics)

By

Anil P Gore

Vice-President, Bio-statistical Services

CYTEL Statistical Software and Services Pvt. Ltd. Pune

Sharayu Paranjpe

Principal Statistician, CYTEL Statistical Software and Services Pvt. Ltd. Pune

Madhav B Kulkarni

Associate Professor and Head, Department of Mathematics and Statistics

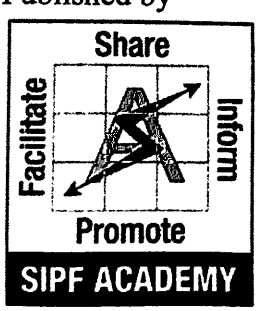
B.Y.K.(Sinnar) College of Commerce, Nashik

SIPF ACADEMY

Second Edition

Reprint: 4 April 2015

Published by



SIPF ACADEMY

Publishers and Consultants,
Mrs. Madhuri Madhav Kulkarni

2, Swargandh Apartments, Opposite Tanishq show room,
Samarth Nagar, Nashik 422 005, (Maharashtra State, India)

Rs. 300/- or **20 $** out of India

E-mail: madhurimk19@gmail.com
E-mail: madhavbk@gmail.com

Medical Statistics

Preface

This book is generally about human beings and their health. That description is not precise enough. In fact the book discusses statistical aspects of human populations with emphasis on health related issues. Like in all science, the broad aim here is to identify patterns and relationships among variables. The material is statistical in the sense that patterns or rules will necessarily be about groups of people and not about individuals. This is for two reasons. There are some variables that simply do not make sense in the context of one individual. Consider birth rates and death rates. An individual does not have a birth or death rate. These concepts arise only in the context of populations. But even if the variable can be meaningful for just one subject, the only patterns possible may be for groups. Consider lifespan. It is duration between birth and death. It can be calculated for any (dead) individual. However, we will only examine patterns in life span for a population. All the rules identified will be about populations.

This book is intended to be a text book for a one semester elective course for students of Third Year B.Sc. (Statistics) in Indian Universities. The elective course is designed to prepare students for work in the field of health, with particular focus on clinical trials. (Draft syllabus and suggestions for evaluation are given below.) Writing of this book is prompted by the substantial increase in the volume of outsourced work on analysis of clinical trials data, being done in India. This increase has occurred mainly during the first decade of the new century. Many multinational companies in the pharmaceutical field (Pfizer, Glaxo-Smith-Klein, Novartis and Bristle-Meyer-Squibb etc) have found it economic to get statistical work (related to clinical trials) done in India. They have either set up their own in-house facility or given out work to contracting companies, leading to an increased demand for young statisticians. This new kind of employment opportunity is expected to grow further in near future. However, there are hardly any training facilities for statistics students to prepare them for such work.

It is the opinion of authors of this book that an elective course can and should be offered to students in their final year of undergraduate education so that they are better equipped to seek jobs in this field. In view of this perspective, second half of the book is devoted to clinical trials. The book tries to combine theory with practice and information. It begins with a discussion about populations. This is the first part of the book. In this part there is some 'theory' about modeling population growth. Then it offers a general description of the population of India. The part includes discussion of models of mortality and hazard rates etc. The second part is more specifically about diseases and epidemics. There is coverage of how to measure morbidity and also how to quantify role of risk factors. Third part of the book focuses on statistics of clinical trials. Process of drug discovery is described briefly. Various technical terms used in the field are explained and then issues related to statistical inference are taken up. Many exercises are provided and it is expected that teachers will show in class, work on one example and then ask everyone to try more problems on their own. In addition to using problems provided in

this book, teachers will have to make up more problems along the way. The book refers to many relevant websites and it is necessary that students as well as teachers visit the websites and read the material available there.

We give here a proposed syllabus for the course.

Title of the course: **Medical statistics.**
The syllabus is in three parts. I Population study II Epidemiology and III Clinical trials. Weight for three parts is approximately, 25%, 25% and 50%.

Part I. Population study- India's population and census, population growth and models for population growth, birth and death rates, survival function, hazard rate (age specific mortality rate), use of exponential and Weibull distribution for modeling hazard rate. Study the following website- http://www.iipsindia.org/dps.htm , www.icmr.nic.in, www.**censusindia**.net, www.**mrcindia**.org , www.nari-icmr.res.in , http://ntiindia.kar.nic.in/ , http://www.nfhsindia.org/, www.**cdc**.gov , Examine a case paper from a general practitioner, an admission paper from a general hospital and plan summary of data from such documents

Part II. Epidemiology, a brief historical review. Use of contingency tables, Odds, odds ratio, relative risk. Estimation of OR. CI for OR. Relation with parameter in a logit model, symmetry in square contingency tables, collapsing tables and Simpson's paradox

Part III. Clinical trials- This part is to be covered in 6 weeks.
Week 1- General information on history of drug discovery including Louis Pasteur (rabies) , Edward Jenner (small pox), Ronald Ross (malaria), Alexander Fleming (penicillin), Jonas Salk (polio), John Snow (Cholera), asthma, diabetes, blood pressure, heart attack, arthritis.
Week2- phases of clinical trial, purpose, duration, cost, drug regulatory bodies, ICH, randomization
Week 3- parallel designs, factorial design, cross over design
Week 4- PK/PD
Week 5- bioequivalence and bio-availability, non-inferiority trial
Week 6- sample size, power

Suggestions for evaluation

It would be appropriate to evaluate students of this course on three different aspects. They are (a) statistical theory, (b) data analysis and interpretation and (c) project. Depending on the flexibility of the examination system, evaluation can be done in many different ways. Every university requires that a final examination be held. In such an examination, questions can be asked on selected sections 1.1, 1.2, 1.3, 2.3, 3.3, 3.5, 3.6, 3.7 and 3.8. Here questions can be of the form 'Prove that..' or 'Derive the following..'. There should also be questions about interpreting outputs of analysis. The question should have a description of a situation, data set and some computer output such as 'ANOVA' table. Short answer questions about what different entries in the output mean are good for testing a student. Another way is to erase some entries and ask students to

compute them using other entries. This tests their understanding of interrelation among entries.

If a practical or laboratory component is allowed in the examination, then students can be asked to carry out specific procedures. The issue of making a good choice of procedures should be made part of the theory paper. In the practical part, that choice should be given in the question itself. Execution and interpretation of results is what a student should be tested on. If a practical component is not feasible, examination should include very small data sets to be manipulated using electronic calculators. Implementation should be easy for a student that understands the material. Some part of the question should be explicitly about interpretation. Lastly, a project would be a very valuable part of the course. Students should visit websites indicated in the text and similar other websites and learn more about the real world. If possible there should be presentations on these materials by students. Lastly, the project component could include work with medical practitioners and hospitals/pathology laboratories etc. Students can try to use case papers and other kinds of documents available to summarize information. 'Survey of oral health among consumers of gutkha' and such contemporary topics can be used for actual field work. This should be done for credit (i.e. as a part of the examination). If that is not feasible, it should be done any way because it constitutes a very valuable educational experience for the students.

In this course, there will be some situations in which new statistical methods (i.e. new to students of T Y B Sc class) will be introduced. In some other situations, methods known to them will be used. Students will be expected to analyze data sets using some widely available software such as R or EXCEL or MYSTAT. They can also search on the internet and look for freely downloadable software. Checking the software will then be one more exercise.

In this book many new terms will be introduced and students will be expected to read about those terms on the internet. Many websites will be mentioned and students will have to visit the websites and find out information available. Such work should be shared. In other words, different students (groups) should study different websites and report about them in the class (instead of each student visiting the same website for familiarity. If a program on a website is to be used, then of course all the students will have to visit it.

It is hoped that teachers will be able to develop liaison with a public hospital nearby and take up some assignment involving analysis of data gathered there. Such a project will be of immense educational value and participants should get academic credit.

Lastly we touch upon the issue of software. It is very useful for students to be familiar with statistical software packages in use. Many colleges may find it beyond their means to buy and provide enough copies of commercial soft wares like MINITAB, SPSS and SAS. Hence it is good to focus on software packages that are free or inexpensive. Students should learn use of EXCEL that is widely available. Another valuable package that is free is R. For epidemiology, EpiInfo is a package that is freely downloadable and

is used widely. Again, familiarity with this package will make students capable of helping a wide range of users of statistics in the medical and public health field. Teachers should encourage students to search freely downloadable packages from internet and then to test them. This can be a very useful project.

It is our pleasure to bring the second edition of the book for a wonderful year 2013. This is a very special and meaningful year for all the users of statistics because of the 2013 International Year of Statistics (http://www.statistics2013.org). We dedicate this book to those who are participating enthusiastically in promoting statistics to improve and enrich human life.

Anil Gore
Sharayu Paranjpe
Madhav Kulkarni

7 Mar 2013

***

Addendum: One input from some teachers who used this book is that in the part II, there is some history which poses a challenge in terms of examination. Teachers are puzzled as to what kind of questions can be asked. We have three responses. One is that examination is just one part of the whole learning experience. Knowledge should be the priority. We should not let convenience of examination determine what students should learn. Second response is that there is enough material on contingency tables on which traditional questions can be set. After all this part is only 25% of the whole course. Last response is that if the need is critical, teachers can perhaps include some simple models (differential equations) in epidemiology in the course. Plenty of material is available on such models in text books and internet.

April 4, 2015

Table of Contents

Chapter 3: Statistical Aspects of Clinical Trials

Chapter 1

Population Study

This material constitutes 25% of the course and should be covered in 3 weeks. This part is further divided into four sections. Section 1 is on models of population growth. Section 2 is on life tables and related data and Section 3 is on Survival curves. The last section describes population and changes in population size in India.

1.1 Models of Population Growth.

In nature, populations (of animals or plants or humans) are never static. They change. Sometimes they grow and at other times they decline. By and large, in recent history, populations of humans have tended to increase. There are occasional dips, caused by epidemics or war. But a dominant trend is that of growth. In this section our aim is to discuss some mathematical models commonly used for describing growth of a human population. In fact the same models also work for populations of other animals (and plants as well). Specifically there are two aspects of interest in these models. The first is rate of growth and the second is attainment of a steady state (equilibrium). Equilibrium is a population size at which growth/change rate becomes zero. As is inevitable in any modeling exercise, we shall keep the framework very simple (and somewhat unrealistic). We ignore/neglect effects of immigration and emigration. We shall thus be dealing with a geographically **'closed'** population. We also ignore ages of individuals, at least to begin with. If needed, these restrictions can be relaxed thus making the model more realistic. This realism comes at the expense of extra mathematical complexity. Population size is generally denoted by N. Since population size changes with time, the dependence on time is shown explicitly by writing N as a function of t, N(t). Aim of the model is to express change in N(t) over time.

1.1.1 Linear Growth

Growth in the size of a population is often modeled using a differential equation

$$\frac{dN(t)}{dt} = f(N(t)),$$

where N(t) is the population size at time t and f is any suitable function. The simplest function is of course a constant. In that case the equation becomes

$$\frac{dN(t)}{dt} = c.$$

It has the solution $N(t) = ct + d$, where d is a constant of integration to be determined by initial condition. Let the population size corresponding to $t = 0$ be $N(0)$. Then, under the above model we get, $N(t) = N(0) + ct$.

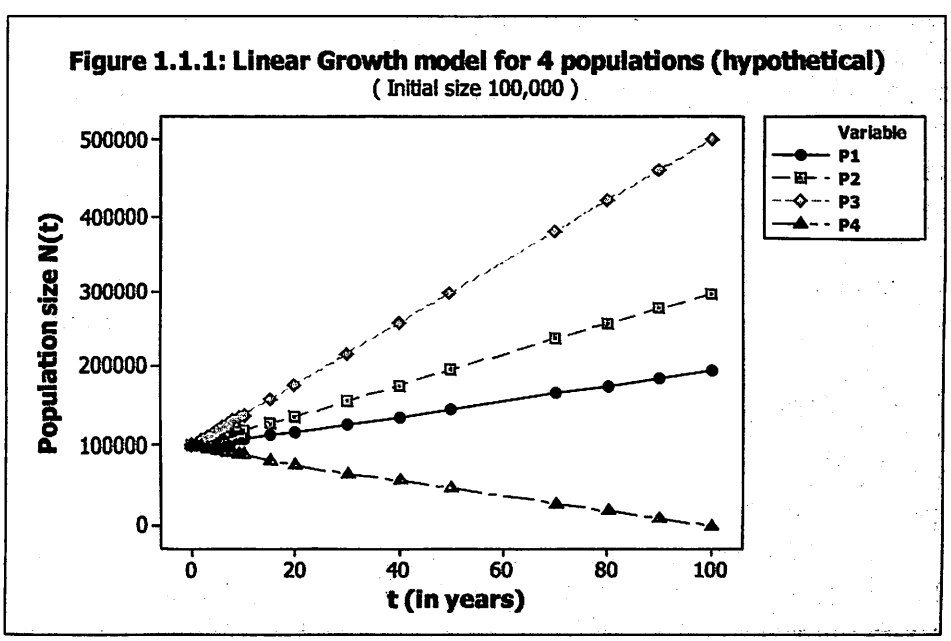
This is the model for linear growth. If c is negative, the population progressively declines to zero. If c is positive, N(t) increases without any upper limit. Also, the increment per unit time is constant. These are the features of this model.

Study the data in Table 1.1.1 below and also the graph (see Figure 1.1.1) on population size. Can you guess the growth rate in these hypothetical populations?

Table 1.1.1: Population sizes with linear growth (hypothetical)

t (years)	P1	P2	P3	P4
0	100000	100000	100000	100000
1	101000	102000	104000	99000
2	102000	104000	108000	98000
3	103000	106000	112000	97000
4	104000	108000	116000	96000
5	105000	110000	120000	95000
6	106000	112000	124000	94000
7	107000	114000	128000	93000
8	108000	116000	132000	92000
9	109000	118000	136000	91000
10	110000	120000	140000	90000
15	115000	130000	160000	85000
20	120000	140000	180000	80000
30	130000	160000	220000	70000
40	140000	180000	260000	60000
50	150000	200000	300000	50000
70	170000	240000	380000	30000
80	180000	260000	420000	20000
90	190000	280000	460000	10000
100	200000	300000	500000	0

(starting population size 100,000).



These features may seem to be rather unrealistic. Usually a larger population has a larger number of individuals in the reproductive phase and hence the number of births is

also larger. Thus the number of individuals in the population is expected to influence its rate of growth. This can be incorporated into the model by assuming that (instead of growth rate) per capita growth rate (growth rate divided by the size of the population) is constant, i. e. by choosing $f(N(t)) = r^*N(t)$, so that the growth equation becomes

$$\frac{dN(t)}{dt} = r^*N(t).$$

This leads to exponential growth.

1.1.2 Exponential Growth

The constant r in the above equation, interpreted as per capita instantaneous growth rate, is a parameter of considerable interest. This equation is easily solved by separating variables as

$$\frac{dN(t)}{N(t)} = r\,dt,$$

and integrating, to get $\ln(N(t)) = rt + d$. Using initial condition this can be written as

$$N(t) = N(o)\, e^{rt}.$$

This is known as the model for exponential growth. Per capita instantaneous growth rate r is sometimes called 'intrinsic rate of increase'. It is also sometimes called 'Malthusian parameter' after the British scholar Thomas Robert Malthus (1766-1834). Malthus argued in his 'Essay on the Principles of Population' written in 1798, that in the absence of any constraints, human population will grow in a multiplicative manner and will eventually outstrip available food. In his pessimistic scenario, this would cause famine and mass deaths unless natural disasters other than food shortages decimate the numbers.

Charles Darwin was inspired by the Malthusian concept of exponential growth. He calculated that a single pair of elephants would have at least 15 million descendants after five centuries if indeed elephant population grew exponentially. Such calculations made amply clear the inevitability of competition among individuals of the same species for limited resources. This competition is the keystone of Darwin's theory of evolution through natural selection.

Notice that in the equation for exponential growth, if r is positive, the population increases exponentially beyond limit. Take the case of human populations in many countries today. Almost nowhere does one encounter the so called 'zero population growth' which is equivalent to r = 0. Some countries have a rather low growth rate of say 1% per year, others experience a moderate growth rate of say 2% per year while high growth rates of 3 or 4 % are not rare. A growth rate of 2% per year can be represented by the equation $N(t+1) = (1.02)\, N(t)$.

Here time is measured in years. The coefficient of N(t) in such equations is often denoted by R (here it is 1.02). Malthusian parameter r is nothing but $ln(R)$, the natural logarithm of R.

Population explosion implicit in exponential growth is sometimes difficult to grasp in an intuitive way. If so, consider the following legend.

A King who was pleased with the skill exhibited by a chess player offered to give him a reward. Thoughtfully the player asked for some grains of wheat to put on the chessboard. King granted him his wish. The player said that he wanted one grain on the first square, two grains on the second square, four on the third and so on, doubling the number every time till the 64th square on the chess board was included. It seemed like a trivial demand. But the Chief of Royal Treasury was alarmed. He advised the King that it was well beyond the kingdom's resources to satisfy the chess player. Can you estimate how much wheat (in thousands of tons!) is needed to fulfill the request?

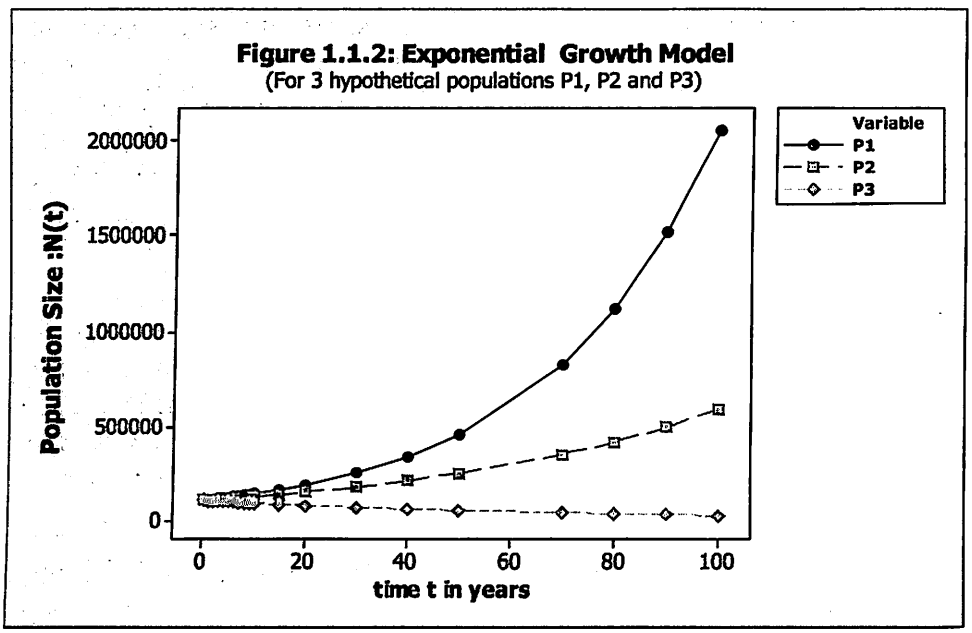
Example: Study the following data on changing population sizes of three different hypothetical populations P1, P2, P3. In each case initial population size N(0) = 100,000. When plotted (see Figure 1.1.2), We see the exponential nature of the change in population. Verify that population 1 and 2 are increasing exponentially whereas population 3 is decreasing exponentially. Also find crude estimate of population growth rate.

Table 1.1.2: Population sizes with exponential growth (hypothetical)

t (years)	P1	P2	P3
0	100000	100000	100000
1	103066	101786	98511
2	106226	103603	97045
3	109483	105454	95600
4	112840	107337	94176
5	116300	109253	92774
6	119865	111204	91393
7	123541	113190	90032
8	127328	115212	88692
9	131232	117269	87372
10	135256	119363	86071
15	157302	130408	79852
20	182942	142476	74082
30	247441	170063	63763
40	334678	202993	54881
50	452673	242298	47237
70	828130	345216	34994
80	1120097	412061	30119
90	1514999	491848	25924
100	2049129	587085	22313

Initial population: 100,000

Is exponential growth a common feature of populations? After all any population growing exponentially must, sooner or later, encounter shortages of resources since earth is finite. Then is the model of much use? Was Malthus at all right?



The answer is that there are phases in the growth of a population, during which resources are more than adequate relative to the size of the population at that time. The exponential model can usefully describe growth in such situations.

1.1.3 Sigmoidal Growth

All observers of nature agree that populations do not grow without limit. Question is what limits them? It may be shortage of essential resources or internal competition. Whatever the cause, populations instead of increasing indefinitely, often tend to reach a plateau. This frequently observed pattern can be modeled using a variety of equations. The most popular among those is the Logistic Equation.

How can the equation for exponential growth be modified so that it generates a plateau? Instead of assuming that the per capita instantaneous growth rate r is constant, we let it depend on the population size. Thus

$$\frac{1}{N(t)}\frac{dN(t)}{dt} = f(N(t))$$

The simplest form for f (N(t)) is linear. Remember that the per capita growth rate has to decrease as population size increases. Hence we have

$$\frac{1}{N(t)}\frac{dN(t)}{dt} = r - cN(t)$$

This can be rewritten as

$$\frac{dN(t)}{N(t)} = r\,dt - c\,N(t)dt.$$

The solution is $\ln(N(t)) = rt - c(N(t))^2 + d$. When N(t) is close to zero, we can ignore the term $(N(t))^2$. Therefore the equation becomes $\ln(N(t)) = r.t + d$, which is an equation of exponential model. However, as N(t) increases, the per capita growth rate falls and reaches zero at N(t) = r/c. This constant (r/c) is often called the 'carrying capacity' of the environment and is denoted by K. Substituting back c= r/K we get

$$\frac{dN(t)}{dt} = r\,(K - N(t))\,/\,K.$$

To solve this differential equation we separate the variables. Thus

$$\frac{K\,dN(t)}{N(t)(K - N(t))} = rdt$$

This can be written using partial fractions as

$$\left[\frac{1}{N(T)} + \frac{1}{K - N(t)}\right] dN(t) = rdt$$

This, on integrating, gives $\ln\left(\frac{N(t)}{K - N(t)}\right) = r.t + d$. Here the constant of integration d is clearly equal to

$$\ln\left(\frac{N(0)}{K - N(0)}\right)$$

Hence we get $\frac{N(t)}{K - N(t)} = e^{r.t}\,\frac{N(0)}{K - N(o)}$, this implies

$$N(t) + \frac{e^{rt}}{q} N(t) = K\frac{e^{rt}}{q},$$

which on simplification gives

$$N(t)\,(1 + \frac{e^{rt}}{q}) = K\frac{e^{rt}}{q}\ .$$

It can be rewritten as

$$N(t) = \frac{K}{1 + qe^{-rt}}$$

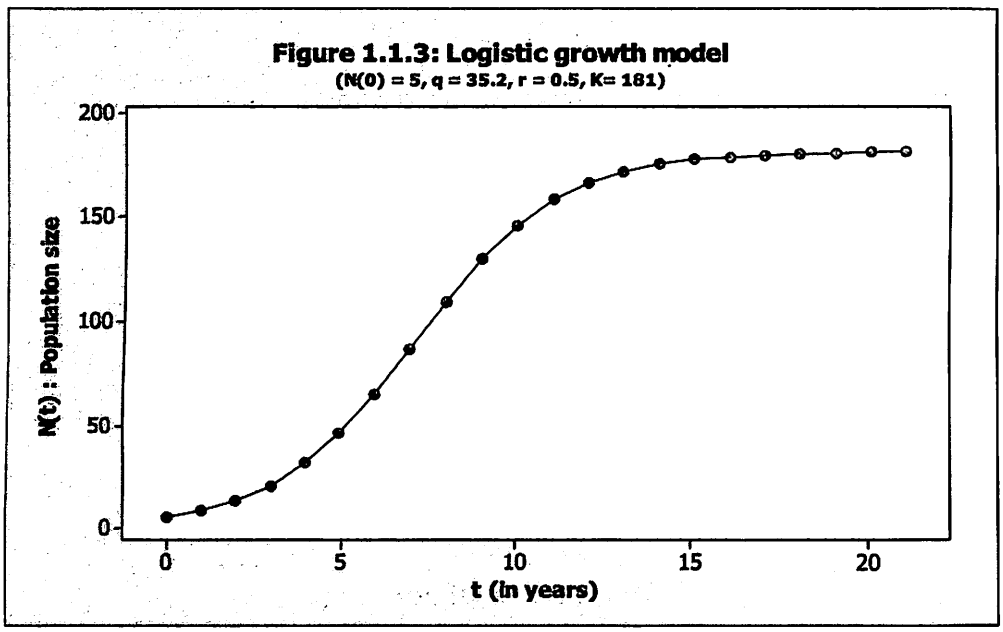
where $q = (K - N(0))\,/N(0)$.

This is the equation of logistic growth. Note that as $t \to \infty$, $N(t) \to K$. The population does not grow without limit. The curve represented by this equation is called sigmoidal because it is shaped like the letter S stretched at both ends.

Table 1.1.3: Growth of Eucalyptus Plantations

T	N(t)	t	N(t)	t	N(t)	t	N(t)
0	5.000	6	65.758	12	166.475	18	180.217
1	8.098	7	87.739	13	171.903	19	180.524
2	12.976	8	110.050	14	175.371	20	180.711
3	20.442	9	130.119	15	177.543	21	180.825
4	31.403	10	146.301	16	178.888		
5	46.537	11	158.237	17	179.713		

We reproduce the data on the growth of eucalyptus plantations over the years (see Table 1.1.3). We have fitted logistic growth model to these data, the values of N(0) = 5, q = 35.2, r = 0.5 and carrying capacity K = 181, estimated graphically. The logistic growth model is an appropriate fit for the above data. See figures 1.1.3



Even in cases where exponential growth is observed for a while, intuitively it is expected that populations should eventually experience a decreased growth rate. Darwin's elephants in fact do not grow to millions. Today in Indian Forests, even in favorable areas, there are just a couple of elephants per sq. km.

Returning to equation that we started with, namely $\frac{1}{N(t)}\frac{dN(t)}{dt} = r - cN(t)$. Note that growth rate is zero at N(t) =0 and also at N(t) = K. These values are therefore called equilibria. Of these N(t) = 0 is unstable because, once population size becomes positive, in this set up, it never returns to zero. By contrast K is called a stable equilibrium. Further, because this stable equilibrium is attained only in the limit as $t \to \infty$, it is called an 'asymptotically stable equilibrium'.

1.1.4 Exercises

E1.1.1 (a) Draw graphs for linear, exponential and logistic growth superimposed together.

(b) For the logistic growth model obtain the equilibria and discuss their stability.

(c) Check what happens to the sign of the growth rate when population deviates from equilibrium. For stability, growth rate should become negative as soon as population increases beyond the equilibrium level and should become positive as population falls below it.

(b) Verify the following:

(i) If $N(t) < K$, $\frac{dN(t)}{dt}$ is positive and increasing initially. It reaches a maximum at $N(t) = K/2$.

(ii) If $N(t) > K$, $\frac{dN(t)}{dt}$ is negative and N(t) decreases to K.

(iii) $\frac{dN(t)}{dt}$ is symmetric around K/2.

E 1.1.2. The table below gives data on growth of Indian population in the last century.

Table 1.1.4: Variation in India's Population Size Since 1901

Year	Total	Rural	Urban
1901	238,396,327	212,544,454	25,851,873
1911	252,093,390	226,151,757	25,941,633
1921	251,321,213	223,235,043	28,086,170
1931	278,977,238	245,521,249	33,455,989
1941	318,660,580	274,507,283	44,153,297
1951	361,088,090	298,644,381	62,443,709
1961	439,234,771	360,298,168	78,936,603
1971	548,159,652	439,045,675	109,113,977
1981	683,329,097	523,866,550	159,462,547
1991	846,302,688	628,691,676	217,611,012
2001	1,028,737,436	742,490,639	286,119,689

For each column (total, rural and urban) fit various models of growth and check which gives a better fit.

E1.1.3 Here is a set of crude estimates of human population size (in billion) at various points of time in history. Assuming the world population grew exponentially throughout this period at an approximately constant rate, estimate this rate.

Table 1.1.5: World Human Population (in billions)

t	1650	1700	1800	1920	1950	1975	1985	2000	2010	2020
N(t)	0.5	0.8	0.9	2.0	3.0	4.0	5.0	6.0	7.0	8.0

E1.1.4 For the table 1.1.5 check analytically whether the following is true:
A population's approximate doubling time is found by dividing 70 by its annual percentage population growth. For example, at a 4 percent growth rate, a country's population will double in about 18 years; at a 1 percent growth rate, it will take about 70 years.

E1.1.5 Table below gives population growth rates for many countries in the world. (a) Locate some of the countries with high growth rates on the world map. (b) Locate some of the countries with negative growth rates on the world map. Do you see any patterns? Have you heard about 'north' and 'south' groups of nations?

Table 1.1.6 Population Growth Rates by Country (2009-2013)								
Rank	Country	Gr	Rank	Country	Gr	Rank	Country	Gr
1	Oman	9.1	43	Luxembourg	2.4	93	Haiti	1.4
2	Qatar	7.1	49	Cote d'Ivoire	2.3	93	Kazakhstan	1.4
3	South Sudan	4.3	49	Yemen, Rep.	2.3	93	Morocco	1.4
4	Kuwait	3.9	51	Ghana	2.2	98	Azerbaijan	1.3
5	Niger	3.8	51	Jordan	2.2	98	Colombia	1.3
6	Uganda	3.4	51	Papua New Guinea	2.2	98	Dominican Republic	1.3
7	Eritrea	3.3	51	Turks and Caicos Islands	2.2	98	India	1.3
8	Burundi	3.2	51	Vanuatu	2.2	98	Iran, Islamic Rep.	1.3
8	Gambia, The	3.2	56	Solomon Islands	2.1	98	Norway	1.3
8	Zambia	3.2	56	Sudan	2.1	98	Peru	1.3
11	Angola	3.1	58	Central African Republic	2.0	98	South Africa	1.3
11	United Arab Emirates	3.1	58	Honduras	2.0	98	Turkey	1.3
13	Chad	3.0	58	Syrian Arab Republic	2.0	98	Turkmenistan	1.3
13	Mali	3.0	61	Algeria	1.9	108	Bangladesh	1.2
13	Tanzania	3.0	61	Bahrain	1.9	108	Canada	1.2
13	West Bank and Gaza	3.0	61	Lao PDR	1.9	108	Guam	1.2
17	Burkina Faso	2.9	61	Macao SAR, China	1.9	108	Hong Kong SAR, China	1.2
17	Malawi	2.9	61	Maldives	1.9	108	Indonesia	1.2
17	Senegal	2.9	61	Namibia	1.9	108	Mexico	1.2
17	Somalia	2.9	61	Saudi Arabia	1.9	108	Nepal	1.2
17	Timor-Leste	2.9	61	Sierra Leone	1.9	108	St. Kitts and Nevis	1.2
22	Equatorial Guinea	2.8	69	Cambodia	1.8	116	Cyprus	1.1
22	Madagascar	2.8	69	Israel	1.8	116	French Polynesia	1.1
22	Nigeria	2.8	71	Australia	1.7	116	Lesotho	1.1
22	Rwanda	2.8	71	Bhutan	1.7	116	St. Martin (French part)	1.1
26	Benin	2.7	71	Bolivia	1.7	116	Switzerland	1.1
26	Congo, Dem. Rep.	2.7	71	Cayman Islands	1.7	116	Vietnam	1.1
26	Kenya	2.7	71	Egypt, Arab Rep.	1.7	122	Antigua and Barbuda	1.0
26	Liberia	2.7	71	Kyrgyz Republic	1.7	122	Curacao	1.0
26	Sao Tome and Principe	2.7	71	Malaysia	1.7	122	Lebanon	1.0
26	Zimbabwe	2.7	71	Pakistan	1.7	122	Seychelles	1.0
32	Congo, Rep.	2.6	71	Paraguay	1.7	122	Tunisia	1.0
32	Ethiopia	2.6	71	Philippines	1.7	127	Argentina	0.9
32	Guinea	2.6	81	Ecuador	1.6	127	Botswana	0.9
32	Togo	2.6	81	New Caledonia	1.6	127	Brazil	0.9
36	Cameroon	2.5	81	Panama	1.6	127	Chile	0.9
36	Guatemala	2.5	81	Sint Maarten (Dutch part)	1.6	127	Kosovo	0.9
36	Iraq	2.5	85	Bahamas, The	1.5	127	St. Lucia	0.9
36	Mauritania	2.5	85	Djibouti	1.5	127	Suriname	0.9
36	Mozambique	2.5	85	Kiribati	1.5	134	Cabo Verde	0.8
36	Singapore	2.5	85	Mongolia	1.5	134	Fiji	0.8
36	Tajikistan	2.5	85	Nicaragua	1.5	134	Libya	0.8
43	Afghanistan	2.4	85	Swaziland	1.5	134	Malta	0.8
43	Belize	2.4	85	Uzbekistan	1.5	134	Monaco	0.8
43	Comoros	2.4	85	Venezuela, RB	1.5	134	Myanmar	0.8
43	Gabon	2.4	93	Brunei Darussalam	1.4	134	Samoa	0.8
43	Guinea-Bissau	2.4	93	Costa Rica	1.4	141	Belgium	0.7

Table 1.1.6 Population Growth Rates by Country (2009-2013) cont.								
Rank	Country	Gr	Rank	Country	Gr	Rank	Country	Gr
141	El Salvador	0.7	160	Netherlands	0.4	186	Spain	0.0
141	Isle of Man	0.7	160	Russian Federation	0.4	186	Sri Lanka	
141	Liechtenstein	0.7	160	Tonga	0.4	186	St. Vincent and the	0.0
141	Palau	0.7	169	Albania	0.3	192	Belarus	-0.1
141	Sweden	0.7	169	Austria	0.3	192	Bosnia and Herzegovina	-0.1
141	United States	0.7	169	Jamaica	0.3	192	Faeroe Islands	-0.1
148	Andorra	0.6	169	Thailand	0.3	192	Greenland	-0.1
148	Guyana	0.6	169	Trinidad and Tobago	0.3	196	Japan	-0.2
148	New Zealand	0.6	169	Uruguay	0.3	196	Ukraine	-0.2
148	San Marino	0.6	175	Armenia	0.2	198	American Samoa	-0.3
148	United Kingdom	0.6	175	Georgia	0.2	198	Croatia	-0.3
153	Barbados	0.5	175	Ireland	0.2	198	Greece	-0.3
153	China	0.5	175	Slovak Republic	0.2	201	Estonia	-0.4
153	Finland	0.5	175	Slovenia	0.2	201	Portugal	-0.4
153	France	0.5	175	Tuvalu	0.2	201	Romania	-0.4
153	Iceland	0.5	181	Czech Republic	0.1	204	Hungary	-0.5
153	Korea, Dem.	0.5	181	Macedonia, FYR	0.1	204	Serbia	-0.5
153	Korea, Rep.	0.5	181	Marshall Islands	0.1	204	Virgin Islands (U.S.)	-0.5
160	Aruba	0.4	181	Montenegro	0.1	207	Bulgaria	-0.6
160	Bermuda	0.4	181	Northern Mariana	0.1	208	Puerto Rico	-0.7
160	Denmark	0.4	186	Cuba	0.0	209	Latvia	-1.2
160	Dominica	0.4	186	Micronesia, Fed. Sts.	0.0	210	Lithuania	-1.3
160	Grenada	0.4	186	Moldova	0.0	211	Germany	-1.7
160	Mauritius	0.4	186	Poland	0.0	212	Italy	-2.0

(c) One guess is that high growth rates go together with high birth and death rates and low per capita incomes. Check whether it is true. (Note that you will need to search for some more data from internet before you can complete this exercise.)

E1.1.6 Carry out the exercise about estimating the quantity of wheat needed to satisfy the demand of the chess player.

E1.1.7 Fit an appropriate growth curve to the following data:

Table 1.1.7: Forecasting Population Growth			
Year	Population (in millions)	Year	Population (in millions)
1910	10.6	1980	80.0
1920	14.5	1990	96.0
1930	20.5	2000	102.5
1940	25.8	2005	116.0
1950	35.0	2010	121.5
1960	49.0	2020	Can you estimate it?
1970	65.0	2050	Can you estimate it?

Estimate the population size in year 2020 and in year 2050.

E1.1.8 Following are the data on population growth of a certain species observed in different locations. It is a common experience that the increase in the abundance of a species in a large habitat is also associated with the increase in the locations at which the species is found. So it is also of interest to model the growth in the number of locations where the individuals of the species are sighted. This may be looked as a diffusion process. Can you model (a) the number of individuals and (b) the number of locations by an exponential model assuming constant rate of growth? Estimate this rate.

Table 1.1.8: Growth of a Species in Numbers and Locations

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Individuals	3	9	21	75	140	300	559	1200	3600	8000
Locations	1	2	4	12	25	40	85	95	130	160

1.2 Life Tables

The population models studied above take a very broad view. All details about birth, death etc are ignored. When the interest is in these more detailed parameters, we have to use different models. We will do this in two ways. One way is to prepare a table of various events called a life table. The other way is to build special models to describe these phenomena. We will take it up in the next section.

The term 'life table' arises from the fact that data on survivorship, age composition etc. are given in a tabular form. These tables constitute the basis on which all calculations of life insurance are based. There are two kinds of life tables commonly used. In both, variable x is used to denote age. Body of the table contains information about the number of people in various age groups and also about number of deaths in those age groups.

One major use of life tables is for designing insurance policies by taking into account probability of death at various ages.

1.2.1 Cohort life table

Here we have a cohort or group of individuals all born at the same time, say $x = 0$. The first column of a cohort life table has age classes usually with equal intervals. The second column l_x generally gives the number of individuals alive at age x (instead of proportions) out of say 1000. The third column is the number of deaths in the interval (x, x+1], denoted by d_x where $d_x = l_x - l_{(x+1)}$. Here lx is of course a non-increasing sequence. Fourth column is the proportion of deaths, $q_x = \frac{d_x}{l_x}$ in the interval. Fifth column is L_x, the number of individuals multiplied by time units lived in the interval (this could be called person years lived). Theoretically with unit intervals, $L_x = \int_x^{x+1} l(u)du$, but it is approximated by $(l_x + l_{\{x+1\}})/2$. This could be man years or insect days or whatever. Sixth column is T_x, total life remaining to all individuals alive at x, given by $T_x = \sum_{j=x}^{w} Lj$, where w

is the last age class beyond which everyone is dead. Naturally the average remaining life for an individual alive at x, e_x is given by $e_x = T_x / l_x$ which is the last column. Following is an illustrative life table.

Table 1.2.1: Cohort Life Table

X	l_x	d_x	$q_x=d_x/l_x$	L_x	T_x	e_x	X	l_x	d_x	$q_x=d_x/l_x$	L_x	T_x	e_x
0	100000	2629	0.02629	98685.5	6549130	65.5	39	92576	261	0.00282	92445.5	2821486	30.5
1	97371	141	0.00145	97300.5	6450445	66.2	40	92315	295	0.0032	92167.5	2729041	29.6
2	97230	107	0.0011	97176.5	6353144	65.3	41	92020	332	0.00361	91854	2636873	28.7
3	97123	63	0.00065	97091.5	6255968	64.4	42	91688	408	0.00445	91484	2545019	27.8
4	97060	63	0.00065	97028.5	6158876	63.5	43	91280	414	0.00454	91073	2453535	26.9
5	96997	69	0.00071	96962.5	6061848	62.5	44	90866	464	0.00511	90634	2362462	26
6	96928	69	0.00071	96893.5	5964885	61.5	45	90402	532	0.00589	90136	2271828	25.1
7	96859	52	0.00054	96833	5867992	60.6	46	89870	587	0.00653	89576.5	2181692	24.3
8	96807	54	0.00056	96780	5771159	59.6	47	89283	680	0.00762	88943	2092116	23.4
9	96753	51	0.00053	96727.5	5674379	58.6	48	88603	702	0.00792	88252	2003173	22.6
10	96702	33	0.00034	96685.5	5577651	57.7	49	87901	782	0.0089	87510	1914921	21.8
11	96669	40	0.00041	96649	5480966	56.7	50	87119	841	0.00965	86698.5	1827411	21
12	96629	47	0.00049	96605.5	5384317	55.7	51	86278	885	0.01026	85835.5	1740712	20.2
13	96582	61	0.00063	96551.5	5287711	54.7	52	85393	974	0.01141	84906	1654877	19.4
14	96521	86	0.00089	96478	5191160	53.8	53	84419	1082	0.01282	83878	1569971	18.6
15	96435	105	0.00109	96382.5	5094682	52.8	54	83337	1088	0.01306	82793	1486093	17.8
16	96330	83	0.00086	96288.5	4998299	51.9	55	82249	1213	0.01475	81642.5	1403300	17.1
17	96247	125	0.0013	96184.5	4902011	50.9	56	81036	1344	0.01659	80364	1321657	16.3
18	96122	133	0.00138	96055.5	4805826	50	57	79692	1423	0.01786	78980.5	1241293	15.6
19	95989	149	0.00155	95914.5	4709771	49.1	58	78269	1476	0.01886	77531	1162313	14.9
20	95840	154	0.00161	95763	4613856	48.1	59	76793	1572	0.02047	76007	1084782	14.1
21	95686	138	0.00144	95617	4518093	47.2	60	75221	1696	0.02255	74373	1008775	13.4
22	95548	163	0.00171	95466.5	4422476	46.3	61	73525	1784	0.02426	72633	934402	12.7
23	95385	168	0.00176	95301	4327010	45.4	62	71741	1933	0.02694	70774.5	861769	12
24	95217	166	0.00174	95134	4231709	44.4	63	69808	2022	0.02897	68797	790994	11.3
25	95051	151	0.00159	94975.5	4136575	43.5	64	67786	2186	0.03225	66693	722197	10.7
26	94900	149	0.00157	94825.5	4041599	42.6	65	65600	2261	0.03447	64469.5	655504	10
27	94751	166	0.00175	94668	3946774	41.7	66	63339	2371	0.03743	62153.5	591035	9.3
28	94585	157	0.00166	94506.5	3852106	40.7	67	60968	2426	0.03979	59755	528881	8.7
29	94428	133	0.00141	94361.5	3757599	39.8	68	58542	2356	0.04025	57364	469126	8
30	94295	160	0.0017	94215	3663238	38.8	69	56186	2702	0.04809	54835	411762	7.3
31	94135	149	0.00158	94060.5	3569023	37.9	70	53484	2548	0.04764	52210	356927	6.7
32	93986	152	0.00162	93910	3474962	37	71	50936	2677	0.05256	49597.5	304717	6
33	93834	160	0.00171	93754	3381052	36	72	48259	2811	0.05825	46853.5	255120	5.3
34	93674	199	0.00212	93574.5	3287298	35.1	73	45448	2763	0.0608	44066.5	208266	4.6
35	93475	187	0.002	93381.5	3193724	34.2	74	42685	2710	0.06349	41330	164200	3.8
36	93288	212	0.00227	93182	3100342	33.2	75	39975	2848	0.07125	38551	122870	3.1
37	93076	228	0.00245	92962	3007160	32.3	76	37127	2832	0.07628	35711	84319	2.3
38	92848	272	0.00293	92712	2914198	31.4	77	34295	2835	0.08267	32877.5	48608	1.4
39	92576	261	0.00282	92445.5	2821486	30.5	78	31460	2803	0.0891	15730	15730	0.5

X: Age class lower limit, l_x: number alive at age X, d_x: number of deaths in the interval, q_x: proportion of deaths in the interval, L_x: number of units lived, T_x: Total residual life, e_x: average residual life.

1. In this cohort, the numbers of individuals studied are 100,000. All these individuals were alive at age 0. Here $l_0 = 100000$
2. Of theses 100,000 individuals, $d_0 = 2629$ individuals died before reaching age 1. This gives $q_0 = \mathbf{0.026290}$, proportion of deaths in the interval (0,1].
3. $L_0 = \mathbf{98685.5}$ are the number of individuals survived in the age group (0,1].
4. T_0, the total residual life of these l_0 individuals is **6549130** and therefore average residual life of these l_0 individuals is **65.6** years.

Similarly other entries in the table below should be read. Consider row corresponding to age 71. Of the initial 100,000, individuals, 50936 reached the age 71 years. Hence $l_{71} =$ 50,936

2. Of theses 50,936 individuals, $d_{71} = 2677$ individuals died before reaching age 72. This gives $q_{71} = \mathbf{0.052556}$, proportion of deaths in the interval (0,1].
3. $L_{71} = \mathbf{49597.5}$ are the number of individuals survived in the age group (71, 72].
4. T_{71}, the total residual life of these l_{71} individuals is **304717** and therefore average residual life of these l_{71} individuals is **6.0** years.

Thus the life table is a document that discloses the life history of a group of individuals all born at the same point of time. It shows that how many of these individuals survived to celebrate each birth day, the probability of surviving from one age to another, and so on. The life table is a summary of the mortality experience of any population group during a specified time period.

1.2.2 Current life table:

Data required to construct a cohort life table are somewhat difficult to get especially for organisms with long lifespan. Humans do have a long life. Also, over recent decades and centuries, it is getting longer. In 1815 Joshua Milne suggested an intuitive method of estimating parameters of interest from current data on a given population. This approach leads to the so called current life table. This is really cross sectional data on the age composition of a population at a point of time and numbers of deaths occurring among them in one unit of time. (In contrast, cohort life table has data over time or `longitudinal data'). Current life tables are most common for human populations in which censuses are conducted rather thoroughly. Current life table is equivalent to cohort life table if population is stationary.

The current life table begins with a column for age interval, followed by number of individuals in that interval say Y(x). The third column is number of deaths among individuals in the age group, D(x). The fourth column is age specific death rate $M(x) = D(x)/Y(x) = [l_x - l_{x+1}]/ Y(x) = 2[l_x - l_{x+1}]/ [l_x + l_{x+1}]$

What is a stable population? And what is a stationary population? What is the significance of these concepts?

A population where fertility rate and mortality rate are constant is called a stable population. This type of population will show an unvarying age distribution and will grow at a constant rate. Where fertility and mortality are equal, the stable population is

stationary. This population remains constant both in terms of numbers and proportions. Needs of a population are decided by these two aspects. If proportions in different age groups change, requirements change. If the proportion of older people increases (graying of the population) need for healthcare of elderly becomes very important. If total count increases but proportions in different age groups remain constant, then planners have to provide for more and more resources but the share of different age groups can be kept constant. All such considerations make these features relevant and interesting.

What are fertility and mortality rates?

So the way we really calculate fertility is like this: For a given year, we count the percentage of women of any given age who had a child that year, total up these percentages for all ages, and divide by 100. For example, suppose -- just to make things simple for an example -- that women only have children when they are between the ages of 18 and 21, and that there are no twins or other circumstances where a woman has more than one child in a year. Then say that 20% of the 18-year-olds had a baby, 40% of the 19-year-olds, 50% of the 20-year-olds, and 30% of the 21-year-olds. Adding these, we get 140%, so the average woman in this hypothetical society has 1.4 children in her lifetime. This is how age specific fertility rates are added to give a fertility rate over the entire reproductive lifespan of a woman.

Age-specific fertility rates are calculated by dividing the number of live births in each age group by the total female population (in thousands) in each age group.

Mortality rate is typically expressed in units of deaths per 1000 individuals per year. Age specific mortality is a rate limited to a particular age group, in which the numerator is the number of deaths in that age group, and the denominator, the number of persons in that age group in the population

If we assume that the population is stationary then the number of deaths in the age group (x, x+1], denoted by D(x), is given by $D(x) = l(x) - l(x+1)$. This is trivial. D(x) is simply the difference between number of persons alive upto age x and those alive upto age (x + 1). Therefore age specific mortality rate M(x) the ratio of number of deaths among the age group (x, x+1] to the number of individuals in the age group (x, x+1]. This gives

$$M(x) = \frac{D(x)}{Y(x)} = \frac{l(x) - l(x+1)}{[l(x) + l(x+1)]/2}.$$

Problems involving life tables are dealt with in detail in text books of demography such as Impagliazzo (1985) and Keyfitz (1977).

Life tables are the basis of all mathematical demography. They provide a broad picture of the mortality and survivorship pattern in a population. Use of probability models for l(x) has seen considerable development in the field of statistics in recent years.

Note: Western Europe, Russia, China, Canada, Australia, Japan have declining populations. USA, Brazil Argentina Iran Turkey Indonesia, Mongolia, Kazakhstan etc have stable populations. Mexico, South Africa, India, Malaysia, Philippines etc have slowly increasing populations. Much of Africa, Arabia, Pakistan, Central America have rapidly growing populations. See http://www.pregnantpause.org/numbers/fertility.htm#mx

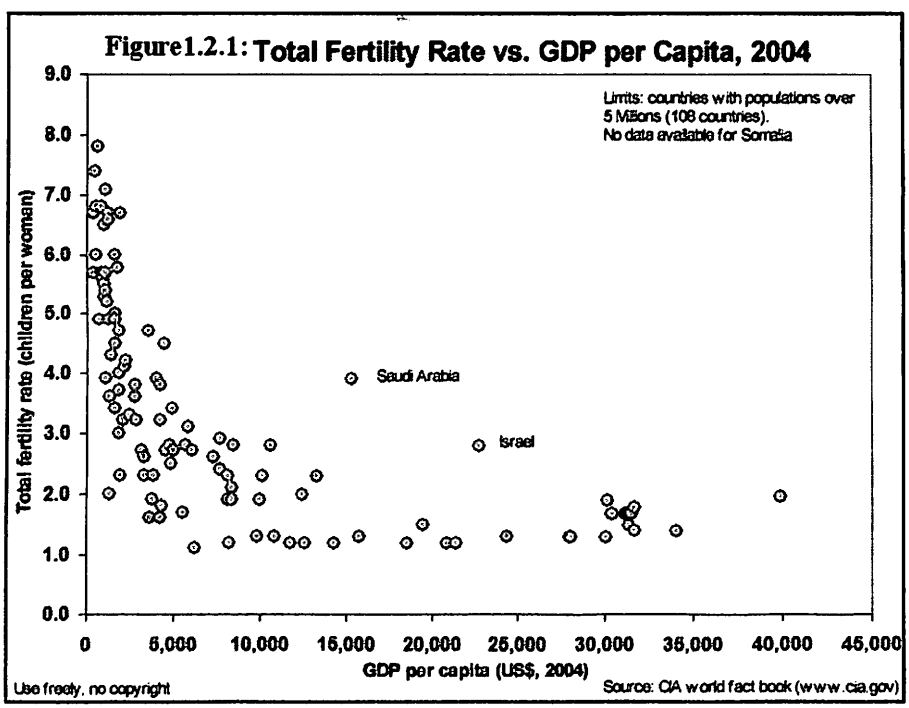
1.2.3 Exercises

E1.2.1 Data for Cohort life tables are not easy to obtain particularly for organisms with long life spans. For example Sharma and Tomar (1964) had to wait patiently for 16 years to study survivorship of bamboo culms. Following table gives their data on survivorship of a cohort of 439 individual culms of Dendrocalamus strictus. All culms were 1 year old at the beginning of observation. Prepare a cohort life table from this data.

Table 1.2.2: Life Table for Dendrocalamus Strictus

Age (Years)	# Survivors	Age (Years)	# Survivors
1	439	9	229
2	438	10	168
3	433	11	106
4	427	12	46
5	410	13	20
6	380	14	9
7	340	15	1
8	286	16	0

E1.2.2 Fertility and mortality rates tend to be high in poorer countries than in wealthy countries. The following graph supports this suggestion.



What equation would you use to fit to this data set? Justify your choice.

E1.2.3 Table below gives India's birth and death rates over nearly a century.

YEAR	Year	BIRTH RATE	DEATH RATE	Year	BIRTH RATE	DEATH RATE
1901-1910	**1905**	49.2	42.6	**1984**	33.9	12.6
1911-1920	**1915**	48.1	47.2	**1985**	32.9	11.8
1921-1930	**1925**	46.4	36.3	**1986**	32.6	11.1
1931-1940	**1935**	45.2	31.2	**1987**	32.2	10.9
1941-1950	**1945**	39.9	27.4	**1988**	31.5	11.0
1951-1960	**1955**	41.7	22.8	**1989**	30.6	10.3
1961-1970	**1965**	41.2	19.0	**1990**	30.2	9.7
1971-1980	**1975**	37.1	14.8	**1991**	29.5	9.8
1981	**1981**	33.9	12.5	**1992**	29.2	10.1
1982	**1982**	33.8	11.9	**1993**	28.5	9.2
1983	**1983**	33.7	11.9			

Fit a suitable regression model and estimate rates for all later years up to 2011 and compare your estimates with actual figures where possible.

1.3 Summarizing Survivorship Data

In this section we describe models for summarizing data in the form of l_x, using probability distributions. The summary is then in terms of one or a few parameters of the probability distribution. These are defined below.

Survival function: Suppose a random variable X denotes the new born's age at death, i.e., X is the life length of an individual. X is a continuous non-negative random variable. Its distribution can be represented in many ways. The distribution function of X is defined as $F(x) = P(X \leq x)$, and $F(x) = 0$ for $x < 0$. We define survival (or survivorship) function as $S(x) = 1 - F(x) = P(X \geq x)$. It is the probability that the individual survives up to age x.

Properties of S(x)

1. $S(0) = 1$
2. Lim $S(x) = 0$ as $x \to \infty$
3. S(x) is non-increasing, i.e. for $x_1 \leq x_2$, $S(x_1) \geq S(x_2)$
4. S(x) is right continuous always, it is continuous for continuous random variable

Interpretation of S(x): S(x) is the probability that an individual is surviving at time x. If there is a population of individuals with identically distributed lifetimes, then S(x) is the expected fraction of the population that has survived at time x.

S(x) is useful for comparing the survival patterns of several populations. If $S_1(x) \geq S_2(x)$ for all x then it can be concluded that the individuals in the population 1 are 'longer living' to those in population 2.

Estimation of F(x) (and S(x)): The simplest estimate of F (x) is the empirical distribution function F_n (x) given by $F_n(x)$ = (# of deaths at or before age x) /n, where n is the size of cohort at x = 0.

Similarly S_n (x) = 1 - F_n (x) (which is nothing but lx) gives a natural estimate of S(x). To consider further condensation we look for suitable parametric models to represent F(x). This is most fruitfully done via hazard function or age specific mortality rate corresponding to F(x). Before studying more details of the hazard function, here is an illustration of a survival model.

The random variable of interest is the time until death for a person aged x. Suppose T(x) denotes the time until death of an individual of age x. This is a residual life of a person of age x. The distribution function and the survival function of T(x) are respectively denoted by ${}_tq_x$ and ${}_tp_x$ (these are the standard International actuarial notations).

Note that ${}_tq_x = P(T(x) \leq t)$, $t \geq 0$ and ${}_tp_{x=1-{}_tq_x} = P(T(x) \geq t)$, for $t > 0$. If t = 1, prefix 1 in ${}_tq_x$ and ${}_tp_x$ may be dropped.

Suppose a survival model is defined by the following values of p_x.

Table 1.3.1; Survival Model

Age (x)	0	1	2	3	4	5	6
p_x	0.95	0.90	0.80	0.50	0.30	0.10	0.00

Obtain corresponding values of S(x) for x = 0, 1, 2, ...6 and 7. Assume l_0 = the number of new born, also called as (radix) = 100,000, find values of l_x and d_x (expected number of deaths between x and x +1) for x = 0, 1, 2, ...6 and 7. Verify that $\sum_x d_x = l_0$.

Solution:

S(0) = 1 always and $p_x = S(x+1)/S(x)$. Hence the recursive relation S(x+1) = p(x)* S(x) can be used to obtain S(x+1). We therefore have S(1) = 0.95, S(2) = 0.855, S(3) = 0.684, S(4) = 0.342, S(5) = 0.1026, S(6) = 0.01026 and S(7) = 0. Further, l_x can be obtained using the relation $l_x = l_0 S(x)$ and d_x is obtained using $d_x = l_x - l_{x+1}$. These results are presented in the table below.

Table 1.3.2: Estimation of Survival

x	p_x	S(x)	l_x	d_x
0	0.95	1.00000	100000	5000
1	0.90	0.95000	95000	9500
2	0.80	0.85500	85500	17100
3	0.50	0.68400	68400	34200
4	0.30	0.34200	34200	23940
5	0.10	0.10260	10260	9234
6	0.00	0.01026	1026	1026
7		0.00000	0	0

Here the limiting age is 7 units (why? This is because S(7) = 0 for the first time). Also note that addition of the last column equals l_0 (= 100000), this must always happen.

Hazard rate: The hazard rate is one of the popular representations for lifetime distributions. Hazard rate is defined as h (x) = f (x) / (1 - F(x)) where f(x) is the derivative of F(x) To interpret h(x), notice that h(x)dx is approximately the conditional probability of death in the age interval (x, x + dx) given that the individual is still alive at age x. It has also been called 'force of mortality' at age x. Note that a probability density function is not a probability. Similarly hazard function is not a probability. Hazard rate r(t) is the ratio of pdf to survival function Note that h (x) in turn determines F (x) completely. To get F (x) given h (x) we see that the integrated form of h (x), called cumulative hazard function H (x) is $H(x) = \int_0^x h(t)dt = \int_0^x \frac{f(t)dt}{(1-F(t))} = \int_0^{F(x)} \frac{du}{1-u} = -\ln(1 - F(x))$ or in other words

$$F(x) = 1 - e^{\{-H(x)\}}$$

This also indicates the constraints to be satisfied by h(x). Since F(x) is non-negative, continuous, non-decreasing and since it approaches 1 as $x \to \infty$, H (x) must be non-negative, non-decreasing and must go to∞ as x d oes so. (see exercise 4) Hence any parametric form chosen for h (x) must be checked on this basis.

Properties of the hazard rate:

1. $h(x) \geq 0$ for all $x \geq 0$
2. $\int_0^\infty h(x)dx = \infty$
3. $S(x) = \exp\{\int_0^x h(x)dx\}$

Illustration: Suppose X follows exp(λ). So $F(x) = 1 - \exp(-\lambda x)$ and $S(x) = \exp(-\lambda x)$ And $h(x) = f(x)/S(x) = \lambda$.

Mean residual life function: Another way of representing survivorship data is through mean residual life function. In the life table it is represented by e(x). Let r (x) be the expected remaining life of an individual alive at age x so that we can write

$$r(x) = \int_x^\infty S(u)du / S(x).$$

r (x) is related to the hazard rate in the following way.

$$\frac{dr(x)}{dx} = r'(x) = \{S(x)(-S(x)) - \int_x^\infty S(u)du \ (-f(x))\}/\{S^2(x)\}$$

$$= -1 + [f(x)/S(x)][\int_x^\infty S(u)du / S(x)]$$

$$= -1 + h(x).r(x)$$

So finally, $\{r'(x) + 1\}/r(x) = h(x)$.

Any parametric form chosen for r (x) must therefore satisfy

(i) $r'(x) \geq -1$,

(ii) $0 \leq r(x) < \infty$,

(iii) $\int_0^{\infty} \frac{dx}{r(x)} = \infty$

(iv) $\{r(x)\}/\{x\ln(x)\} \to 0$ as $x \to \infty$

Let us consider hazard functions for some commonly used distributions in survival analysis.

1.3.1 Exponential Distribution and lack of memory

The simplest form of h (x) is of course a constant function $h(x) = \lambda > 0$. Hence $H(x) = \lambda x$ and $F(x) = 1 - e^{-\lambda x}$. This is the well known exponential distribution, the only distribution with the so called lack of memory property namely,

$$P(X > x + a | X > a) = P(X > x), \quad x, \ a > 0.$$

Let us verify that exponential distribution has this property. Note that the right hand side is $\exp(-\lambda x)$. The left hand side is simply the ratio $P(X > x+a)/P(X > a)$ which in this case is $\exp(-\lambda(x+a))/\exp(-\lambda a)$. Hence the result.

To prove the converse, we see that the given condition namely

$$\{1 - F(x + a)\}/\{1 - F(a)\} = 1 - F(x)$$

can be rewritten as

$$\{F(a) - F(x + a)\}/\{1 - F(a)\} = -F(x).$$

On differentiating with respect to x we get

$$\{f(x + a)\}/\{1 - F(a)\} = f(x).$$

As $x \to 0$ this reduces to

$$\{f(a)\}/\{1 - F(a)\} = f(0),$$

Or alternatively

$$\{d(-\ln(1 - F(a)))\}/\{da\} = f(0).$$

On integrating,

$$\ln(1 - F(a)) = -f(0)\,a + c.$$

But since $F(a) \to 0$ as $a \to 0$, the constant of integration c must be zero. Hence

$$F(a) = 1 - e^{-f(0)a}$$

which is the distribution function of the exponential distribution with $\lambda = f(0)$.

What is the interpretation of this lack of memory property of the exponential distribution? Roughly speaking, the property tells us that chance of living 'a' units of time beyond the current epoch is the same at all ages. Chance that you will live 10 more years is the same whether you are just 5 years old or 95 years old. It means there is no effect of aging on mortality rates. **This is of course not true in humans**. This model therefore is suitable only if the predominant cause of death is unrelated to age. If you consider lifespan of an adult before senescence sets in then for that limited period the model is appropriate. In that portion of life, getting older marginally does not matter. But if you get really old, chance of death keeps climbing.

Here we have assumed that death is possible at any positive age. If there is a lower age limit say θ below which death cannot occur or is not observable then the appropriate form of exponential distribution has the density

$$f(x,\ \theta,\ \lambda) = \lambda e^{-\lambda(x-\theta)} \quad x > \theta > 0,\ \lambda > 0.$$

How plausible is this situation? If we are modeling data on school going children and mortality among them, then everybody is at least 5 years old. This is the constraint of the data set. So, it is clear that $x > 5$.

If we have n independent observations $x_1, x_2, \ldots, x_n$ on lifetimes under such a model, the maximum likelihood estimators of θ and λ are given by

$\hat{\theta}$ = minimum of $(x_1, x_2, \ldots, x_n)$ and $\hat{\lambda} = (\bar{x} - \hat{\theta})^{-1}$

where $\bar{x}$ is the sample mean. If θ is known to be zero $\hat{\lambda}$ reduces to $1/\bar{x}$.

1.3.2 Weibull Distribution and monotone hazard rate

While the constant hazard rate is an attractive property mathematically, it is not always realistic. Sooner or later, old age sets in and every unit of time elapsed makes an individual more prone to death. The hazard rate therefore should increase with age. Weibull distribution which is a generalization of the exponential distribution, has this property. The density function of the Weibull distribution is

$$f(x,\ \alpha,\ c) = \{c/\alpha\}\ (x/\alpha)^{\{c-1\}} \exp[-(x/\alpha)^c],\ x,\ \alpha,\ c > 0.$$

The corresponding distribution function is

$$F(x) = 1 - \exp\{-(x/\alpha)^c\}.$$

Notice that for $c = 1$, the hazard rate becomes a constant as it should be. If $c > 1$, $h(x)$ increases with x. Hence it is useful to model situations in which greater age implies higher proneness to death. Survivorship during adult life of many animals can therefore be modeled using Weibull distribution. If $c < 1$, $h(x)$ is a decreasing function of x. Hence the same distribution can be used to model high mortality which is often observed in infancy.

1.3.3 Bath Tub shaped hazard rate

The hazard rate for the entire life span is not expected to be constant or even monotone. It is generally decreasing in infancy, constant during adulthood and increasing in old age. We should recognize that to use the Weibull model it is necessary to discard either data on infancy or on old age to make it suited to a monotone hazard model. A realistic hazard function for the whole life span has to have what has been called a bath tub shape. So which probability distribution should we use? Unfortunately most of the common probability distributions do not have a bath tub shaped hazard rate. (See Rajarshi and Rajarshi (1988)). A pragmatic strategy for summarizing therefore chooses a suitable hazard function based on sample data and derives the corresponding probability distribution.

We already know that $F_n(x)$, the sample distribution function is a good estimator of F(x). Since the integrated hazard function H(x) is nothing but $-\ln(1 - F(x))$, a natural estimator for it is $H_n(x) = -\ln(1 - F_n(x))$. The data to be summarized therefore are n pairs $(x_i, H_n(x_i))$, $i = 1, 2, \ldots, n$. To these data points we can fit the integrated hazard function chosen.

Suppose we want to fit a monotone hazard function $h(x) = ax + b$. Clearly $b \geq 0$ otherwise for small x, h (x) may become negative. Similarly 'a' must also be nonnegative or for large x, h (x) may fall below 0. Further the integrated hazard function $H(x) = ax^2/2 + bx + c$ must also be zero at the origin so that the constant of integration c =0. The simplest way of estimating 'a' and 'b' is by least squares provided the resulting estimates satisfy the above restrictions. If not, closest admissible values are used. This is the method suggested by Bain (1990).

To get a bath tub shape for the hazard rate we can use a quadratic function given by $h(x) = a + bx + cx^2$. The corresponding cumulative hazard function with constant of integration zero, is $H(x) = ax + bx^2/2 + cx^3/3$.

We can fit this cubic to the data $(x_i, H_n(x_i))$ by least squares. Again there are natural restrictions on a, b and c. For h (x) to be nonnegative, a and c must be nonnegative and the minimum of h (x) (at $x = -b/(2c) \geq 0$) must also be non negative i. e. $4ac \geq b^2$. Thus $-2\sqrt{ac} \leq b \leq 2\sqrt{ac}$. If the fitted value of b turns out to be positive, we conclude that the hazard rate is monotone and not bathtub. If the fitted value of b is below $-2\sqrt{\{ac\}}$, then we discard the estimated value and use $-2\sqrt{\{ac\}}$ instead.

Another model proposed in literature for describing non-monotone hazard rates is due to Hjorth (1980). Here following Hjorth we define

$$h(x) = [\{\theta\}/\{1 + \beta x\}] + \delta x, \quad x \geq 0, \ \beta, \ \delta, \ \theta \geq 0$$

This function can be increasing (e. g. if $\beta = 0$), decreasing (e.g. if $\delta = 0$) and bath tub (when both $\beta, \sim\sim\sim \delta > 0$). Hence Hjorth has called it IDB distribution. It corresponds to the probability density function

$$f(x) = \{ ((1+\beta x)\, \delta + \theta) \exp(-\delta x^2/2) \} / \{ (1+\beta x)^{\{\theta/\beta\}} \}$$

and distribution function

$$F(x) = 1 - \{\exp(-\delta x^2/2)\} / \{(1+\beta x)^{\{\theta/\beta\}}\}$$

In this hazard function, δx represents the effect of old age while $\theta/(1+\beta x)$ represents the effect of infancy. The cumulative hazard is

$$H(x) = \{\theta\}/\{\beta\} \ln(1+\beta x) + \delta x^2/2.$$

Fitting this equation to data is a nontrivial task, to say the least. One simple way to get at least a reasonable initial solution, which can be used in a nonlinear least squares program is to assume various values of β and obtain least squares estimates of θ and δ in each case. The globally least sum of squares indicates the right choice of the triplet of estimators $\hat{\beta}, \hat{\theta}$ and $\hat{\delta}$.

The **hazard ratio** is the effect on this hazard rate of a difference, such as group membership (for example, treatment or control, male or female), as estimated by regression models which treat the log of the hazard rate as a function of a baseline hazard $h_0(t)$ and a linear combination of explanatory variables:

Such models are generally classed proportional hazards regression models (they differ in their treatment of $h_0(t)$, the underlying pattern the hazard rate over time), and include the Cox semi-parametric proportional hazards model, and the exponential, Gompertz and Weibull parametric models.

For two individuals who differ only in the relevant membership (e.g. treatment vs control) their predicted log-hazard will differ additively by the relevant parameter estimate, which is to say that their predicted hazard rate will differ by e^{β}, i.e. multiplicatively by the anti-log of the estimate. Thus the estimate can be considered a **hazard ratio**, that is, the ratio between the predicted hazard for a member of one group and that for a member of the other group, holding everything else constant.

1.3.4 Exercises

E1.3.1 Show by integration of the p.d.f. that the expression for distribution function of the Weibull distribution is correct. Also show that the function satisfies all properties of a distribution function. (What are those properties?)

E1.3.2 Show that for c = 1, weibull distribution reduces to exponential distribution.

E1.3.3 Show that the hazard function of the Weibull distribution is $h(x) = \{c/\alpha\} (x/\alpha)^{\{c-1\}}$.

E1.3.4 Prove the assertion about H(x). "Since F (x) is non-negative, continuous, non-decreasing and since it approaches 1 as $x \to \infty$, H (x) must be nonnegative, non -decreasing and must go to ∞ as x does so."

E1.3.5 Prove that the distribution function of the IDB distribution has the pdf and hazard function as given in the text.

1.4 Population of India

1.4.1 General Information: The last item in this part of the book is some information on population of India. Census is a survey that collects information through a complete enumeration of people. Modern censuses began in India under the British rule. The first census was organized in 1872. It is the practice to conduct a census every 10 years. The mammoth task is carried out by the office of Registrar General of India. Their website should be visited and examined.

Table 1.4.1 (based on 2001 census) gives some basic numbers. Our population has crossed the one billion mark. This is about one sixth of the world population. We have more males than females. This may be due to female fetuses being aborted, female infanticide, female child getting less attention etc. This has to be improved. (Find out how different states differ in this respect. Kerala in fact has more females than males.) Please find data about other countries on this matter and compare with India.

Table 1.4.1 :Trends in Census Population of India (1901-2000)

S. No.	Year	Total Population in Lakhs Males	Females	Total	Sex ratio Females/ 1000 Males	Density[3] of Population/ Sq K M	Decennial change (%)[7]	Average Exponential Growth Rate	% of Urban Population to Total Population
1	1901	1207.9[2]	1173.6	2384.0[1]	972[4]	77	N.A.	N.A.	10.85
2	1911	1283.9	1237.1	2520.9	964	82	5.75	0.56	10.29
3	1921	1285.5	1227.7	2513.2	955	81	0.31	0.03	11.18
4	1931	1429.3	1357.9	2789.8	950[4]	90	11.00	1.04	11.99
5	1941	1636.9	1546.9	3186.6	945[4]	103	14.22	1.33	13.86
6	1951	1855.3	1755.6	3610.9	946	117	13.31	1.25	17.29
7	1961	2262.9	2129.4	4392.3	941	142	21.64	1.96	17.97
8	1971	2840.5	2641.1	5481.6	930	177	24.80	2.22	19.91
9	1981[1]	3533.7	3299.5	6833.3	934	216	24.66	2.20	23.33
10	1991[4]	4393.6	4070.6	8464.2	926	267	23.87	2.14.	25.70
11	2001[5]	5321.6	4964.5	10286.1	933	325	21.52	1.95	27.82

Table 1.4.2: Estimated Birth Rate, death rate and Natural Growth Rate in India 1901-2001										
Serial No.	Year	Birth Rate			Death rate			Natural growth rate		
		Rural	Urban	Combined	Rural	Urban	Combined	Rural	Urban	Combined
	1	2	3	4	5	6	7	8	9	10
1	1991	30.9	24.3	29.5	10.6	7.1	9.8	20.3	17.2	19.7
2	1992	30.9	23.1	29.2	10.9	7.0	10.1	20.0	16.1	19.1
3	1993	30.4	23.7	28.7	10.6	5.8	9.3	19.8	17.9	19.4
4	1994	30.5	23.1	28.7	10.1	6.7	9.3	20.4	16.4	19.4
5	1995	30.0	22.7	28.3	9.8	6.6	9.0	20.2	16.1	19.3
6	1996	29.3	21.6	27.5	9.7	6.5	9.0	19.6	15.1	18.5
7	1997	28.9	21.5	27.2	9.6	6.5	8.9	19.3	15.0	18.3
8	1998	28.0	21.0	26.5	9.7	6.6	9.0	18.3	14.5	17.5
9	1999	27.6	20.8	26.1	9.4	6.3	8.7	18.2	14.5	17.4
10	2000	27.6	20.7	25.8	9.3	6.3	8.5	18.3	14.0	17.3
11	2001	27.1	20.3	25.4	9.1	6.3	8.4	18.1	14.0	17.0
12	2002	26.6	20.0	25.0	8.7	6.1	8.1	17.9	13.9	16.9
13	2003	26.4	19.8	24.8	8.2	6.0	8.0	17.8	13.8	16.8
14	2004	25.9	19.0	24.1	8.1	5.8	7.5	17.8	13.3	16.6
15	2005	25.6	19.1	23.8	8.1	6.0	7.6	17.5	13.1	16.3
16	2006	25.2	18.8	23.5	8.1	6.0	7.5	17.2	12.8	16.0
Note: Rate Per 1000 population per Annum										

1.4.2 The age composition of our population deserves attention. About 57% people are in the working age group (if we assume that people stop working at the age of 60 years). The percentage of senior citizens is about 8 and that of children (0-14) is about 35. If these two groups are regarded as dependent on others, 43 % of our population is dependent. Find similar information about other countries. Indian population is regarded as relatively young, whereas some other countries (e.g. Japan) have a more graying population.

Table 1.4.3: Age X Sex Composition						
Age Groups	Persons(* 10^7)	%	Males(* 10^7)	%	Females(* 10^7)	%
6 years and below	16.38	15.92	8.50	15.97	7.88	15.87
7 to 14 years	19.98	19.42	10.45	19.63	9.53	19.19
15 to 59 years	58.56	56.93	30.34	57.01	28.23	56.86
60 years and	7.66	7.45	3.78	7.10	3.89	7.83
Age not stated	0.27	0.27	0.15	0.28	0.12	0.25
Total	102.87		53.22		49.65	

Visit the website http://censusindia.gov.in/ to get more information.

Now let us turn to issues concerning health.

1.4.3 Infant mortality rate (IMR)

Death of a child is of course a traumatic event for any family. Public health efforts have to focus on reduction of such events. Table below gives values of infant mortality rates (estimated number of deaths per 1000 live births, in the first year of life) in states. Over the years these are declining, but too slowly. Are IMR values high? Yes, by one order of

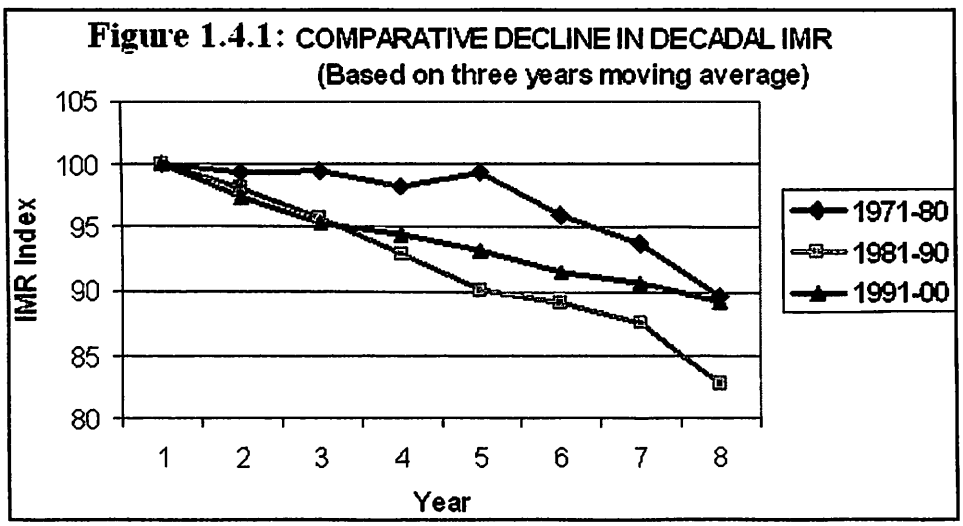
magnitude! In other words, the best figures in the world are about a tenth of our national average. Within India, there is huge variation. The lowest figure is in Kerala. It is comparable with corresponding figures for industrialized countries.

Table 1.4.4: Provisional Estimates of Infant Mortality Rate (2000)

	Total	Rural	Urban	Disparity
India	68	74	43	31
Bigger States				
Andhra Pradesh	65	74	36	38
Assam	75	78	35	43
Bihar	62	63	53	10
Gujarat	62	69	45	24
Haryana	67	69	57	12
Karnataka	57	68	24	44
Kerala	14	14	14	0
Madhya Pradesh	88	94	54	40
Maharashtra	48	57	33	24
Orissa	96	99	66	33
Punjab	52	56	38	18
Rajasthan	79	83	58	25
Tamil Nadu	51	57	38	19
Uttar Pradesh	83	87	65	22
West Bengal	51	54	37	17

Source: Sample Registration System Bulletin. Vol 32(2) Oct 2001
N.A.: Not available due to part-receipt of returns.
Note: Infant mortality rates for Bigger States are estimates of 1999 figures, and for Smaller States and Union Territories are for the period 1998-2000

Figure 1.4.1 shows the overall pattern of decline in IMR. Note that we are treating the value at the beginning of the decade as 100. Do not read the value 100 as the mortality rate. The rate was not the same in 1971, 1981 and 1991.



The graph shows how infant mortality rate is declining. In each decade, it starts at 100 and declines to 90 (82 for 1981-90). In spite of this continuous decline the level is still too high.

Of the estimated 9.7 million children in the world dying before completing five years of age, million, or 21%, are in India . 50% of child morality is due to neo-natal reasons, as opposed to 37% across the world. The other causes are pneumonia (19%), diarrhoea (17%), malaria (8%) etc. Other findings of the report are: 9.4 million children in India are not immunized. 8.3 million children weigh less than 2,500 gm at birth. One-third of all underweight children in the world are in India. States with high rates of underweight children are Madhya Pradesh, Jharkhand, Bihar, Gujarat, Orissa, Chhattisgarh, Uttar Pradesh and Meghalaya. The maternal mortality rate is 450 per 100,000 live births, according to the report, although domestic sources put it at 301

1.4.4 Causes of Death

Registrar General of India is the office that keeps track of causes of death (Table 1.4.6). These statistics show where effort is needed to give relief to population. Similar statistics are also available for world population (Table 1.4.5). These can be compared.

Table 1.4.5-A: Deaths Due to Major causes for Groups of Countries by Income (Low and Middle)

	Low-income countries			Middle-income countries	
Disease group	Deaths in millions	% of deaths	Disease group	Deaths in millions	% of deaths
Lower respiratory infections	2.94	11.2	Stroke and other	3.47	14.2
Coronary heart disease	2.47	9.4	Coronary heart disease	3.40	13.9
Diarrhoeal diseases	1.81	6.9	Chronic obstructive	1.80	7.4
HIV/AIDS	1.51	5.7	Lower respiratory infection	0.92	3.8
Stroke and other cerebrovascular	1.48	5.6	Trachea, bronchus, lung	0.69	2.9
Chronic obstructive pulmonary	0.94	3.6	Road traffic accidents	0.67	2.8
Tuberculosis	0.91	3.5	Hypertensive heart disease	0.62	2.5
Neonatal infections	0.90	3.4	Stomach cancer	0.55	2.2
Malaria	0.86	3.3	Tuberculosis	0.54	2.2
Prematurity and low birth weight	0.84	3.2	Diabetes mellitus	0.52	2.1

Table 1.4.5-B: Deaths Due to Major causes for Groups of Countries by Income (High and Worldwide)

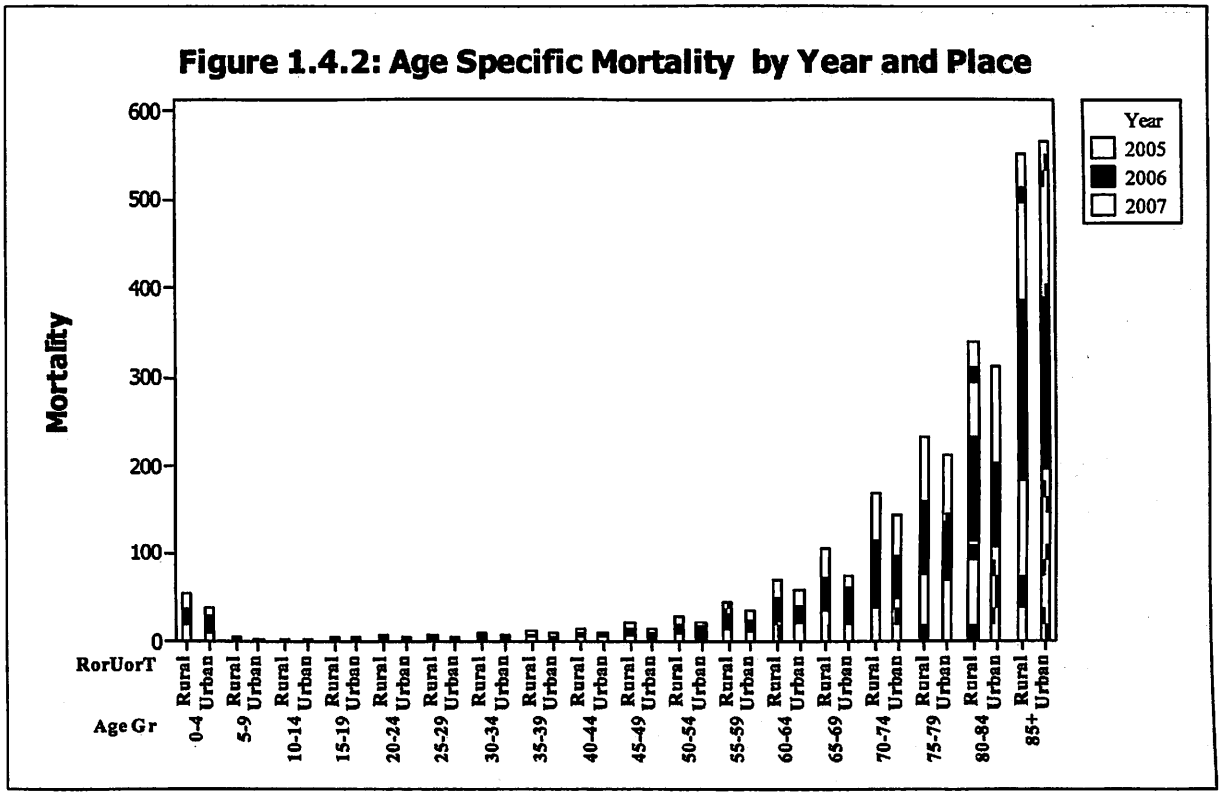
	High-income countries			World	
Disease group	Deaths in millions	% of deaths	Disease group	Deaths in millions	% of deaths
Coronary heart disease	1.33	16.3	Coronary heart disease	7.20	12.2
Stroke and other cerebrovascular diseases	0.76	9.3	Stroke and other cerebrovascular diseases	5.71	9.7
Trachea, bronchus, lung cancers	0.48	5.9	Lower respiratory infections	4.18	7.1
Lower respiratory infections	0.31	3.8	Chronic obstructive pulmonary	3.02	5.1
Chronic obstructive pulmonary	0.29	3.5	Diarrheal diseases	2.16	3.7
Alzheimer and other dementias	0.28	3.4	HIV/AIDS	2.04	3.5
Colon and rectum cancers	0.27	3.3	Tuberculosis	1.46	2.5
Diabetes mellitus	0.22	2.8	Trachea, bronchus, lung cancers	1.32	2.3
Breast cancer	0.16	2.0	Road traffic accidents	1.27	2.2
Stomach cancer	0.14	1.8	Prematurity and low birth weight	1.18	2.0

Table 1.4.6: Percent Deaths Due to Major causes (India Rural)		
Disease group	Year	
	1993	1998
Unknown	20	18
Respiratory system (e.g. pneumonia)	14	17
Infectious and parasitic (e.g. malaria)	13	13
Circulatory system (e.g. heart attack or blood pressure)	7	9
Perinatal period (e.g. infant tetanus or jaundice)	10	7
Venomous animal (snakes or dogs etc)	5	7
Central nervous system inflammation (e.g. meningitis)	4	0
Viral infection	4	4
Neoplasm (cancers)	3	3
Blood and blood forming organs (e.g. anemia)	3	2
Digestive system (e.g. hepatitis b or c)	1	2
Others	16	18
Total	**100**	**100**

1.4.5 Age specific mortality rates

The table below shows age specific mortality for three years. The same information is displayed in the bar chart below. You will notice that there is substantial mortality for age group 0-4 years. Then the hazard declines and stays low till about 55 after which there is a persistent increase.

Table 1.4.7: Age Specific Mortality Rate in India										
		2005			2006			2007		
Sr.No.	Age Gr.	Rural	Urban	Total	Rural	Urban	Total	Rural	Urban	Total
1	0-4	19.5	10.3	17.3	16.3	17.7	17	18	9.6	16
2	5-9	1.8	0.9	1.6	1.3	1.5	1.4	1.5	0.7	1.3
3	10-14	1.3	0.7	1.1	0.9	1	1	1.1	0.6	1
4	15-19	1.9	1.3	1.7	1.5	1.6	1.6	1.5	1	1.4
5	20-24	2.2	1.6	2	2	2.1	2	2.2	1.5	2
6	25-29	2.5	1.7	2.2	2.6	2	2.3	2.2	1.7	2.1
7	30-34	2.7	2	2.5	3.1	2.2	2.7	2.8	2	2.6
8	35-39	3.4	2.8	3.2	4.2	2.5	3.4	3.6	2.8	3.3
9	40-44	4.3	3.5	4	5.3	3	4.2	4.7	3.8	4.4
10	45-49	6.4	5.3	6	7.4	4.2	6	6.5	5.2	6.1
11	50-54	8.6	7.8	8.4	9.9	5.9	7.7	8.9	7.7	8.6
12	55-59	14	11.7	13.3	15.4	10.2	13	13.5	11.9	13.1
13	60-64	22	19	22.2	24.4	18	21.2	23.6	20	22.6
14	65-69	33.4	13.3	32.6	38.7	28.9	33.7	33.6	30.2	32.7
15	70-74	54	49.2	52.8	56.8	47.6	52.1	59.1	48.8	56.6
16	75-79	72.8	77.2	74	84.4	67	75.5	76.7	68.9	74.6
17	80-84	106.4	107.8	106.7	118.3	98.3	107.8	115.8	107	113.6
18	85+	166.7	176.9	169.3	190.7	188.6	189.6	194.4	198.7	195.5
All	Ages	8.1	6	7.6	8	7	7.5	8	6	7.4
Source: Statistical Report, Registrar General of India, 2007.(Latest)										



Following table shows trend in life expectancy over time. There are two patterns to be noticed. Firstly values for female are higher at each time point. Secondly for each gender values increase gradually over time.

Serial No.	Census Year	At Birth	
		Male	Female
	1	2	3
1	1991-95[2]	59.7	60.9
2	1992-96[2]	60.1	61.4
3	1993-97[2]	60.4	61.8
4	1994-98[2]	60.6	62.2
5	1995-99[2]	60.8	62.5
6	1996-2000	61.0	62.7
7	1997-2001	61.3	63.6
8	1998-2002	61.6	63.3
9	1999-2003	61.8	63.5
10	2000-2004	62.1	63.7
11	2001-2005	62.3	63.9
Notes: Based on SRS, Relates to the period 1991-95			

1.4.6 Exercises:

E1.4.1 Table 1.4.9 gives information on statewise slum population in India. Try to find a variable in that is well correlated with %population living in slums. Is it total population of the state or the number of cities or %urban population or what?

Table 1.4.9: State/UT wise Urban and Slum population in India 2001 (Census)

S. No.	State/UT	No of cities/towns reporting slums	Total urban population	Population of cities/ towns reporting slums	Total slum population	Percentage of slum Population to total Urban population	Population of cities/ towns reporting slums
	India	640	283,741,818	184,352,421	42,578,150	15.00	23.10
1	Andhra pradesh	77	20,808,940	16,090,585	5,187,493	24.90	32.20
2	Assam	7	3439,240	1,371,881	82,289	2.40	6.00
3	Bihar	23	8681,800	4,814,512	531,481	6.10	11.00
4	Chhatisgarh	12	4185,747	2,604,993	817,908	19.50	31.40
5	Goa	2	670,577	175,536	14,482	2.20	8.30
6	Gujarat	41	18,930,250	12,697,360	1,866,797	9.90	14.70
7	Haryana	22	6,115,304	4,296,670	1,420,407	23.20	33.10
8	Jammu and Kashmir	5	2,516,638	1,446,178	268,513	10.70	18.60
9	Jharkhand	11	5,993,741	2,422,943	301,569	5.00	12.40
10	Karnataka	35	17,961,529	11,023,376	1,402,971	7.80	12.70
11	Kerala	13	8,266,925	3,196,622	64,556	0.80	2.00
12	Madhya Pradesh	43	15,967,145	9,599,007	2,417,091	15.10	25.20
13	Maharashtra	61	41,100,986	33,635,219	11,202,762	27.30	33.30
14	Meghalaya	1	454,111	132,867	86,304	19.00	65.00
15	Orissa	15	5,517,238	2,838,014	629,999	11.40	22.20
16	Punjab	27	8,262,511	5,660,268	1,159,561	14.00	20.50
17	Rajasthan	26	13,214,375	7,668,508	1,294,106	9.80	16.90
18	Tamilnadu	63	27,483,998	14,337,225	2,866,893	10.40	20.00
19	Tripura	1	545,750	189,228	29,949	5.50	15.80
20	Uttar Pradesh	69	34,539,582	21,256,870	4,395,276	12.70	20.70
21	Uttarkhand	6	2,179,074	1,010,188	195,470	9.00	19.30
22	West Bengal	59	22,427,251	15,184,596	4,115,980	18.40	27.10
23	A & N Islands	1	116,198	99,984	16,244	14.00	16.20
24	Chandigarh	1	808,515	808,515	107,125	13.20	13.20
25	Delhi	16	12,905,780	11,277,586	2,029,755	15.70	18.00
26	Pudicherry	3	648,619	513,010	73,169	11.30	14.30

E1.4.2 Fit trends to birth rates and death rates given in table 1.4.2. Which of these rates declines faster? Can you suggest why?

E1.4.3 What is the type of hazard rate you see in the table 1.4.7 for age specific mortality? Is it constant or increasing or decreasing?

E1.4.4 Fit a suitable (linear or polynomial regression) model to data in each column in the following table (except the first column!).

Table 1.4.10: Trends in Selected Indicators: India

Year	CBR	CDR	IMR	GNP at 93-94 prices in Rs (Crores)	Per Capita NNP at 93-94 prices in Rs (Crores)	Annual growth Rate of NNP	Female Literacy Rate
1950-51	39.9	27.4	148	139912	3687.1	2.4	8.86
1960-61	41.7	22.8	138	205196	4429.4	7.0	15.35
1970-71	41.2	19.0	120	293933	5001.8	5.1	21.97
1980-81	37.2	15.0	110	401970	5352.2	7.5	29.76
1990-91	32.5	12.5	80	683670	7320.7	5.4	39.30
2000-2001	25.8	8.5	63	1186246	10306.0	4.2	54.16

E1.4.5 There is a notion that wealthier communities have lower population growth rates (and lower birth rates and death rates) on the other hand poorer communities have higher birth rates, death rates and growth rates. . Following data give us the income levels in states.

Table 1.4. 11: State wise per capita income (Rs.) by year and IMR per 1000 live births (2006)

State	1999	2000	2001	2002	2003	2004	2005	IMR (2006)
Andhra Pr.	15507	17243	18630	19568	22041	23729	26211	56
Arunachal Pr.	14107	15538	17893	17434	19707	22542	23788	40
Assam	12269	12797	13153	14600	15687	16825	18598	67
Bihar	5766	6396	6197	6928	6913	7467	7875	60
Jharkhand	12747	10294	10972	11865	12941	17493	19066	49
Goa	42296	43735	44110	48839	54577	66135	70112	15
Gujarat	18864	18392	19823	22683	26922	29468	34157	53
Haryana	21966	24138	26077	28259	31509	35044	38832	57
HimachalPr.	20806	22795	24608	26627	28333	31140	33805	50
J & K	13745	14301	15031	16452	17528	18630	NA	52
Karnataka	16758	17464	17776	19041	20515	24199	27291	48
Kerala	19294	19917	21047	23207	25645	27864	30668	15
Madhya Pr.	12384	11862	12697	12303	14306	14534	15647	74
Chattisgarh	11761	10985	12443	13145	16098	18068	20151	61
Maharashtra	23340	22992	24450	26697	29770	32979	37081	35
Manipur	13260	12369	12801	13250	14728	18386	20326	11
Meghalaya	14611	16100	17936	18756	20729	21915	23420	53
Mizoram	16443	17826	19430	20896	21963	22417	NA	25

Relate income levels to various parameters of population growth given in the vital statistics table that is reported below.

Table 1.4.12: Vital Rates based on Sample Registration System – States and union Territories and India (2009) per 1000 Population

States/ union Territories	Birth Rate			Death rate			Natural Growth rate		
	Total	Rural	Urban	Total	Rural	Urban	Total	Rural	Urban
India	22.5	24.1	18.3	7.3	7.8	5.8	15.2	16.3	12.5
Andaman and Nicobar islands	16.3	16.5	16.1	4.1	4.4	3.6	12.2	12.0	12.5
Andhra Pradesh	18.3	18.8	17.0	7.6	8.5	5.5	10.6	10.3	11.5
Arunachal Pradesh	21.1	22.6	14.9	6.1	7.0	2.5	15.0	15.6	12.4
Assam	23.6	24.9	15.9	8.4	8.8	5.9	15.2	16.1	10.1
Bihar	28.5	29.3	22.2	7.0	7.2	5.8	21.5	22.1	16.6
Chandigarh	15.9	22.1	15.3	3.9	3.9	3.9	12.0	18.2	11.4
Chhatisgarh	25.7	27.2	19.0	8.5	8.1	6.4	17.6	18.8	12.6
Dadar and Nagar Haveli	27.0	26.5	28.8	4.8	5.2	3.7	22.1	21.3	25.2
Daman and Diu	19.2	19.6	18.8	5.1	5.4	4.4	14.2	14.1	14.2
Delhi	18.1	19.9	17.8	4.4	4.8	4.3	13.8	15.0	13.5
Goa	13.5	13.0	13.9	6.7	8.2	5.8	6.8	4.8	8.1
Haryana	22.7	23.8	20.1	6.6	7.1	5.8	16.0	16.7	14.5
Himachal Pradesh	17.2	17.8	11.7	7.2	7.4	4.9	10.0	10.4	6.8
Jammu and Kashmir	18.6	19.9	13.7	5.6	6.0	4.7	12.8	13.9	9.0
Jharkhand	25.6	27.1	19.2	7.0	7.4	5.3	18.6	19.3	13.9
Karnataka	19.5	20.6	17.6	7.2	8.3	5.3	12.3	12.3	12.3
Kerala	14.7	14.6	14.9	6.8	6.8	6.5	7.9	7.8	7.3
Lakshadweep	15.0	15.5	14.6	5.8	5.4	6.3	9.2	10.1	8.3
Madhya Pradesh	27.7	29.7	20.8	8.5	9.2	6.1	19.2	20.5	14.8
Maharashtra	17.6	18.1	16.9	6.7	7.6	5.5	10.9	10.6	11.4
Manipur	15.4	15.5	15.5	4.7	4.6	5.0	10.7	10.8	10.5
Meghalaya	24.4	26.4	15.0	8.1	8.6	5.7	16.3	17.9	9.4
Mizoram	17.6	21.7	13.2	4.5	5.0	4.1	13.0	16.7	9.2
Nagaland	17.2	16.4	17.3	3.6	3.7	3.1	13.6	13.7	13.2
Orissa	21.0	21.9	15.7	8.8	9.2	6.8	12.2	12.7	8.9
Pudicherry	16.5	17.1	16.2	7.0	7.4	6.8	9.5	9.7	9.4
Punjab	17.0	17.7	15.8	7.0	7.8	5.8	9.9	9.9	10.1
Rajasthan	27.2	28.4	23.2	6.6	6.7	6.1	20.6	21.7	17.1
Sikkim	18.1	18.4	16.0	5.7	6.0	3.9	12.3	12.4	12.1
Tamilnadu	16.3	16.5	16.0	7.6	8.5	6.6	8.6	8.0	9.4
Tripura	14.8	15.5	11.6	5.1	5.0	5.5	9.7	10.5	6.1
Uttar Pradesh	28.7	29.7	24.7	8.2	8.6	6.5	20.5	21.1	18.3
Uttarakhand	19.7	20.6	16.3	6.5	6.9	5.2	13.2	13.7	11.0
West Bengal	17.2	19.1	12.1	6.2	6.1	6.4	11.0	13.0	5.7

Source: Office of the Registrar General, India

E1.4.6 Compare population growth rate for small against the large states.

E1.4.7 http://www.cbhidghs.nic.in/writereaddata/linkimages/Demographic%20

Indicators133750

Visit the above website- the source for many of these tables.

Chapter 2

Epidemiology

2.1 What is epidemiology?

Here is how Wikipedia (the free on-line encyclopedia) defines the term.

2.1.1 Epidemiology is the study of factors affecting health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine. It is highly regarded ... for identifying risk factors for disease and determining optimal treatment approaches to clinical practice.

The key words are 'populations', 'disease' and 'risk factors'. These bring the study field very close to statistics. In fact epidemiology is one of the most important parts of the interface between medicine and statistics.

Epidemiology, "the study of what is upon the people," is derived from the Greek terms *epi* = among; *demos* = people; *logos* = study; suggesting that it applies only to human populations. But the term is widely used in studies of zoological populations (veterinary epidemiology), although the term 'epi-zoology' is available, and it has also been applied to studies of plant populations (botanical epidemiology).

The Greek physician Hippocrates (Do you know about the oath that doctors are supposed to take? If not, check the internet and read about it.) is sometimes regarded the father of epidemiology. He examined the relationships between occurrence of disease and environmental influences. He coined the terms *endemic* (for diseases usually found in some places but not in others i.e. incidence varying in space) and *epidemic* (for diseases that are seen at some times but not others i.e. incidence varying in time).

In the medieval Islamic world, physicians discovered the contagious nature of infectious disease. In particular, the Persian physician Avicenna, considered a "father of modern medicine," (1020s), discovered the contagious nature of tuberculosis and sexually transmitted diseases, and the distribution of disease through water and soil. He introduced the method of quarantine as a means of limiting the spread of a contagious disease. He also used the method of risk factor analysis.

Epidemiology is concerned with the incidence of disease in populations and does not address the question of the cause of an individual's disease. Another objective is study of trends in diseases. Is the incidence on the rise? Is it on a decline? Was some public health measure effective in controlling a disease? In India, malaria was a serious problem at the time of independence. In the next twenty years malaria was brought under control in most parts of the country by extensive use of DDT. In fact the elaborate administrative machinery for malaria

eradication was wound up. Now malaria has revived. Hence the efforts to control have also been revived.

It is important for students to appreciate that we are discussing serious matters here. When an epidemic hits a society, things turn upside down. In April 2009, an epidemic of swine flu hit Mexico. Visitors from Mexico to other countries suddenly became suspects as possible carriers of swine flu. The government shut down everything (schools, factories, offices etc) for 5 days. If this is not serious, we do not know what is. In August 09 the sickness came to India and everyone was worried. Good long term data can give valuable guidance when such situation arises.

We need to know formal definitions of two words viz. incidence and prevalence of a disease in a community. **Incidence** is the proportion of new cases of a disease in a community in a year. Incidence rate of 1% implies that in a year, there is one new case per hundred people in the community. **Prevalence** of a disease in a community is the total number of cases divided by the size of the community. Prevalence is a measure that is more relevant to a disease that is long lasting (e.g. HIV AIDS) while incidence is more appropriate for a short duration sickness such as common cold or chicken pox.

Viewed in another way, incidence of a disease is the rate at which **new** cases occur in a population during a specified period. Sometimes measurement of incidence is complicated by changes in the population at risk during the period when cases are ascertained, for example, through births, deaths, or migrations. This difficulty is overcome by relating the numbers of new cases to the *person years at risk*, calculated by adding together the periods during which each individual member of the population is at risk during the measurement period. Thus incidence is defined as:[Number of new cases]/[Total person years at risk].

A person, who leaves the area of interest after six months, will only contribute half person year to the denominator. A new arrival in the last quarter will add ¼ person year to the denominator and so on.

Analogously, another way to explain prevalence of a disease is that it is the *proportion of a population that is 'cases' at a point in time.* Prevalence is an appropriate measure only in such relatively stable conditions, and it is unsuitable for acute disorders.

2.1.2 Precision and bias

Precision in epidemiological variables is a measure of random error. Precision is inversely related to random error, so that to reduce random error is to increase precision. Confidence intervals are computed to demonstrate the precision of relative risk estimates. The narrower the confidence interval, the more precise the relative risk estimate.

There are two basic ways to reduce random error in an epidemiological study. The first is to increase sample size of the study. In other words, add more subjects to your study. The second is to reduce variability in measurement in the study. This might be accomplished by using a more accurate measuring device or by increasing the number of measurements.

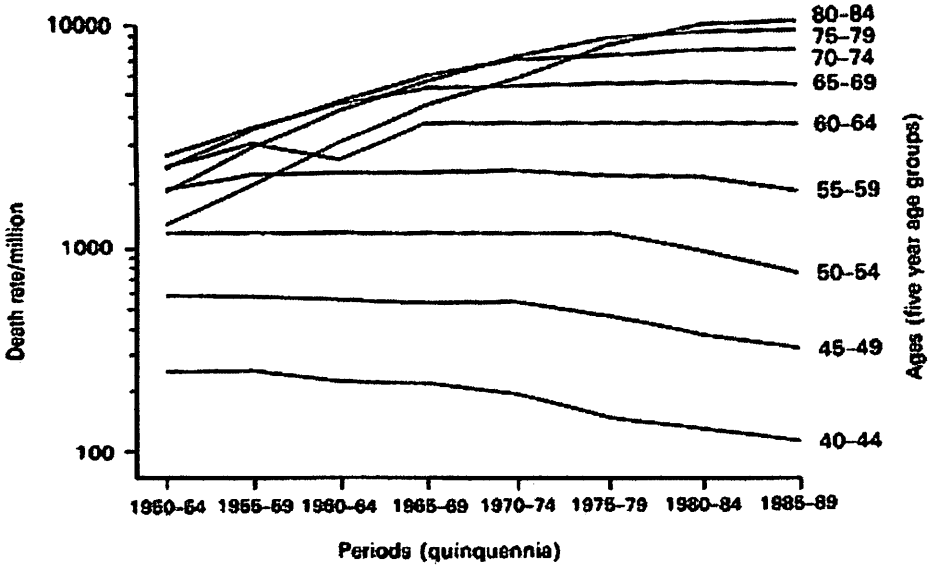
Note that if sample size or number of measurements is increased, or a more precise measuring tool is purchased, the costs of the study are usually increased. There is usually an uneasy balance between the need for adequate precision and the practical issue of study cost.

Bias is a systematic error. Bias occurs when there is a difference between the true value (in the population) and the observed value (in the study) from any cause other than sampling variability. An example of systematic error is zero error of an instrument. Because the error happens in every instance, it is systematic. Conclusions you draw based on that data may still be incorrect. But the error can be reproduced in the future (e.g. by using the same instrument that has a bad setting.) and hence correction may be relatively easy. Bias can slip into data because of the method of selection of subjects in a study. If volunteers are used in medical studies, they may be healthier (motivated by desire to help) or more sick (hence keen to get help). If many of the selected subjects are absent or unwilling to participate in the study that can cause a bias too.

2.1.3 Crude and specific rates

A crude incidence, prevalence, or mortality (death) rate is one that relates to results for a population taken as a whole, without subdivision or refinement. The crude mortality from lung cancer in men in England and Wales during 1985-89 was 1034/million/year compared with 575/million/year during 1950-54 (see Figure 2.1.1).

Figure 2.1.1: *Mortality from lung cancer in men in England and Wales, 1950-89*



So one thing is clear, namely that cancer deaths increased hugely in 40 years. However, it does not tell you the whole story which is more complex. In fact the mortality figures seen by age show that mortality from lung cancer was declining in younger men while going up in the elderly. In other words, aggregated data hide patterns that exist in subgroups.

It is often helpful to break down results for the whole population to give rates specific for age and sex, but it is important to give results using suitable age classes. Either decadal classes e.g. 5-14, 15-24, and so on or five yearly classes e.g. 5-9, 10-14, and so on are recommended. Overlapping classes (5-10, 10-15) should be avoided.One warning is in order. When we break aggregated data into subgroups, the patterns become clearer as in this case; but sometimes the patterns seen may get reversed. Such a situation is called Simpson's paradox and we will visit it again in a later section.

Traditionally, statistical inference has been based on hypothesis testing. A *null hypothesis* about the target population is formulated. Then p value is calculated. This is the probability of obtaining an outcome in the study sample as extreme as that observed, simply by chance (assuming that the null hypothesis holds). Suppose in a city half the population drinks water from bore wells and the other half drinks water from municipal supply. Suppose there is an outbreak of gastroenteritis. Each patient is asked about the source of drinking water. Suppose 5 patients in a row answer that they drink well water. Should we feel concerned that well water may be the cause of the sickness? Here our null hypothesis is that a patient has probability 0.5 of being a user of well water. Is the evidence enough to reject the null hypothesis? What is our p value? It is the probability of getting all 5 patients to be well water consumers. The probability is to be calculated assuming truth of the null hypothesis. The probability is $(0.5)^5$ which is 0.03125. A p value of <0.05 would imply that under this null hypothesis, the probability of selecting a random sample as extreme as that observed in the study would be below one in 20. The lower the p value, the higher the inclination to reject the null hypothesis and adopt a contrary view. In this case we may conclude that too many patients are well water drinkers and it would be good to get that water tested for presence of pathogens. In other words a p value below a stated threshold (for example, 0.05) is deemed to be (*statistically*) significant, but this threshold is arbitrary. There is no reason to attach much greater importance to a p value of 0.049 than to a value of 0.051. A p value depends not only on the magnitude of any deviation from the null hypothesis, but also on the size of the sample in which that deviation was observed. Failure to achieve a specified level of statistical significance will have different implications according to the size of the study. A common error is to weigh "positive" studies, which find an association to be significant, against "negative" studies, in which it is not. Two case-control studies could indicate similar odds ratios, but because they differed in size one might be significant and the other not. Clearly such apparently opposite findings would not be incompatible. Typically, as sample size goes on increasing, p value goes on declining and eventually, when sample size is large enough, rejection of the null hypothesis is virtually assured.

Because of the limitations of the p value as a summary statistic, epidemiologists today prefer to base statistical inference on confidence intervals.A *statistic* of the study sample, such as an odds ratio or a mean haemoglobin concentration, provides an estimate of the corresponding population *parameter*. Because the study sample may by chance be atypical, there is uncertainty about the estimate. A confidence interval is a range within which, assuming there are no biases in the study method, the true value for the population parameter might be expected to lie. Most often, *95% confidence intervals* are calculated. The formula for the 95% confidence interval is set in such a way that on average 19 out of 20 such intervals will include the population parameter. Large samples are less prone to chance error than small samples, and therefore give tighter confidence intervals.

Whether statistical inference is based on hypothesis testing or confidence intervals, the results must be viewed in context. Assessment of the contribution of chance to an observation should also take into account the findings of other studies. An epidemiological association might be highly significant statistically, but if it is completely at variance with the balance of evidence from elsewhere, then it could still legitimately be attributed to chance. For example, if a cohort study with no obvious biases suggested that smoking protected against lung cancer, and no special explanation could be found, we would probably conclude that this was a fluke result. Unlike p values or confidence intervals, the weight that is attached to evidence from other studies cannot be precisely quantified.

Statistical calculations should be done using a good software package. For problems in epidemiology, the best option is EpiInfo. It is developed by the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia (USA). It has been in existence for over 20 years. It can be used for data entry and analysis including t-tests, ANOVA, nonparametric statistics, cross tabulations, estimates of odds ratios, risk ratios, and risk differences, logistic regression (conditional and unconditional), survival analysis (Kaplan Meier and Cox proportional hazard), and analysis of complex survey data. The free software can be downloaded from http://www.cdc.gov/epiinfo. Here are some more free packages that can be explored depending upon interest of students. CSPro, OpenEpi , X-12-ARIMA , AnSWR , Epi Map .

2.2 Some discoveries based on quantitative analysis

It is important to recognize that in this topic we are going to encounter situations in which quantitative or qualitative data are used to draw major conclusions in the field of medicine. There are many instances in history when data interpretation led to crucial insights. Let us visit some of them.

2.2.1 William Harvey: We begin with the story of blood circulation. This very fundamental feature of our bodies was discovered in 1628 by William Harvey. Remember today we inject a medication in one part of the body and assume that it will be distributed all over. We take the fact of blood circulation for granted. It was not known till Harvey used a quantitative argument

and guessed that blood must circulate. Prior to him people believed that blood was used up in different organs as food. Harvey did simple calculations to see how much blood would pass through the heart each day. In this effort he used estimates of the capacity of the heart, how much blood is expelled with each pulse/beat of the heart, and the number of times the heart beats in a half hour. All of these estimates were purposefully low, so that people could see the vast amount of blood the body would have to produce (if the idea that blood was used up as food were right). He estimated that the capacity of the heart was 1.5 ounces, and that every time the heart pumps, 1/8 of that blood is expelled. This led to Harvey's estimate that about 1/6 of an ounce of blood went through the heart every time it pumped. The next estimate he used was that the heart beats 1000 times every half an hour, which gave 10 pounds6 ounces of blood in a half an hour, and when this number was multiplied by 48 half hours in a day he realized that the liver would have to produce 540 pounds of blood in a day. Clearly it was impossible for the body to produce this much fresh blood every day. The inescapable conclusion was that the same blood came back to the heart. This clashed with the accepted model going back to Galen, the Roman physician, who identified venous (dark red) and arterial (brighter and thinner) blood, each with distinct and separate functions. Venous blood was thought to originate in the liver and arterial blood in the heart; the blood flowed from those organs to all parts of the body where it was consumed. Finally, Harvey's arguments prevailed. We should look for such instances in which data analytic arguments gave rise to interesting and profound medical conclusions.

2.2.2 Florence Nightingale: The next example is that of Florence Nightingale. We know of her as the lady with the lamp, the compassionate nurse who helped British soldiers in the Crimean war in the nineteenth century. In fact her great achievement was not in hospital wards but elsewhere. She examined records and pointed out that more soldiers died in barracks than in battle fields. The more potent cause of death was not the gun but the gutter. Soldiers fell ill because of filthy living conditions and died. This was so alarming and unexpected that the British government swung into action to ensure better living conditions for redcoats. This was not one accidental discovery. Florence Nightingale came to India, the crown jewel among the British colonies and again found that British soldiers were killed less by the sword and more by the swamp and cesspool. Again, the colonial government acted with alacrity and developed well planned and hygienic residential areas (called cantonments) all over India. They continue to be better planned parts of many Indian cities even today.

2.2.3 Cholera in London: Now we turn to cholera. Cholera is an illness caused by a germ invading the bowels. The disease is usually spread by contaminated water supplies. The main symptom is watery diarrhea which leads to fluid depletion and death from dehydration. It has been a killer disease in Asia for over 1,000 years but the first of a series of seven pandemics arrived in Europe in 1817 with devastating effects. The situation in European cities in terms of public health was far from ideal.To get some idea of how bad the sanitation situation was in Britain in general and London in particular, let us read a report in a London newspaper.Henry Mayhew, *Morning Chronicle* (24th September 1849):

"We then journeyed on to London Street, down which the tidal ditch continues its course. In No. 1 of this street cholera first appeared seventeen years ago, and spread up it with fearful virulence; but this year it appeared at the opposite end, and ran down it with like severity. As we passed along the reeking banks of the sewer the sun shone upon a narrow slip of the water. In the bright light it appeared the colour of strong green tea, and positively looked as solid as black marble in the shadow - indeed it was more like watery mud than muddy water; and yet we were assured this was the only water the wretched inhabitants had to drink".

As we gazed in horror at it, we saw drains and sewers emptying their filthy contents into it; we saw a whole tier of door-less privies (toilettes) in the open road, common to men and women, built over it; we heard bucket after bucket of filth splash into it, and the limbs of the vagrant boys bathing in it seemed by pure force of contrast, white as marble.

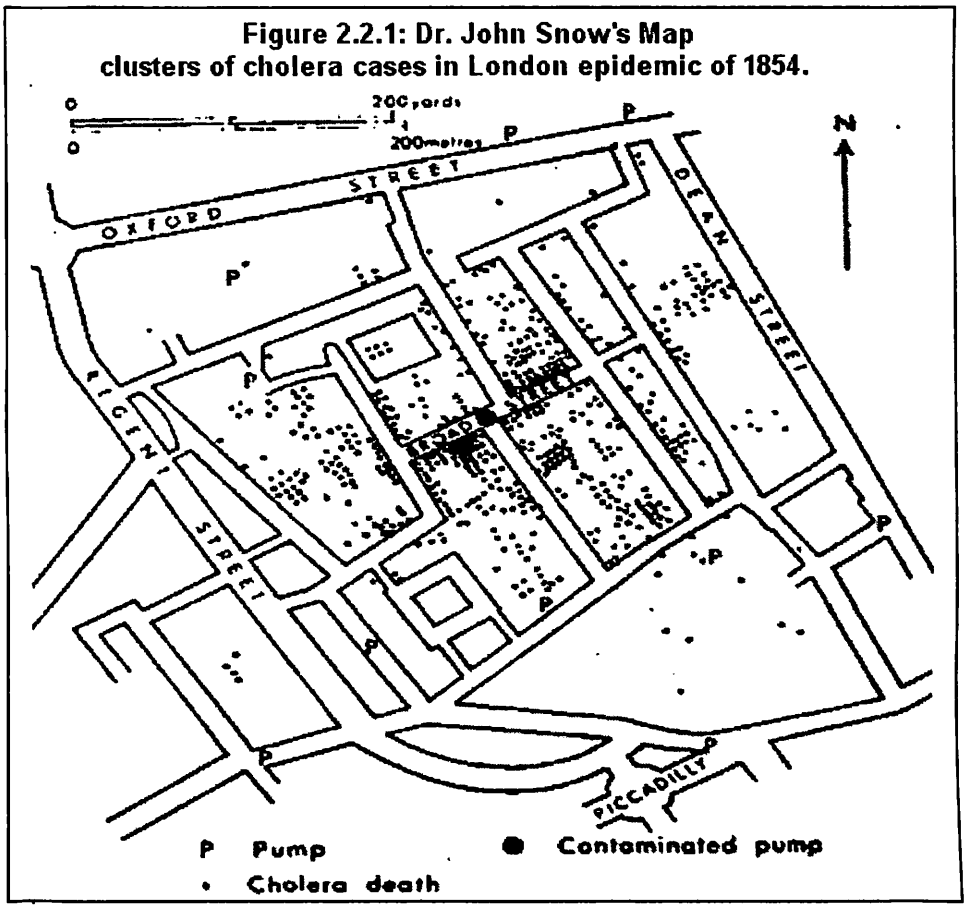
In this wretched place we were taken to a house where an infant lay dead of the cholera. We asked if they really did drink the water. The answer was, "They were obliged to drink the ditch, without they could beg or thieve a pailful of water." But have you spoken to your landlord about having it laid on for you? "Yes, sir and he says he will do it, and do it, but we know him better than to believe him."

Did you know that the situation of poor people in London was so bad even in times when Britain was the most powerful nation on earth? No wonder people became victims of cholera in droves. Here are some numbers. In the summer of 1849 over 33,000 people died of cholera in Britain in three months. Around 13,000 of those who died lived in London. Until the second-half of the 19th century, about 50 per cent of the people who caught cholera died of the disease.

The cause of cholera was unknown. It is remarkable that a breakthrough for residents of London came not from using fancy medicines or special surgical instruments or applying some deep theories from science, but instead by using graphical representation of data. Perhaps it is right to say that the graph helped save many lives by first identifying the 'cause' in 1854. This path breaking graph was drawn by Dr. John Snow. He is famous for the suppression of an 1854 outbreak of cholera in London's Soho district. He plotted the locations of houses that had a patient of cholera. The points seemed to cluster around some public hand pumps (shown by crosses) for water. People drew water from river Thames using those pumps. It turned out that the river water drawn by these pumps was polluted since raw sewage was dumped into it just upstream. So a guess was that polluted water could be responsible. Going further into details, he zeroed in on a public water pump on Broad Street. Notice that the cross on the Broad Street sits in the middle of a large cluster of points. This is in contrast with other crosses which are relatively in empty spaces i.e. there are few cases in areas around those pumps. If this diagnosis was correct, then remedial action was obvious. Stop people from using water from that source. To ensure this he had the handle of the pump removed. There was a decline in the number of

cases and finally the outbreak ended. Just a graphical representation of data about patients was enough to identify the steps needed. This must be one of the most influential uses of a graph. (Some people argued that Snow was not quite the key to control of the outbreak .The epidemic was already in decline when Snow took action.) This has been perceived as a major event in the history of public health and as the founding event of the field of epidemiology.

Figure 2.2.1: Dr. John Snow's Map clusters of cholera cases in London epidemic of 1854.



Of course if water supply is polluted, we get sickness. But many other environmental factors can affect health. We have a study that showed higher mortality levels for higher ambient temperatures. See exercise 1 in 2.2.8.

2.2.4 Puerperal Fever: Fourth example is that of maternal mortality in Austria. Ignaz Philipp Semmelweis (July 1, 1818 - August 13, 1865), was a Hungarian physician who was the Director of the maternity clinic at the Vienna General Hospital in Austria. In those days maternity clinics had such a bad reputation that they were sometimes referred to as deathtraps for women. The death rate during childbirth was high. People suggested more than once that lives could be saved simply by closing the clinics.

There were two maternity clinics at the Vienna General Hospital. Both clinics were free and attractive to poor women. In return, the women agreed to be treated by doctors and midwives under training (interns). The first clinic had an average maternalmortality rate due to puerperal

fever of about 4% and the second clinic's rate was lower, averaging 2.5%. This fact was known outside the hospital. The clinics admitted on alternate days but women begged to be admitted to the second clinic due to the bad reputation of the first clinic. Semmelweis described desperate women begging on their knees not to be admitted to the first clinic.(Students should note that 4% and 2.5% may seem to be not too different. But for the women ready to deliver their babies, the difference seemed literally like difference between life and death. Moral of the story is that numbers by themselves cannot be considered 'similar' or 'close to each other' without reference to context.)

It severely troubled and literally sickened Semmelweis that his First Obstetrical Clinic had a much higher mortality rate due to puerperal fever (name for the sickness of women in childbirth) than the Second Clinic. It "made me so miserable that life seemed worthless". The clinics used almost the same techniques. The only major difference was the individuals who worked there. First Clinic was the teaching service for medical students, while the Second had been selected in 1841 for the instruction of midwives only.

It was in this setting that Semmelweis began eliminating several of the proposed causes of puerperal fever or childbed fever (the sickness that women fell prey to). "To me, it appeared logical that patients who experienced street births would become ill at least as frequently as those who delivered in the clinic. [...] What protected those who delivered outside the clinic from these destructive unknown endemic influences?"

He excluded "overcrowding" as a cause because clinic two was always more crowded as stated above but the mortality was lower. He eliminated climate as a cause because the climate was not different in two rooms in the same building separated only by a common anteroom, and so on. The breakthrough for Ignaz Semmelweis occurred in 1847 following the death of his good friend JakobKolletschka who had been accidentally poked with a student's scalpel while performing a postmortem examination. Kolletschka's own autopsy showed a pathological situation similar to that of the women who were dying from puerperal fever. This was most unexpected. It suggested that the fever experienced by the women in maternity ward may have nothing to do with gender or maternity. Semmelweis immediately proposed a connection between cadavericcontamination and puerperal fever. He then went back 100 years into the history of all the major hospitals in Europe and demonstrated that the increase in the incidence of childbed fever had followed the introduction of regular performance of autopsies in each of these hospitals. This is what statistical evidence can do. It supported the suspicion that the secret of puerperal fever lay in autopsies. *(Of course such data have to be collected painstakingly for a long time before it can throw light on some puzzle. Also, the story underlines the importance of examining historical data. In India, it is very typical of many organizations and institutions to collect data meticulously and then ignore it. Enterprising statistics students can take up the challenge of summarizing such data and finding interesting uses for it.)*

He concluded that he and the medical students carried "cadaverous particles" on their hands from the autopsy room to the patients they examined in the First Obstetrical Clinic. This explained why the student midwives in the second clinic who were not engaged in autopsies and had no contact with corpses experienced a much lower mortality rate.

We shall use the data on death in the Vienna maternity ward in our study of contingency tables. Having identified the cause of the higher mortality rate in the first clinic, Semmelweis could come up with a simple solution for the problem: Washing of hands with chlorinated lime solutions! This was required for interns who had performed autopsies before assisting a woman in child birth. So simple indeed! The mortality rate dropped and became comparable to the Second Clinic's. The mortality rate in April 1847 was 18.3 percent (see table below), hand washing was instituted mid-May, the rates in June were 2.2 percent, July 1.2 percent, August 1.9 percent and, for the first time since the introduction of anatomical orientation, the death rate was zero in two months in the year following this discovery.

Hence people called Semmelweis *"savior of mothers"*. Despite the clarity of his arguments and the concrete proof demonstrated by the reduction in mortality when hand washing procedures were conscientiously followed, Semmelweis faced a wall of opposition within his professional community. Doctors could not face the fact they had been the ones responsible for the epidemic of childbed fever and the death of thousands of young women throughout Europe. Semmelweis was largely ignored, rejected, or ridiculed. He was dismissed from his position as Director of the maternity hospital and offered a lower position which he refused. He then had difficulty finding employment as a medical doctor. Nevertheless he volunteered his services in other hospitals spreading his doctrine and saving lives. Later he began to publicly attack the indifference of the medical community after a practical solution to the problem had long since been demonstrated. He began writing open and increasingly angry letters to prominent European obstetricians, at times denouncing them as irresponsible murderers. His contemporaries, including his wife, believed he was losing his mind and he was in 1865 committed to an asylum (mental institution). Semmelweis died there only 14 days later, possibly after being severely beaten by guards.

Semmelweis' practice earned widespread acceptance years after his death, when Louis Pasteur developed the germ theory of disease which offered a theoretical explanation for Semmelweis' findings. Semmelweis is considered a pioneer of antiseptic procedures.

During 1848 Ignaz Semmelweis widened the scope of his washing protocol to include all instruments coming in contact with patients in labor, and used mortality-rate time series to document his success in virtually eliminating puerperal fever from the hospital ward.

2.2.5 Smoking and cancer: Let us now turn to a very recent and famous example in which numbers pointed to health hazards long before basic science could confirm that the dark hints emerging from statistics are in fact true. It is the case of connection between smoking and cancer.

In the early twentieth century, number of cancer cases began to rise in some European countries. Balance of sexes changed and incidence became higher in men. Proportion of lung cancer cases among cancer cases increased. The largest increase in lung cancer came in men over 45 years. Incidence increased six times between 1930 and 1945. What was the cause? Could it be work related? Could it be some environmental aspect? A review of post-mortem certificates showed that neither could be the explanation. This was in 1947. So, Medical Research Council in England decided to launch a large scale statistical study of past smoking habits of lung cancer patients and control groups. Statistical Research Unit at the London School of Hygiene and Tropical Medicine became the home base of the study. Its findings were published by Austin Bradford Hill and Richard Doll in the British Medical Journal (1950). The authors concluded that there was real association between lung cancer and smoking and smoking was indeed a factor. It became a watershed study. It involved comparison of cases and controls from 20 London hospitals. [Cases were of course patients of lung cancer and controls were patient of other diseases with similar background as cases.] The same pair of researchers carried out a prospective study on the same question between 1951 and 1956.[Prospective study is one in which individuals who do not have the disease are kept under observation for a long period and occurrence of disease is recorded] It compared deaths among English doctors in the smoking, non-smoking and ex-smoking groups. It concluded that death rate increased as the amount of smoking increased. Follow up of this study continued for the next 50 years.

R. A. Fisher argued that Hill and Doll had only shown association and that does not mean causation. He was himself a smoker and was outraged by the proposal in the article of Hill and Doll that people should be discouraged from smoking. (His position on the issue was suspect because he was advisor to the Tobacco Manufacturers' Standing Committee to assist research.) Later, Hill spelt out conditions under which a causal interpretation of correlation would be justified. These are sometimes called Hill's postulates. All applied statisticians should be familiar with these postulates.

Discovery of connection between smoking and lung cancer was a watershed in the acceptability of chronic disease epidemiology to provide legitimate forms of scientific explanation. It was a major paradigm shift to a statistical mode of explanation and causation at the expense of laboratory science. R A Fisher criticized this approach. One reason was that correlation was used to show causation. This was against the wisdom of statistics. Second reason was libertarian. He believed that people should be free to choose. He was outraged by the recommendation of Doll and Hill that people should be discouraged from smoking. Others argued that he was influenced by the fact that he was Adviser to the Tobacco Manufacturers' Standing Committee, set up in 1956 to assist research.

But response of the British government to the recommendation of Hill and Doll was denial and delay because of the tax income from tobacco and also influence of the tobacco industry. Tobacco tax constituted 16% of central revenue in 1950.

2.2.6 Malaria in Mumbai Some of you perhaps wonder why all the above examples come from western sources. Is nothing Indian available? Did Indians do nothing of the kind? We must apologize for our limited knowledge. Perhaps some of you can take it upon yourselves to search for stories from India. We will try to give a couple of examples that we are familiar with.

The first example is about malaria in Mumbai. We know about this material because we were involved in the analysis of the concerned data set. The work is done by a team from Tata Institute of Fundamental Research led by DrSulabha Pathak. This team examined data on malaria cases in Mumbai (admitted to one of two selected government hospitals) in recent years. Of course we already know many things about epidemiology of malaria. It is caused by a microbe called Plasmodium. There are several species of this organism, of which two are common in India. They are P.vivax and P. falciparum. These are carried from one person to another by mosquitoes. Hence malaria can be controlled by controlling mosquito populations. We use bed nets and mosquito repellent coils etc to save ourselves from malaria. But not everything is known. That is why research in epidemiology of malaria is still of interest. TIFR team found that in the medical records of malaria patients in Mumbai, there were more men than women. This was intriguing. Why should it be so? Perhaps there are more men than women in Mumbai. Many men come to Mumbai for work while their wives stay back in the villages. (We are following the example of Semmelweis who tried to understand why death rate in one ward in his hospital in Vienna was higher. He considered one potential cause at a time and ruled it out. Remember epidemiology can be thought of as a detective work.) This explanation does not work because we are not comparing number of malaria cases among men with number of malaria cases among women. We are comparing the *proportion* of male patients taking treatment for malaria (men taking treatment for malaria/ all men taking some treatment) with similarly computed proportion among women. Perhaps men go to doctor more readily and women suffer quietly. Even this explanation does not work (again for the same reason as before). Perhaps among poor people (Generally, people from lower economic stratum dominate the use of government health facilities.) men sleep in the open and are more exposed to mosquito attack. This potential explanation for difference between genders seems plausible. Well, rainy season in Mumbai forces everyone to sleep indoors. So, if the argument is right, then any differences between genders should vanish in rainy season. It does not happen. Differences persist. This line of argument is not taking us anywhere. We can try something else.

How about comparing genders at different ages? Well, this was done. What did we find? Among children, there was no gender difference. Among children visiting health facility, proportion sick with malaria was similar for boys and girls. But as age progressed, differences set in. Among adults, the proportion was higher among males. Later, as we move to older age groups, gradually the difference disappears. Statistical tests can be used to confirm that the gender differences are statistically significant and not just due to chance. So, it appears that there are two groups. One group is adult men. The second group is everyone else, children, women and old men. The first group has a higher incidence than the second group. This is as far as we

can go with data analysis alone. Now the result has to be interpreted. Here, statistical skill is not enough. Doctors and malaria experts must contribute. That is why statisticians have to work in collaboration with experts in the field of application. What is common to people in group 1 that is absent among group 2? What do adult males have that women, boys and old men do not? It is the male sex hormone 'testosterone'. As the levels of this hormone go up, individuals become more prone to malaria. It seems that there were indications of this in earlier studies, but the evidence was not as clear as in the Mumbai study. Implication of this study is clear. Working men need more protection against malaria than other members of the family. Perhaps any practice of using short pants and T shirts should be changed to use of long pants and long sleeve shirts especially during evenings and nights when mosquitoes are more active. In other words, men should accept the handicap and adjust their behavior when possible.

2.2.7 Pneumonia in Gadchiroli

The next example concerns infant mortality among tribal groups in the state of Maharashtra. This work is done by DrAbhay Bang and Dr Rani Bang, a couple which works in the tribal district of Gadchiroli in eastern part of the state. They run an institution called SEARCH (Society for Education And Research in Community Health). Their website http://www.searchgadchiroli.org/ has huge amount of valuable information including research papers. It is worth studying this website.

So, what is the interesting story emerging from the work of Abhay and Rani Bang? We will try to describe one feature of their work. It concerns their successful campaign to reduce infant mortality rate in Gadchiroli area. Note first that infant mortality is a very important index of the well being of a society. Higher the rate poorer the society! It is generally measured as number of deaths among infants (age below 12 months) per 1000 births. It is regarded as a very good indicator of development of the society. Wealthy west European and North American nations have very low rates (say about 10) while poverty stricken African and Asian countries have very high rate (say 200). In India, state of Kerala has the lowest rate while rural UP and rural Bihar may have the highest. (Some information relevant to this point may be found in the tables about population of India in the first part of the book.) Generally tribal groups have high infant mortality. In the area served by the doctor couple, the rate was about 120 in year 1988. It reduced linearly to about 30 by the year 2003. This is a phenomenal achievement! It has attracted the attention from all over the world. How did they manage it? We cannot go into all the aspects this work. We will discuss only one aspect, namely pneumonia. One major cause of infant death is pneumonia. If detected early, it can be treated successfully using antibiotics. Key is early diagnosis. A direct method of checking will need a blood test. That is not possible in remote forest areas. A rough and simple alternative is to count respiratory rate. If in case of a child with cough the rate exceeds 50, it is a good indication of pneumonia. Well, this seems like a very easy way to diagnose the disease without any expertise in medicine. All you do is to count. What could be easier?

The problem was that the only person who visited the family in the village, namely a village health worker, was typically an illiterate woman who could not count to 50. She could count only up to 10. The doctor couple accepted this as a fact of life and thought of a suitable innovation to overcome the difficulty. What was their innovation? It was a simple device fabricated locally. It contained a sand clock that emptied in a minute and also an abacus-like string of beads. The worker turned the sand clock upside down and started counting the respiration rhythm. After a count of 10, one bead was pushed to the other end of the string. Sixth bead was red. If it had to be pushed, diagnosis was that of pneumonia. Very simple numbers, but profound impact!

2.2.8 Declaring an epidemic

Now we will give one example that may be relevant in many places. It came to us from DrShyamAshtekar, a physician active in the field of public health. At present he is associated with the YashwantRaoChavan Maharashtra Open University, Nashik. Think of a remote village that has great difficulty in securing medical help. If there is an occasional case of major sickness, someone walks to the nearest primary health center to get some medication. (Do you know what a PHC is? If not, find out.) If there is, instead, an epidemic of some sort, this mode of operation is quite unsuitable. Instead it may be more efficient to organize a small party of medical personnel and take it to the village so that curative and preventive activities can be launched quickly. If you accept this line of argument, then the critical step is judging whether a village has an epidemic. How do we do that? There has to be a specified number or count. If the count of cases (say in a week) exceeds that number, we declare an epidemic. That is easy enough. The key question is how to decide this number. If the number is too small, we will trigger the alarm often and a lot of preparation may be made when it is not warranted. If the number is too large, by the time the health team arrives on the scene, prevention may be very difficult. Clusters of cases sometimes occur in one place within a short time simply by chance. There can be errors in diagnosis. If you get many such false positives, you end up thinking of an epidemic where there is none. Such a situation has been termed pseudo-epidemics.

One possibility of selecting the number is using Poisson distribution. Sickness is not a frequent event and the Poisson model may be suitable for the number of cases in a week. We can estimate the mean of this Poisson distribution from historical data. Using this estimate, we can calculate the probability that the count of sick people exceeds the observed count. If this probability is small, our trigger should be pulled.

We hope you see that understanding data can be extremely useful in improving health status of a community. One problem with data analysis is that it can, at best, show some connection (association) between disease and some other factors. We know that association (correlation) cannot be treated as causation. But this is not an absolute principle. Since it is very difficult to establish causation directly, applied statisticians always want to go a little beyond

claiming association. This happened in a major way in the case of tobacco and cancer controversy as we have noted earlier.

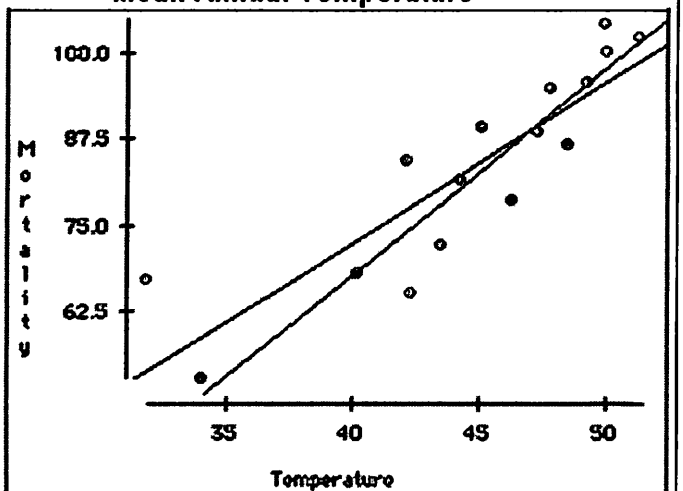
2.2.9 Exercises

E2.2.1 A 1965 report discussed the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women. The subjects were residents of certainThe subjects were residents of certain regions of Great Britain, Norway, and Sweden.

Table 2.2.1: Temperature and Mortality Among Breast Cancer Patients (Great Britain, 1965)

Mortality	102.5	104.5	100.4	95.9	87	95	88.6	89.2	78.9	84.6	81.7	72.2	65.1	68.1	67.3	52.5
Temperature	51.3	49.9	50	49.2	48.5	47.8	47.3	45.1	46.3	42.1	44.2	43.5	42.3	40.2	31.8	34

Figure 2.2.2: Breast Cancer Mortality and Mean Annual Temperature



A simple regression of mortality index on temperature shows a strong positive relationship between the two variables. The data set contains a single outlier. The scatter plot shows the regression line for the entire data set and a separate regression line that excludes the outlier.

Fit a regression line, examine residuals and see if you can interpret the relationship. Is there any scientific basis for it or is it spurious relationship? You may have to use Google for help in answering the question.

E2.2.2 Here is real life data that can be used as example of chi-square tests. For each year, we compare mortality in two clinics. We can also treat this as a three way contingency table and see if things changed over time. Further we can use this table for calculation of odds ratios etc.

Table 2.2.2: Puerperal fever mortality rates for two clinics at the Vienna General Hospital 1841-1846.

	First clinic				Second clinic		
Year	Births	Deaths	Rate (%)		Births	Deaths	Rate (%)
1841	3,036	237	7.8		2,442	86	3.5
1842	3,287	518	15.8		2,659	202	7.6
1843	3,060	274	9		2,739	164	6
1844	3,157	260	8.2		2,956	68	2.3
1845	3,492	241	6.9		3,241	66	2
1846	4,010	459	11.4		3,754	105	2.8

a. Construct a 2X2 contingency table for each year, with clinic as the column variable and survive/death as rows. Remember that the number of deaths (of women) is given above. But # surviving is not given. What you have is the total number of women who gave birth
b. What is the hypothesis of interest? Which is the test suitable for that hypothesis?
c. Carry out the test.

E2.2.3 While describing the achievements of DrAbhay and Dr Rani Bang it was pointed out that "In the area served by the doctor couple, the (infant mortality) rate was about 120 in year 1988. It reduced linearly to about 30 by the year 2003". Produce estimates for all intermediate years

E2.2.4 Sample size and p value: Given below are two 2X2 contingency tables. Entries in the second table are multiples of corresponding entries in the first table. Carry out a test of independence in each case, comment on p values.

Table 1		**Table 2**	
5	10	50	100
15	20	150	200

Verify that proportions are the same in two tables but p value of the test of independence is over 0.5 in the first table and less than 0.05 in the second table. Interpret this result.

E2.2.5 We have read above how in the 19th century, doctors did not wash hands and caused infection to women giving birth. One amazing aspect of this hand wash story is the reluctance/refusal of the medical community to accept the truth and change behavior. But that was a century ago. What about modern times? In the data set from USA given below (Table 2.2.3), the response observed is number of times the nurse used gloves (column 2) out of the number of occasions on which she was observed, without her knowledge (column 3).

You have to formulate ideas on how to analyze these data. Formulate questions that seem relevant and can be answered from the data. Here are a couple of examples. (a) Do experienced

nurses ignore the matter of wearing gloves? (b) Does compliance level increase as time (since training) passes? Try different graphical presentations to bring out any trends in the data. The important point is that students have to decide themselves how to summarize data and how to plot it for good effect. Remember, you have to do two things: make up reasonable questions and then answer them after necessary analysis.

Table 2.2.3: Experience Level and Practice of Glove Using

Period	Observed	Gloves	Experience	Period	Observed	Gloves	Experience	Period	Observed	Gloves	Experience
1	2	1	15	3	3	2	8	4	5	5	2
2	7	6	15	4	4	2	8	1	1	0	5
3	1	1	15	1	4	4	9	1	1	0	15
1	2	1	2	2	4	4	9	2	3	3	15
2	6	5	2	1	4	0	9	1	1	1	3
3	11	10	2	2	4	4	9	2	2	2	3
4	9	9	2	3	2	1	9	3	1	1	3
1	5	5	3	1	2	0	15	4	1	1	3
2	13	13	3	2	3	2	15	1	1	0	14
3	8	7	3	3	1	1	15	4	1	1	14
4	15	14	3	4	2	1	15	2	2	2	14
1	2	0	10	1	6	1	8	3	3	3	14
2	2	2	10	2	1	1	8	4	1	1	14
3	2	2	10	3	2	2	8	2	1	1	8
4	5	4	10	1	3	0	8	3	1	1	8
1	12	0	20	2	4	3	8	4	1	1	8
2	2	2	20	3	8	6	8	2	2	1	3
3	3	3	20	4	2	0	8	2	1	1	6
4	3	0	20	1	2	0	2	2	1	1	3
1	3	0	8	2	3	3	2	3	2	2	1
2	8	8	8	3	8	8	2	4	1	0	6

Period: Observation period (1 = before intervention, 2 = one month after intervention, 3 = two months after, 4 = 5 months after intervention) **Observed:** Number of times the nurse was observed, **Gloves:** Number of times the nurse used gloves

Experience: Years of experience of nurse

E2.2.6 Here are monthly data on childbed deaths in the Vienna hospital.

a. Draw a time plot of % deaths.
b. Fit a linear regression and obtain the residual sum of squares.
c. Now fit separate regression lines for 3 phases. Phase 1 is the first 16 rows. This was the phase in which old practice (no hand wash) was in force. Phase 2 is rows 17-24. Here the hand wash policy was introduced. Phase 3 is rows 25-28. Here the hand wash policy was strictly enforced.
d. Compute the residual ss for the single regression fitted to all points and also compute sum of residual ss for each of 3 regression lines. Which value should be larger? Why?

Table 2.2.4: Childbed Deaths in Vienna Hospital					
Sr. No.	Year	Month	Births	Deaths	%deaths
1	1846	Jan	336	45	13.4
2		Feb	293	53	18.1
3		Mar	311	48	15.4
4		Apr	253	48	19
5		May	305	41	13.4
6		Jun	266	27	10.2
7		Jul	252	33	13.1
8		Aug	216	39	18.1
9		Sep	271	39	14.4
10		Oct	254	38	15
11		Nov	297	32	10.8
12		Dec	298	16	5.4
13	1847	Jan	311	10	3.2
14		Feb	312	6	1.9
15		Mar	305	11	3.6
16		Apr	312	57	18.3
17*		May	294	36	12.2
18		Jun	268	6	2.2
19		Jul	250	3	1.2
20		Aug	264	5	1.9
21		Sep	262	12	4.6
22		Okt	278	11	4
23		Nov	246	11	4.5
24		Dec	273	8	2.9
25**	1848	Jan	283	10	3.5
26		Feb	291	2	0.7
27		Mar	276	0	0
28		Apr	305	2	0.7

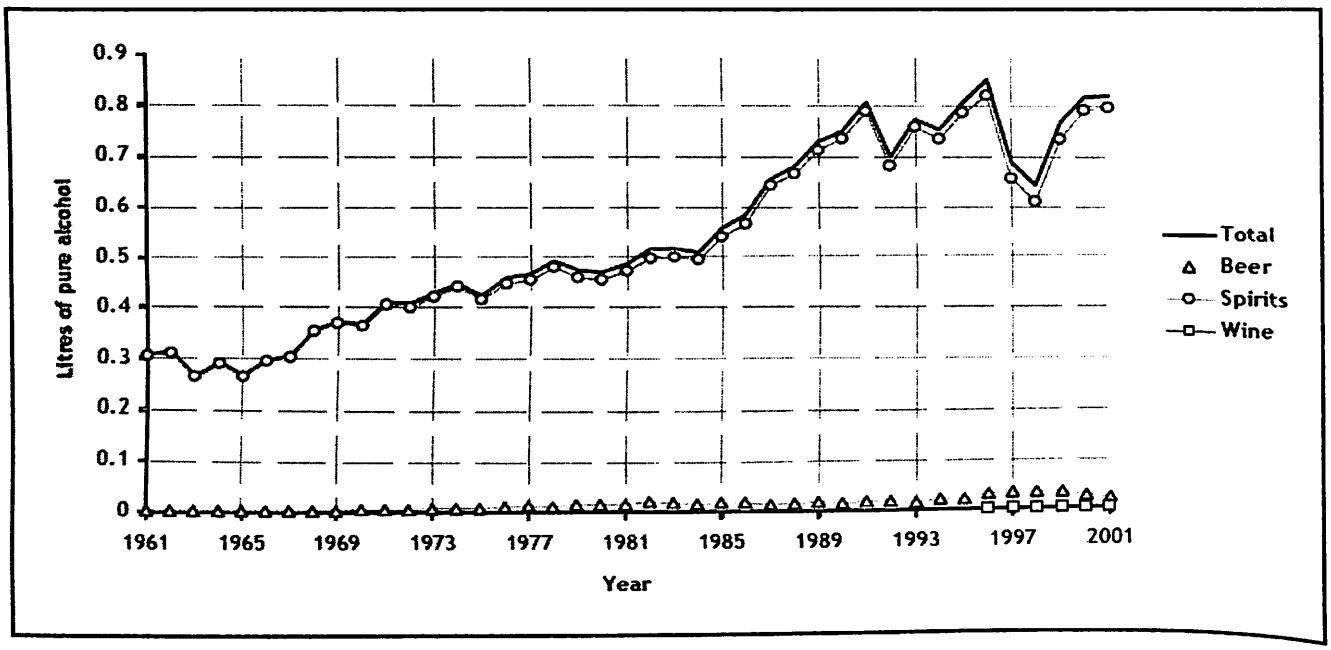
* Hand washing policy instituted in mid May

** Strict controls enforced on students' negligent hand wash

E2.2.7 Treat the above table as a before-after data set. We have to compare proportions and see if hand washing was useful in reducing death rate among women. Prepare relevant contingency tables and carry out tests.

E2.2.8 Given below is the trend in alcohol consumption (per year) in India. Fit a regression line, estimate slope and intercept and obtain projected value for 2010 (together with CI). Comment on the graph and your calculations.Use this information about a WHO survey in India and obtain CI for proportion of males that never drink alcohol and also for females.

Figure 2.2.3: Recorded adult per capita consumption (age 15+)



Sources: FAO (Food and Agriculture Organization of the United Nations), World Drink Trends 2003

According to the 2003 World Health Survey (total sample size $n = 9540$, males $n = 4605$ and females $n = 4935$; sample population aged 18 years and above), the rate of lifetime abstainers was 89.6% (total), 80.2% (males) and 98.4% (females).[2]

E2.2.9 Suppose the average number of new cases of a disease in a village is 1. We want a trigger (new case count) that is exceeded with a probability of no more than 0.05. What would be that score?

2.3 Some statistical issues and concepts

2.3.1 Correlation and causation

"Correlation does not imply causation" is a common theme for much of the epidemiological literature. Epidemiologists use gathered data and a broad range of biomedical and psychosocial theories in an iterative way to generate or expand theory, to test hypotheses, and to make educated, informed assertions about which relationships are causal, and about exactly how they are causal.

Hill criteria

In 1965 Austin Bradford Hill detailed criteria for assessing evidence of causation. These guidelines are sometimes referred to as the *Bradford-Hill criteria*. Hill said "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required sine qua non." So we have to take these ideas with caution.

1. Strength: A small association does not mean that there is not a causal effect.

2. Consistency: Consistent findings observed by different persons in different places with different samples strengthen the likelihood of an effect.

3. Specificity: Causation is likely if a very specific population at a specific site and disease shows association with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.

4. Temporality: The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay).

5. Biological gradient: Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence. In such a case, claim of causality is weakened.

6. Plausibility: A plausible mechanism between cause and effect is helpful (but Hill noted that knowledge of the mechanism is limited by current knowledge).

7. Coherence: Coherence between epidemiological and laboratory findings increases the likelihood of an effect. However, Hill noted that "... lack of such [laboratory] evidence cannot nullify the epidemiological affect on associations"

8. Experiment: "Occasionally it is possible to appeal to experimental evidence"

9. Analogy: The effect of similar factors may be considered

Today we definitely know that smoking causes cancer. But people had no idea of this when the epidemiological research on lung cancer began. Richard Doll, the first author of the Doll-Hill study was asked about this and he declared that discovery of the link between smoking and cancer came as a surprise. "I suspected that if we could find a cause it was more likely to have something to do with motor cars and tarring of roads." This quote shows the strength of the epidemiological and statistical methods. A totally unexpected cause was identified here which led to major long term gains in public health

The second phase and what may be called a great leap forward in tobacco control began in 1990s. In 1992, the *Journal of the American Medical Association* published a review of the available evidence regarding the relationship between secondhand smoke and heart disease, and estimated that passive smoking (smoke inhaled by non-smokers because they were near a smoker) was responsible for 35,000 to 40,000 deaths per year in the United States in the early 1980s (see the entry in wikipedia on passive smoking) . This shocked the society into serious action. Until then, while it was known that smoking has a serious adverse impact on health, there was still the escape that each person was free to do whatever she/he wanted with her/his own body and life. Now suddenly it seemed that a smoker caused harm not only to self but to all in

the vicinity. So it was no longer just a matter of personal freedom. Smoking in public places came to be recognized as an antisocial activity. Consequently, restrictions were imposed on it. In airplanes, buses, cinema halls and such crowded places, smoking was allowed only in limited and specified areas. Now those areas have reduced to zero. So, once again, we see that epidemiological findings have had a major impact on society.

The concept of relative risk was first introduced in the smoking and lung cancer work. It introduced emphasis on risk factors in diseases. Earlier there was emphasis on occupation and class. Now it shifted to behavior.

2.3.2 Risk and relative risk.

Risk is nothing but probability of some event. The word is generally used in the context of negative/undesirable events. We do not talk of risk of winning a prize. We do talk of risk of a heart attack or accident. To illustrate the terms, we shall use contingency table below about benefit from taking a pill of aspirin every day. There were 10,200 individuals who took a placebo (a pill without any active medicine) and of these, 200 experienced a heart attack. Another group of 10,100 people took one aspirin tablet each and of these, 100 suffered heart attack.(Do read relevant material in wikipedia about aspirin)

Table 2.2.5: Relative Risk of Heart Attack

Medicine	Risk of Heart Attack	Heart attack		
		Yes	No	Total
Aspirin	Π_1	100	10,000	10,100
Placebo	Π_2	200	10,000	10,200

The two groups of people can be regarded as independent samples from two Bernoulli populations. There are only two possible outcomes (getting a heart attack or being free from it). It is of interest to estimate the probability of getting a heart attack (risk) for each group. Obvious estimates of risk are the corresponding sample proportions p_1 (100/10100) and p_2 = (200/10200). Here the relative risk is the ratio of risks, $\Pi_1/\ \Pi_2$.

Notice that risk is always a number between zero and one, but relative risk can be any positive number. It can become very large if the denominator becomes very small (close to zero). It is estimated using the ratio of corresponding proportions namely p_1 and p_2 .In this case the estimated relative risk of heart attack for those who take aspirin as opposed to placebo is (100/10100)/ (200/10200). This is very nearly ½. So, if this data set is dependable, taking aspirin halves the risk of heart attack. Of course this is an estimate and will vary from sample to sample. It may be of interest to compute a confidence interval for the population relative risk. How do we do this? Here is one possible approach. Each sample proportion can be assumed to have an approximate normal distribution. Then we have a ratio of two normal random variables. We can use Fieller's theorem (to be studied later in the course). For the second possibility see exercise 8..

The attributable risk (AR) of a disease given an exposure is simply the rate of disease (incidence) in the exposed people minus the rate in the unexposed people. So the attributable risk for lung cancer in smokers is, in essence, simply the rate of lung cancer amongst smokers minus the rate of lung cancer amongst non-smokers. In fact AR shows which proportion of the disease in exposed subjects is due to exposure.

A **risk factor** is a variable associated with an increased risk of disease. Risk factors have a correlation with disease status but the relationship is not necessarily causal, because correlation does not imply causation. For example, being young cannot be said to cause measles, but young people are more at risk as they are less likely to have developed immunity during a previous epidemic.

Risk factors are evaluated by comparing the risk of those exposed to the potential risk factor to those not exposed. Let's say that at a wedding, 74 people ate the chicken and 22 of them were ill, while of the 35 people who had the fish or vegetarian meal only 2 were ill. Did the chicken make the people ill?

So the chicken eaters' risk = 22/74 = 0.297and non-chicken eaters' risk = 2/35 = 0.057.

Those who ate the chicken had a risk over five times as high as those who did not, suggesting that eating chicken was the cause of the illness. Note, however, that this is *not* a proof. Statistical methods would be used /needed in a less clear cut case. If a factor increases risk by a minuscule amount, we may choose to ignore that factor. If the impact is sizable, it deserves attention as a risk factor. Even then, no amount of statistical analysis could prove that the risk factor *causes* the disease; this could only be proven using direct methods such as a medical explanation of the disease's roots.

2.3.3 Odds and odds ratio

Let us now learn another pair of related terms. They are odds and odds ratio. In the context of the above 2X2 table, odds of getting a heart attack are defined by Odds(aspirin-group)=$\Pi_1/ (1- \Pi_1)$. Estimated odds are 100/ 10,000= 0.01. We could say that the odds in favor of heart attack are 1%. Odds ratio is a positive number. It could very large if the denominator is small. In the present case the fact that odds are so small suggests that not getting a heart attack is much more likely than getting it. Analogously,

Odds (placebo group) = $\Pi_2/ (1- \Pi_2)$. Estimated odds are 200/ 10,000= 0.02

Here too, odds are pretty small. Whether you take aspirin or not, chance of heart attack is small. What do we do if our interest is comparison of these two small probabilities? We use odds ratio (OR). For the aspirin table OR = $\theta = [\Pi_1/ (1- \Pi_1)]/[\Pi_2/ (1- \Pi_2)] = \Pi_1(1- \Pi_2)]/ \Pi_2 (1- \Pi_1)$. In view of this representation, odds ratio is sometimes called a cross product ratio.(The two way table has row 1 as (Π_1, Π_2) and row two as (1- Π_2), (1- Π_1))

	Response	
Groups	Yes	No
Aspirin	Π_1,	$1-\Pi_1$
placebo	Π_2	$1-\Pi_2$

Relative risk is a ratio of two probabilities while odds ratio is (as the name suggests) ratio of two odds. It can be estimated by use of sample proportions in place of unknown probabilities. So, $\hat{\theta}$ is $[p_1/(1-p_1)]/[p_2/(1-p_2)]$.Using the numbers we have $\hat{\theta} = 0.01/0.02 = ½$.

The quantity 'odds ratio' is used frequently in epidemiology and needs to be understood well. The **odds ratio** measures effect size . It is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. The groups can be an experimental group and a control group, or such dichotomous classification. If the probabilities of the event in each of the groups are p_1 (first group) and p_2 (second group), then the odds ratio is: $p_1(1-p_2)/p_2(1-p_1)$.

An odds ratio of 1 suggests that the event is equally likely in both groups. An odds ratio greater than 1 suggests that the event is more likely in the first group. And an odds ratio less than 1 suggests that the event is less likely in the first group. The odds ratio must be greater than or equal to zero. As the odds of the first group tend to zero, the odds ratio tends to zero. As the odds of the second group tends to zero, the odds ratio tends to infinity

We repeat that odds ratio is a positive number. Odds ratio $\theta = 1$ signifies that in the two groups being compared, the risks are equal. $\theta > 1$ implies that risk is higher in the first row. $0 < \theta < 1$ implies that risk is lower in the first group (in our case, aspirin group) .

The estimated odds ratio is of course a random variable and changes from sample to sample. Hence its standard error and such other properties are of interest. We shall not derive them. We will state them and use them. We need some notation to state the results.

	Response	
Groups	Yes	No
A	n_{11}	n_{12}
B	n_{21}	n_{22}

We have samples of sizes $n_{11}+n_{12}$ and $n_{21}+n_{22}$ from 2 groups. Estimated odds in favor of response 'yes' are n_{11}/n_{12} for group A and n_{21}/n_{22} for group B. Estimated odds ratio is $n_{11}*n_{22}/n_{12}*n_{21}$. If θ is the population odds ratio, then $\ln(\theta)$ is called the log odds. Logarithmic transformation is widely used because typically estimated odds ratio tends to have a very skewed

distribution. Logarithmic transformation makes a skew distribution approximately symmetric. $ln(\hat{\theta})$ has for large samples an approximately normal distribution with mean $\ln(\theta)$ and standard error $\{(1/n_{11})+(1/n_{12})+(1/n_{21})+(n_{22})\}^{0.5}$. Using this result an approximate confidence interval can be obtained for $\ln(\theta)$. It is given by $\ln(\hat{\theta})\pm Z_{\alpha/2}\ SE(\ln(\hat{\theta}))$. Once we have the confidence limits, they can be exponentiated to get limits for θ.

Let us apply this result to the aspirin table. The estimated odds ratio is ½. Its logarithm is -0.6932. The se of log odds is $(1/100+1/200+2/10000)^{0.5}$ which is 0.1233. Let α equal 0.05 (that means we are looking for a 95% CI). Hence $Z_{0.025}$ is to be used. It is 1.96. Hence the limits are -0.6932±1.96*0.1233. In other words the CI is (-0.94,-0.45). When exponentiated we get limits for θ namely (0.39,0.64).

Odds ratio and logit function:

Logit function of a probability π is given by $\text{logit}(\pi)=\ln[\pi/(1-\pi)]$.The related logistic regression model is very popular for connecting a dichotomous variable with a continuous variable. Dichotomous variable can be getting a heart attack (or not getting it) and continuous variable can be body weight. If $\pi(x)$ is the population probability of getting a heart attack when body weight is x kilogram, then the logit model is$\text{logit}(\pi(x))=\ln[\pi(x)/(1-\pi(x))]$.The logistic regression equation is$\ln[\pi(x)/(1-\pi(x))] = \alpha + \beta x$.It is important to understand why this equation is very useful and hence popular in applied statistics. If we wrote a regression equation for $\pi(x)$, say $\pi(x) = a + bx$, then the difficulty is that left hand side is in the restricted range (0,1) while right hand side can cross the range. The logit function extends the range of the left hand side to the entire real line. This facilitates use of a regression type equation. We shall not discuss fitting of this equation and related matters. Our interest is to recognize the connection with odds ratio. We exponentiate both sides and get $[\pi(x)/(1-\pi(x))] = \exp(\alpha + \beta x)$. Left hand side is nothing but the odds. Thus we have a model that expresses odds as an exponential function of x.
Further simplification yields $\pi(x) = \exp(\alpha + \beta x.\ )/[1+ \exp(\alpha + \beta x)$.

2.3.4 Case control studies

Inferences in epidemiology are based on heterogeneity among people. If all the people (sick as well as healthy) have identical background, then we can say very little about cause of disease. If some of them are smokers and others are not, then perhaps we can relate smoking to disease. This can be done in two ways. One way is looking forward and the other way is looking backward. If we start with a group of smokers and a group of non-smokers and keep track of who gets cancer and who does not, it is called a **prospective** study. If we take a bunch of patients and see how many of them are smokers and how many are non-smokers, it is called a **retrospective** study. In either case, our conclusions are based on comparisons.

Case control studies select subjects based on their disease status. The study population is comprised of individuals that are disease positive. The control group should come from the same population that gave rise to the cases. The case control study looks back through time at potential exposures both populations (cases and controls) may have encountered. A 2x2 table is

constructed, displaying exposed cases (A), the exposed controls (B), unexposed cases (C) and the unexposed controls (D). The statistic generated to measure association is the odds ratio (OR), which is the ratio of the odds of exposure in the cases (A/C) to the odds of exposure in the controls (B/D). This is equal to (A*D)/(B*C).

Exposure status	Disease status	
	Cases	Controls
Exposed	A	B
Unexposed	C	D

If the OR is clearly greater than 1, then the conclusion is "those with the disease are more likely to have been exposed," whereas if it is close to 1 then the exposure and disease do not seem to be associated. If the OR is far less than one, then this suggests that the exposure is a protective factor and helps you avoid the disease.Case control studies are usually faster and more cost effective than the other type namely cohort studies (discussed below), but are vulnerable to bias (such as recall bias and selection bias). The main challenge is to identify the appropriate control group; the distribution of exposure among the control group should be representative of the distribution in the population that gave rise to the cases. This can be achieved by drawing a random sample from the original population at risk. This has a consequence that the control group can contain people with the disease under study when the disease has a high attack rate in a population.

2.3.5 Cohort studies

Cohort studies select subjects based on their exposure status. The study subjects should be at risk of the outcome under investigation at the beginning of the cohort study; this usually means that they should be disease free when the cohort study starts. The cohort is followed through time to assess their later outcome status. An example of a cohort study would be the investigation of a cohort of smokers and non-smokers over time to estimate the incidence of lung cancer. The same 2x2 table is constructed as with the case control study. However, the point estimate generated is the Relative Risk (RR) [What is Relative Risk? How is it measured? How can values be interpreted? How is it linked to statistical analysis?

Exposure status	Disease Status		Total
	Case	Non case	
Exposed	A	B	(A+B)
Unexposed	C	D	(C+D)

Explanation is given earlier.], which is the incidence of disease in the exposed group (A/A+B) over the incidence in the unexposed (C/(C+D)). As with the OR, a RR greater than 1

shows association. Here the conclusion can be that "persons with exposure were more likely to develop disease."

Prospective studies have many benefits over case control studies. The RR is a more powerful effect measure than the OR. In a case control study, true incidence cannot be calculated since subjects are selected based on disease status. Temporality can be established in a prospective study, and confounders are more easily controlled for. However, such studies are more costly, and there is a greater chance of losing subjects to follow-up based on the long time period over which the cohort is followed.

2.3.6 Independence and symmetry in 2X2 contingency tables

So far we have frequently used 2X2 tables. We were interested in the hypothesis of homogeneity of two populations. As an example, we asked 'is the probability of heart attack the same among smokers and non-smokers?' The other hypothesis relevant in a 2X2 table is that of independence of two attributes.

Let us take one example. Suppose we check two attributes of 250 randomly selected residents of a village. First attribute is work place. They work either on farm or in a cement factory nearby. The other attribute is whether they have had any lung sickness experience. Results are given in the table below.

Table 2.3.1: Employment and Lung Sickness

Employment	Lung sicknessexperience		Total
	Yes	No	
Farm	15	135	150
Factory	35	65	100
Total	50	200	250

Our suspicion is that work in the factory makes lung sickness more likely. So, our null hypothesis is that the two attributes are independent. Let us draft a corresponding table with population quantities in it.

Employment	Lung sickness experience		Total
	Yes	No	
Farm	Π_{11}	Π_{12}	Π_{1+}
Factory	Π_{21}	Π_{22}	Π_{2+}
Total	Π_{+1}	Π_{+2}	1

Here Π_{11} is the probability that a randomly selected villager is a farm worker and has had some lung sickness. Π_{12} is the probability that a randomly selected person is a farm worker and has not had any lung sickness. Similarly for Π_{21} and Π_{22}. Now a + sign indicates sum over that subscript. So, Π_{1+} is the probability of a randomly selected person being a farm worker (we added the two values for sickness). The null hypothesis of independence is that the cell

probability is product of marginal probabilities, i.e. H_0: $\Pi_{ij} = [\Pi i_+]*[\Pi_{+j}]$. The chi-square test for this hypothesis is well known and we leave it to students to apply it to the lung sickness data.

Now we turn to the hypothesis of symmetry for a 2X2 table.Here is a data set relevant for this problem.Leprosy is a disease with considerable prejudice and stigma associated with it. Some public health experts were of the opinion that education can change attitudes towards leprosy. A film was prepared for a mass education program. A random sample of 100 individuals was asked whether they agree with the statement 'Leprosy is no different from other diseases and it can be cured'. Then the film was shown to the group and the same question was asked again. Two responses were recorded from each individual, one before watching the film and second, after watching it. Following table gives the outcome of this trial.

Table 2.3.2: Agreement (or otherwise) With Statement on Leprosy, Before and After Watching Film

After	Before	
	Agree	Disagree
Agree	33	22
Disagree	5	40

Here the diagonal cells indicate individuals who did not change their opinions. Off diagonal cells have frequencies of people who changed opinion. Clearly 5 people agreed with the statement before watching the film, but after watching they voted 'disagree'. 22 people changed their mind the other way. At first they did not believe in the statement. But once they saw the film, they were convinced. If the film is effective, it should persuade viewers that the statement is correct. Our interest here is not whether rows and columns are independent. Instead we want to know if watching the film made any difference. If not we should have $\Pi_{21} = \Pi_{12}$.If film was very effective, change from disagree to agree should be more than change from agree to disagree. In other words the alternative of interest is $\Pi_{12} > \Pi_{21}$.The test appropriate for this situation is called McNemar's test. The formula for the test statistic is $Z = [n_{12}-n_{21}]/ [n_{12}+n_{21}]^{0.5}$

This value is compared with a suitable cut off point of the standard normal distribution. For the leprosy data, we have $Z= (22-5)/(27)^{0.5} = 3.27$. Clearly, more people have changed from 'disagree' to 'agree' than otherwise. (What is the p value?). For a two sided test, we can use Z^2 and compare it with a suitable cut off point of a chi-square distribution with 1 degree of freedom.

The null hypothesis can alternatively be written as $\Pi_{1+} = \Pi_{+1}$. It automatically implies$\Pi_{2+} = \Pi_{+2}$. This can be paraphrased as 'equality of marginal distributions'. The two statements are equivalent. If the table has 3 or more rows (and columns) then the two hypotheses are not equivalent. (Verify this with a numerical example.) In fact two different tests have to be used.

2.3.7 Symmetry in rXr tables.

Now we can move on to larger square tables. Here is a classic data set on vision grades of two eyes of 7477 women factory workers. Grade 1 represents normal vision and Grade 4 is the weakest vision. The general interest is in any relation between the grade of left eye and right eye. Verify that the two attributes are not independent. Next point of curiosity is whether grades of two eyes of a person are the same. One answer is that all the off-diagonal cells represent cases in which the two eyes of a person had different grades. Obviously (?) grades of two eyes are not always the same. But can we put down some probability here?

Table 2.3.3: Vision Grades of Eyes of Women Workers

Right eye	Left eye			
	1	2	3	4
1	1520	266	124	66
2	234	1512	432	78
3	117	362	1772	205
4	36	82	179	492

So, we know that two eyes are not independent nor are they identical. What next? We ask if the 4X4 contingency table can be regarded as symmetric around the diagonal. We can write this hypothesis as $H_0 : \Pi_{ij} = \Pi_{ji}$, i,j = 1,2,3,4.This is called hypothesis of mirror image symmetry. It asserts that probabilities of cells below diagonal are equal to corresponding probabilities of cells above diagonal. This hypothesis can be tested using Bowker's procedure.

Bowker test.

The test statistic is $T = \Sigma\ (n_{ij}-n_{ji})^2/(\ n_{ij}+n_{ji})$, the sum being taken over i< j. The value is compared with a suitable cut off value of chi-square distribution with r(r-1)/2 degrees of freedom where r is the number of rows.Verify that in this case the value of T equals 19.10 and that the null hypothesis is rejected. In case of tables with more than 2 rows, there is one more way to view symmetry. It is equality of two marginal distributions. It can be stated as H_0: $\Pi_{i+} = \Pi_{+i}$ where i goes from 1 to r (the number of rows). It is sometimes called marginal symmetry. (Convince yourself that the two hypotheses are not equivalent if number of rows is more than 2) The test for this hypothesis is due to A. Stuart. We give below details of the test statistic.

Here is some notation needed. $d_i = n_{i+} - n_{+i}$. Here n_{i+} is the total count of cases in i^{th} row and n_{+i} is the total count of cases in the i^{th} column. Now we define a matrix V with diagonal elements $v_{ii} = n_{i+} + n_{+i} - 2n_{ij}$, $i = 1,2,\ldots,(r-1)$ and off-diagonal terms $v_{ij} = -(\ n_{ij} + n_{ji})$, $i \neq j$. Note that V is a square matrix with r-1 rows and as many columns with elements v_{ij}. Let V^{-1} denote the inverse of V. Elements of this matrix V^{-1} are denoted by v^{ij}. Using this notation we write Stuart's test statistic as

$$T = \sum_{i=1}^{r-1}\sum_{j=1}^{r-1} d_i d_j v^{ij}$$

We reject the null hypothesis if the statistics exceeds a suitable cut off point of the chi-square distribution with r-1 degrees of freedom. Since this test involves use of an inverse, students should learn how to find inverse using EXCEL.

2.3.8 Simpson's paradox.

What is a paradox? It is some sort of self contradiction. We will now see situations in which the same data set, viewed in two ways gives opposite conclusions.This is a real-life example from a medical study comparing the success rates of two treatments for kidney stones.The first table shows the overall success rates and numbers of patients for both treatments (where Treatment A includes all open procedures and Treatment B is percutaneous nephrolithotomy):

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

This seems to suggest that treatment B is more effective. If we include data about kidney stone size, however, the same set of treatments reveals a different answer:

Table 2.3.4: Kidney Stone Size and Treatment Effect

Stone Size	Treatment A	Treatment B
Small	Group 1 93% (81/87)	Group 2 87%(234/270)
Large	Group 3 73% (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	83% (289/350

The information about stone size has reversed our conclusion about the effectiveness of each treatment. Now treatment A is seen to be more effective in both cases. In this example, we did not know, to begin with, that the lurking variable (or confounding variable) of stone size plays any important role. Later when it is included a surprising pattern emerges.Which treatment is considered better is determined by an inequality between two ratios (successes/total). The reversal of the inequality between the ratios, which creates Simpson's paradox, happens because two effects occur together:

1. The sizes of the groups, which are combined when the lurking variable is ignored, are very different. Perhaps doctors tend to give the severe cases (large stones) better treatment (A), and milder cases (small stones) inferior treatment (B). Therefore, totals are dominated by groups 3 and 2, and not by the two much smaller groups 1 and 4.

2. The lurking variable has a large effect on the ratios, i.e. the success rate is more strongly influenced by the severity of the case than by the choice of treatment. Therefore, the group of patients with large stones using treatment A (group 3) does worse than the group with small stones, even if the latter used the inferior treatment B (group 2).

3. To summarize, treatment B appears to be better (when data are aggregated) because it is used mainly for milder cases. For this group, success rate is higher whatever the treatment. This fact dominates the aggregated data.

Let us take one more example to convince readers that there is no magic trick. The paradox shows itself whenever certain conditions are fulfilled.

Two trials for comparing two treatments for an illness: In trial No. 1, treatment A cures 20% of its cases (40 out of 200) and treatment B cures 15% of its cases (30 out of 200). In trial No. 2, treatment A cures 85% of its cases (85 out of 100) and treatment B cures 75% of its cases (300 out of 400). So, in two trials, treatment A scored 20% and 85%. Also in two trials, treatment B scored only 15% and 75%. No matter how many people were in those trials, treatment A (at 20% and 85%) is surely better than treatment B (at 15% and 75%), right? Wrong! That is if you ignore the distinction between two trials. Treatment B performed better in total. It cured 330 (300+30) out of the 600 (200+400) cases, in which it was tried, a success rate of 55%. By contrast, treatment A cured 125 (40+85) out of the 300 cases (200+100), in which it was tried, a success rate of only about 42%.

Table 2.3.5: Data from Two Trials, Separate and Pooled

Trial	Treatment	#Cured	#Not cured	Total	%cured	Better treatment?
1	A	40	160	200	20	A
	B	30	170	200	15	
2	A	85	15	100	85	A
	B	300	100	400	75	
Combined	A	125	175	300	41.67	B
	B	330	270	600	55	

The take home lesson is that comparisons in aggregate can be misleading. It is safer to separate data into meaningful subgroups first and then compute relevant ratios in each category.

2.3.9 Mathematical models

In the early 20th century, mathematical methods were introduced into epidemiology by Ronald Ross, Anderson Gray McKendrick and others. The earliest account of mathematical modeling of spread of disease was carried out in 1766 by Daniel Bernoulli. Trained as a physician, Bernoulli created a mathematical model to defend the practice of inoculating against smallpox (Hethcote, 2000). The calculations from this model showed that universal inoculation against smallpox would increase the **life expectancy from 26 years 7 months to 29 years 9 months** (Bernoulli & Blower, 2004). [We want the students to pause and appreciate these numbers. People died young in most cases. Today, because of advances in science, life expectancy has doubled even in poor countries like India.]

Smallpox has long been an interesting topic for mathematicians to model, primarily because of the huge epidemics that occurred in Europe during the eighteenth century. There was significant disagreement in these earlier times about whether the risks of death from inoculation were worth the benefits of immunity from smallpox. Daniel Bernoulli attempted to study this question using a mathematical model.

Bernoulli's model was very simple. It involved a healthy population, an infected population, and an immune population. Based on observed evidence, Bernoulli estimated the rates at which people became infected and the chances that these infected people could possibly become well on their own.

He then considered the effects of inoculation. What would happen if the healthy population were allowed to become immune to the smallpox virus? This would have to consider the fact that inoculating people could result in possibly killing them. By using his model, Bernoulli calculated that the fatality rate of inoculation must be no more than 1 out of 200 before the technique should be used. Often of course, the rate was much higher and Bernoulli's work sparked controversy

Following Bernoulli, others contributed to modern mathematical epidemiology. Among the most acclaimed of these were A. G. McKendrick and W. O. Kermack, whose paper *A Contribution to the Mathematical Theory of Epidemics* was published in 1927. A simple deterministic (compartmental) model was formulated in this paper. The model was successful in predicting the behavior of outbreaks very similar to that observed in many recorded epidemics (Brauer& Castillo-Chavez, 2001).

By compartments we mean division of the population into homogeneous subgroups. In the simplest such model we split the population into S, I and R groups. S stands for susceptible. 'I' stands for infected and R stands for recovered or removed.

- S(t) is used to represent the number of individuals not yet infected with the disease at time t, or those susceptible to the disease

- I(t) denotes the number of individuals who have been infected with the disease and are capable of spreading the disease to those in the susceptible category

- R(t) is the compartment used for those individuals who have been infected and then recovered from the disease. Those in this category are not able to be infected again or to transmit the infectiontoothers.

The flow of this model may be considered as $S \rightarrow I \rightarrow R$. Using a fixed population, $N = S(t) + I(t) + R(t)$, Kermack and McKendrick derived the following equations:

1.$dS/dT = -\beta SI$. This implies Change in S is due to some susceptible cases getting infected. So, the sign of the change is expected to be negative. In other words, β is assumed positive. There

are S*I pairs with one person infected and the other susceptible. Out of all these interactions, a fraction result in making the susceptible person infected.

2.$dI/dT = \beta SI - \gamma I$. This means The infected group increases to the extent that some members of the susceptible group get infected. That explains the first term on the right side. Now the infected group also loses some members since they recover. The fraction that recovers is γ.

3.$dR/dT = \gamma I$. Lastly, the recovered group increases to the extent that some infected cases recover.

In the kind of situation being considered, two outcomes are possible. One is that disease simply dies out (this is the preferred outcome) or an epidemic occurs. Which of these will occur depends on the parameters of the system. If the population is free of infection (I=R=0) and then one single member gets infected (perhaps acquires infection from a different place during a journey etc), the process begins. If rate of recovery is more than the rate at which new cases are generated, then infection dies out. Otherwise it proliferates. Suppose recovery rate γ is half. In other words, the first infected individual recovers in about two units of time. He has to create at least his own replacement. So, roughly speaking the key quantity is $R_0 = (\beta S)/\gamma$. If this is less than 1, epidemic peters out. Otherwise it expands.

This value quantifies the transmission potential of a disease. If the basic reproduction number falls below one ($R_0 < 1$), i.e. the infective may not pass the infection on during the infectious period, the infection dies out. If R_0 exceeds 1, there is an epidemic. In cases where $R_0 = 1$, the disease becomes endemic, meaning the disease remains in the population at a consistent rate, as one infected individual transmits the disease to one susceptible

We can of course ask how the process will flow. For this we need to know the path that S, I and R will follow over time. Such graphs can be drawn using numerical solutions of the equations. This matter is beyond the scope of the book. We will only note that the system of equations does not have analytical solution but numerical solutions can be obtained. There are websites that provide such numerical solutions of differential equations. We will give address of one such site.

A website http://math.colgate.edu/~wweckesser/solver/DiseaseSIR.shtml,solves SIR models mathematically. It gives numerical solution of SIR model for given choice of parametric values.

2.3.10 Exercises

E 2.3.1 Show that the inverse relation between probability and odds is Π=odds/(1+odds)

E 2.3.2 Prove that the logit function $\ln[\pi(x)/(1-\pi(x))]$ has a range from minus infinity to plus infinity.

E2.3.3 Study the data in table 2.3.1. Test the hypothesis that employment and lung sickness are independent.

E2.3.4 Prove that in case of 2X2 contingency tables, the null hypothesis of marginal symmetry namely $\Pi_{1+} = \Pi_{+1}$ is equivalent to the hypothesis mirror image symmetry. Further prove that two statements are NOT equivalent if the table has 3 or more rows (and columns).

E2.3.5 Verify for the table 2.3.3 on vision grades of two eyes of the same person, that grades of two eyes are not independent.

a. You have a dichotomy among workers- those with same grade for two eyes and others. Now use a Bernoulli distribution and find a confidence interval for the probability that two eyes have the same grade
b. Apply Bowker's test to check if the vision data show mirror image symmetry.
c. Apply Stuart's test to check if the vision data exhibit marginal symmetry.

2.3.6 Data are recorded on pain sensitivity of sibling pairs. Out of 95 brother-sister pairs, each individual was classified as oversensitive or normal or robust. The counts are given in the table below.

Table 2.3.6: Pain Sensitivity of Sibling Pairs

Sister	Brother			
	Oversensitive	Normal	Robust	Total
Oversensitive	12	12	14	38
Normal	9	25	11	45
Robust	5	3	4	12
Total	26	40	29	95

a. Test independence of rows and columns.
b. Apply Bowker's test of mirror image symmetry.

c. Apply Stuart's test of marginal symmetry.

E2.3.7 Smoking and Cancer: Data summarizes a study of men in 25 occupational groups in England. Two indices are presented for each occupational group. The smoking index is the ratio of the average number of cigarettes smoked per day by men in the particular occupational group to the average number of cigarettes smoked per day by all men. The mortality index is the ratio of the rate of deaths from lung cancer among men in the particular occupational group to the rate of deaths from lung cancer among all men.

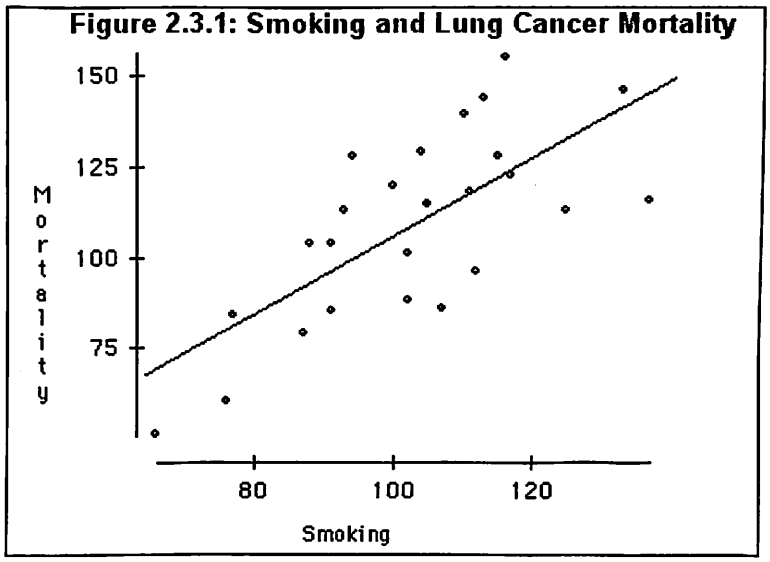


Table 2.3.7: Smoking Index and Lung Cancer Mortality

Occupational Group	Smoking	Mortality
Farmers, foresters, and fisherman	77	84
Miners and quarrymen	137	116
Gas, coke and chemical makers	117	123
Glass and ceramics makers	94	128
Furnace, forge, foundry workers	116	155
Electrical and electronics workers	102	101
Engineering and allied trades	111	118
Woodworkers	93	113
Leather workers	88	104
Textile workers	102	88
Clothing workers	91	104
Food, drink, and tobacco workers	104	129
Paper and printing workers	107	86
Makers of other products	112	96
Construction workers	113	144
Painters and decorators	110	139
Drivers of stationary engines, cranes, etc	125	113
Laborers not included elsewhere	133	146
Transport and communications workers	115	128
Warehousemen, storekeepers, packers	105	115
Clerical workers	87	79
Sales workers	91	85
Service, sport, and recreation workers	100	120
Administrators and managers	76	60
Professionals, technical workers, and artists	66	51

A scatter-plot of the data shows a moderately strong linear association, with a correlation coefficient of 0.716. Residuals from a regression of mortality on smoking are randomly scattered with no outliers or influential observations.'

E2.3.8 Following data on use of beds in hospitals is to be used to give exercises to students.The data were collected by the Department of Health and Social Services of the Government and cover 52 licensed nursing facilities. Answer the following questions:

a. Is the number of inpatient days per bed similar in urban and rural hospitals?
b. Are number of beds similar in rural and urban hospitals?
c. Is revenue per patient similar?
d. Is revenue per patient day similar?
e. Are nursing salaries similar?
f. Is expenditure per patient similar?
g. Is expenditure per patient day similar?
h. Any outliers in the data?

Table 2.3.8: Data on Nursing Facilities in Rural and Urban Area

BED	MCDAYS	TDAYS	PCREV	NSAL	FEXP	RURAL	BED	MCDAYS	TDAYS	PCREV	NSAL	FEXP	RURAL
244	128	385	23521	5230	5334	0	64	91	214	8782	4729	4446	1
59	155	203	9160	2459	493	1	62	146	204	8951	2367	1064	0
120	281	392	21900	6304	6115	0	108	255	366	17446	5933	2987	1
120	291	419	22354	6590	6346	0	62	144	220	6164	2782	411	1
120	238	363	17421	5362	6225	0	90	151	286	2853	4651	4197	0
65	180	234	10531	3622	449	1	146	100	375	21334	6857	1198	0
120	306	372	22147	4406	4998	1	62	174	189	8082	2143	1209	1
90	214	305	14025	4173	966	1	30	54	88	3948	3025	137	1
96	155	169	8812	1955	1260	0	79	213	278	11649	2905	1279	0
120	133	188	11729	3224	6442	1	44	127	158	7850	1498	1273	1
62	148	192	8896	2409	1236	0	120	208	423	29035	6236	3524	0
120	274	426	20987	2066	3360	1	100	255	300	17532	3547	2561	1
116	154	321	17655	5946	4231	0	49	110	177	8197	2810	3874	1
59	120	164	7085	1925	1280	1	123	208	336	22555	6059	6402	1
80	261	284	13089	4166	1123	1	82	114	136	8459	1995	1911	1
120	338	375	21453	5257	5206	1	58	166	205	10412	2245	1122	1
80	77	133	7790	1988	4443	1	110	228	323	16661	4029	3893	1
100	204	318	18309	4156	4585	1	62	183	222	12406	2784	2212	1
60	97	213	8872	1914	1675	1	86	62	200	11312	3720	2959	1
110	178	280	17881	5173	5686	1	102	326	355	14499	3866	3006	1
120	232	336	17004	4630	907	0	135	157	471	24274	7485	1344	0
135	316	442	23829	7489	3351	0	78	154	203	9327	3672	1242	1
59	163	191	9424	2051	1756	1	83	224	390	12362	3995	1484	1
60	96	202	12474	3803	2123	0	60	48	213	10644	2820	1154	0
25	74	83	4078	2008	4531	1	54	119	144	7556	2088	245	1
221	514	776	36029	1288	2543	1	120	217	327	20182	4432	6274	0

BED = number of beds in home, **MCDAYS** = annual medical in-patient days (hundreds) ,
TDAYS = annual total patient days (hundreds), **PCREV** = annual total patient care revenue ($hundreds),
NSAL = annual nursing salaries ($hundreds) , **FEXP** = annual facilities expenditures ($hundreds),
RURAL = rural (1) and non-rural (0) homes

E 2.3.9 CI for relative risk: Show that when sample proportions are small, relative risk is approximately equal to odds ratio. Obtain a 95% CI for population relative risk for the table 2.2.5 on aspirin and heart attack.

E2.3 10 CI for odds ratio: In 1854, in the Vienna General Hospital Clinic I, 237 (out of 3036) mothers died in childbirth. In Clinic II the count of dead mothers was 86 (out of 2442)

Prepare a contingency table. Calculate odds ratio and comment on it. Obtain a 99% confidence interval for the population odds ratio.

E2.3.11 Simpson's paradox: Consider these data on on-time performance for two airlines, Alaska Airlines and America West.It is interesting that at each airport Alaska Airlines has a lower percent delayed than America West but overall America West has a lower percent delayed. What can explain this discrepancy?

Table 2.3.9: Performance of Two Airlines

Alaska Airlines					America West			
Airport	# On-time	# Delayed	% Delayed		Airport	# On-time	#Delayed	% Delayed
Los Angeles	497	62	11.10%		Los Angeles	694	117	14.40%
Phoenix	221	12	5.20%		Phoenix	4840	415	7.90%
San Diego	212	20	8.60%		San Diego	383	65	14.50%
San Francisco	503	102	16.90%		San Francisco	320	129	28.70%
Seattle	1841	305	14.10%		Seattle	201	61	23.30%
Total	3274	501	13.30%		Total	6438	787	10.90%

E2.3.12 Examine the following table for any possible paradox.

Table 2.3.10: Recovery Rate by Gender

Gender	Males	#Recovered	No recovery	Total	%recovered	Better treatment?
Males	Drug T	18	12	30	60	B
	Placebo	7	3	10	70	
Females	Drug T	2	8	10	20	B
	Placebo	9	21	30	30	
Combined	Drug T	20	20	40	50	A
	Placebo	16	24	40	40	

Chapter 3

Statistical Aspects of Clinical Trials

3.1 What is a clinical trial? Sickness is something that each of us undergoes at one time or the other. Then we visit a doctor, get some medicine and all is well once again. Where do these tablets, capsules, syrups come from? How does one discover that a particular drug will give relief from a particular disease? How is this confirmed? Discovery can occur in many ways. But confirmation is generally through trials. Trials of drugs on people are called clinical trials. That is the topic of this third part of the book. We will introduce many new terms here. These terms are typical to the field of clinical trials. We will explain the terms in suitable places. However, if a new term is encountered, students should read about it on the internet using Google, Wikipedia etc. Familiarity with these terms will be an important part of the course. Anyone seeking a job as a statistician in a pharmaceutical company or **CRO** will get considerable advantage in a job interview from familiarity with these terms. Just now you have seen a new term 'CRO'. What does it mean? Google tells you that it is a short form for 'contract research organization'. Wikipedia tells you that 'A Contract Research Organization (CRO) is a service organization that provides support services to the pharmaceutical/biotech industry'. These services are of different types. We are mainly concerned with services to be provided by statisticians. They mainly involve study design, analysis of data recorded and preparation of reports based on the bio-statistical analysis.

Since clinical trials are by the nature of the topic a very interdisciplinary area, it is inevitable that statisticians participating in such trials will have to learn about the medical aspect and participating doctors will have to learn statistical aspects. Together they evaluate the trial, write reports, and work on preparation of materials to be submitted to the Food and Drug Administration (abbreviated as FDA). What is **FDA**? It is the department of the Government of USA that supervises work of pharmaceutical companies and gives permission to sell a product (in USA) as a drug. Its website is very informative. Incidentally, we will mainly pursue 'western' or allopathic medicine in the sense that clinical trials have evolved mainly in the context of western medicine. The principles we learn will be relevant to any form of medicine including Ayurved, Homeopathy etc. (Have you come across the acronym AYUSH? It stands for Ayurved, Unani, Siddha and Homeopathy systems of medicine). Why are we discussing FDA, an American government agency? This is because of two reasons. FDA was a pioneer in developing a systematic approach to regulation of drug research. Hence methods devised by FDA are of interest to all countries. Second reason is economic. USA is one of the largest markets for drugs and also the sponsor of many clinical trials. Hence statisticians in this field are very likely to have to work on a clinical trial aimed at FDA. Of course we will refer to practices in other parts of the world and also in India. You can find dozens or at most hundreds of trials about herbal medicines and alternative therapies listed on the website www.clinicaltrials.gov along with more than 80 thousand trials on allopathic medicines. So, the burden of numbers is what makes us focus on western medicines.

Randomized controlled trials

In most cases to be discussed here, we will examine an RCT (randomized controlled trial). The two adjectives of the word trial are very statistical in nature and quite important. Every student must make sure that the meaning of these adjectives is clear. Let us understand the nature of an RCT step by step.

The first step in a trial naturally is choice of the drug to be tested. Next step is selection of participants in the trial. Criteria for selection of persons (for example, in terms of age, sex, diagnosis, etc) have to be spelt out. The participants have to be representative of the target population in whom we wish to generalize the results. [As an example, for comparison of two treatments for rheumatoid arthritis we chose hospital patients. This may not be right for less severe range of the disease seen commonly.] A rigid rule is that trials have to be on volunteers only. Subjects who satisfy the inclusion criteria must give consent to participation. Further, it must be informed consent. The person must clearly understand what s/he is agreeing to. Sometimes people decline. So, whoever agrees has to be accepted. Then the worry is how far the volunteers that remain can be considered representative of the target population. They might, for example, be younger on average than the ones who refused. (Along with inclusion criteria, there are exclusion criteria also. Any group that is particularly vulnerable to adverse effects is excluded (e.g. pregnant or lactating women). Sometimes a drug for high blood pressure is to be tested but patients should not have any other condition such as diabetes. In some cases the target population is senior citizens. Then younger age group is excluded etc.

Those subjects who agree to participate are then 'randomized' to the treatments under comparison. In other words, the treatment to be given to each subject is decided using an agreed random process. This can mean using published tables of random numbers, or using random numbers generated by computer. What if subjects enter the study sequentially (for instance, as they are admitted to hospital)? Then randomization is often done in blocks or groups. Thus to compare two treatments, A and B, subjects might be randomized in blocks of six. Of the first six patients, three would be given treatment A, and the other three given treatment B - which patient received which treatment being determined randomly. The story then repeats for the next six patients. The advantage of this method is that there is never any large imbalance among different treatments, which otherwise could occasionally occur by chance. It also ensures that the balance between the different treatments is roughly constant throughout the course of the study, thus reducing the chance for confounding by extraneous variables that change over time. We will return to these matters again later.

Students should recall that the main purpose of randomization is averaging out effects of unknown factors. If we notice some feature of the participant's condition that might affect response, we can first form groups depending on that condition. For example, in acute *myocardial infarction* (MI) the presence of certain *dysrythmias* may suggest something about chance of recovery. It is then a *prognostic marker*. Randomization ensures that such markers will tend to be well distributed among the different treatment groups. Sometimes outcome of a treatment is likely to be influenced by other aspects of a patient's management. In such case it is better for people managing the patients to be "*blinded*" to which treatment has been allocated. (Please look up the words in italics on the web).

If this rapid overview is not very helpful, move on without anxiety. We will discuss all these aspects in detail later.

3.2 Some diseases and discoveries: Medicine is a field that is practiced in all societies and at all times. However, in this study we will focus on modern development in medicine in the western world (i.e. Europe and North America). This is not to suggest that western world does not study other medicine systems. In fact clinical trials about efficacy of traditional medicines are very much in vogue.

We have already seen some cases of medical discovery while discussing epidemiology. We know that Florence Nightingale discovered that clean environment saves people from many diseases. John Snow discovered that cholera was connected with polluted water and preventing use of such water had an impact on (controlling) the cholera epidemic. Ignaz Semmelweis discovered that washing hands before assisting a woman in labor could prevent her getting infection. Let us see some more examples.

3.2.1 Small pox was a terrible disease that killed countless number of people in all continents and maimed many more. Edward Jenner, a country doctor in England, discovered in 1796 a way to protect his patients from small pox. He observed that among the villagers, those who worked with cows and got cow-pox (a much milder disease) never got small pox. [Students need to note that we have a very statistical piece of evidence here. It could easily have gone into any modern discussion on epidemiology. See related exercise 2.] He speculated that cow-pox material may be responsible for small pox prevention. He decided to try this theory out on a person. He came across a young woman suffering from cow-pox. He extracted some fluid from her sores. Then he approached a farmer with a proposal. He would introduce the cow-pox material into the body of the farmer's son. If the theory were correct, the son would never get small pox. The farmer agreed to this experiment. Jenner made two small cuts on the hand of the young man and introduced cow-pox material. The young man became mildly sick and recovered in a few weeks. Now Jenner introduced small pox material into the body of the young man. He did not get the disease. This confirmed Jenner's idea. In Latin cow-pox is called vaccinia. Hence Jenner used the word vaccination to describe his procedure. The word vaccine is now widely used to describe any material that renders immunity against a specific disease (and not just small pox).

This can be regarded as a clinical trial. It was an experiment regarding a medicine on a human subject. There are many features of this experiment, which make it rather inappropriate from a modern perspective. Firstly, it is not clear whether the young man whose body became the experimental subject, knew what was going on. His father allowed the experiment. Today clinicians have to obtain consent of the subjects themselves. Secondly, it has to be informed consent. Subjects must know what they are walking into and what risks they run. Lastly, ethical considerations are also involved. Even if subjects give consent, it may still be inappropriate to expose them to unreasonable risks. But who is to decide this? An experimenter is likely to be biased. So, an ethics committee has to be constituted and approval of this committee has to be taken. That is the modern way.

However, in the case of Jenner, the trial did lead to a great discovery. The second major limitation was sample of size 1. Also there is no comparator.

The general idea emerging from Jenner's work is that a weak form of disease agent can make a person immune to the stronger and more virulent disease. Louis Pasteur used that idea to develop a vaccine against rabies. This was a disease without any cure. People were desperate. Pasteur infected rabbits with saliva of rabid dogs. Then he dried the spinal cords of infected rabbits. This was the source of vaccine. In 1885 a nine-year boy bitten several times by a rabid dog was brought to Pasteur. Everyone was sure that the boy will die a miserable death in a few months. Hence people were ready to try anything. Pasteur treated the boy with this rabbit based vaccine. The boy survived. This led to manufacture of rabies vaccine in France. Humanity is obliged to Pasteur for this discovery. But he was not a doctor. Today, laws in many countries are against non-medical people like Pasteur treating patients. So, today he would be prosecuted for what he did without proper license or eligibility! Of course when there are extenuating circumstances, some of the stringent rules for conduct of clinical trials can be relaxed. In critical cases such as advanced cancer, treatments can be tried without any prior safety study on healthy people. A high level of toxicity may be regarded as acceptable etc.

3.2.2 Scurvy is a disease resulting from a deficiency of vitamin C. The chemical name for vitamin C, ascorbic acid, is derived from the Latin name of scurvy, scorbutus. Scurvy leads to the formation of spots on the skin, spongy gums, and bleeding. The spots are most abundant on the thighs and legs, and a person with the ailment looks pale, feels depressed, and is partially immobilized. In advanced scurvy there are open wounds and loss of teeth.

Scurvy was at one time common among sailors, pirates and others aboard ships at sea much longer than perishable fruits and vegetables could be stored, and among soldiers similarly separated from these foods for extended periods. It was described even in ancient times, by Hippocrates (c. 460 BC–c. 380 BC). Herbal cures for scurvy have been known in many native cultures. In 1536, the French explorer Jacques Cartier, exploring the St. Lawrence River in Canada, used the local natives' knowledge to save his men who were dying of scurvy. He boiled needles of the tree Eastern White Cedar to make a tea.(It was later shown to contain 50 mg of vitamin C per 100 grams) Such treatments were not available aboard ships, where the disease was most common.

Lord Anson was the First Lord of the Admiralty (Chief of British Navy), who as a commodore had sailed round the world in 1740 and knew all about the ravages of this disease. Indeed, of the 961 sailors manning his six ships, 626 were dead from scurvy by the time the fleet reached the Juan Fernandez Islands.

James Lind was a surgeon's mate in the Royal Navy. He spent nine years voyaging in the Mediterranean, off West Africa, and in the West Indies. In those days ships were cold, damp, and unwholesome, while the food consisted of putrid beef, rancid pork, and moldy biscuit and foul water. During these years, Lind carefully recorded all his observations, as his later writings show. By 1747 he had been promoted surgeon to HMS Salisbury, and it was

during her cruise in the English Channel that year that there was a severe outbreak of scurvy and he was able to carry out his classic experiments on its treatment.

"On the 20th of May 1747, I selected twelve patients in the scurvy, on board the Salisbury at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of the knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all, viz. water gruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times light puddings, boiled biscuit with sugar, etc., and for supper, barley and raisins, rice and currants, sago and wine or the like.(*Now comes the description of different treatments compared*) Two were ordered each a quart of cyder a day. Two others took twenty-five drops of elixir vitriol three times a day ... Two others took two spoonfuls of vinegar three times a day ... Two of the worst patients were put on a course of sea-water ... Two others had each two oranges and one lemon given them every day ... The two remaining patients, took ... an electary recommended by a hospital surgeon ... The consequence was, that the most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them, being at the end of six days fit for duty ... The other was the best recovered of any in his condition; and ... was appointed to attend the rest of the sick. Next to the oranges, I thought the cyder had the best effects ..."

We notice some interesting features of this study. There is a clear objective, namely finding a cure for scurvy. The author plans comparison of several treatments. He selects subjects that are similar (and all suffering from scurvy). They are subject to similar conditions and the only difference among them is the treatment. Luckily for the doctor, the signal is loud and clear. The small sample size (two per treatment) does not hamper the conclusion.

It is remarkable that James Lind did not get carried away by his own success. He cautioned and said that more evidence should be collected.
" ... but, though a few partial facts and observations may, for a little, flatter with hopes of greater success, yet more enlarged experience must ever evince the fallacy of all positive assertions in the healing art." This remark is of great interest to statisticians. It suggests that a small sample size renders the conclusion rather less convincing. We know that probabilities of two kinds of errors go up if sample size is small. What kind of error would James Lind have made? Findings of his trial suggest that oranges and lemon help in recovery from scurvy. So, he was essentially rejecting the null hypothesis of no difference among treatments. Hence, if an error occurred, it must be of type I, wrongly concluding that oranges and lemons are good for cure of scurvy. Now this possibility suggests that he should have promptly distributed all available oranges among those sick with scurvy. A dramatic recovery of all (if observed) would have given a much stronger support for the claim.

Sir James Lancaster, a non-medical naval captain carried out what may be regarded as the first clinical trial in 1605. He gave two spoonful of lemon juice daily to each sailor on his flagship and none to any other sailor (on 3 other ships under his command). There were no deaths on his ship Dragon and there was 45% mortality on the other 3 ships. So there is some dispute about which is the first clinical trial, this one or that by Lind. For us it suffices to say

that the concept of a trial was already around a couple of centuries ago. [*We recommend that as a co-curricular activity, each student (or group of students) should be asked to read about one specific disease and give a presentation in class.]*

So these are some historical examples of trials. They were very useful. But the methodology was not as well developed as in modern times. In the present times, trials are designed using principles of statistics.

3.2.3 Malaria, polio, asthma: Ronald Ross showed the way to control malaria. He discovered in 1911 in India, that parasites that cause malaria are transmitted to humans by anopheles mosquitoes. This discovery led the way to malaria control measures. Now that the culprit was known, it was just a matter of finding an antidote for it. In 1938 DDT was found to be effective against disease causing insects including mosquitoes. In many parts of the world, where malaria had made life impossible, mosquito control using DDT dramatically reduced the disease and large areas became habitable. This happened around the middle of the twentieth century. Extensive use of DDT caused a sharp decline in malaria in India as well. Regions such as terai at the foothills of Himalaya could be brought under cultivation. Hence the population of such areas increased rapidly. However, in a few decades, malaria returned! Mosquitoes had become resistant to DDT. (Here are some figures. In 1952 there were 75 million cases of malaria and 0.8 million deaths. In 1965, there were only 0.1 million cases and no deaths. In 1976 with new resurgence, there were 6.5 million cases and 59 deaths. Since 1983, the number of cases is between 2 and 3 million in a year. Source- http://mohfw.nic.in/mspnew.pdf)

Jonas Salk worked on the problem of preventing the dreaded polio, which caused severe disabilities in many children. He used formaldehyde to obtain killed polio virus that was still able to trigger the right immune response from human body. (What is immune response?) The vaccine was first tested on monkeys, then on patients and then on human volunteers including he and his family. In 1954 a massive trial was launched involving 2 million children in the age group 6 to 9 years. It was one of the first double blind placebo controlled trials. This is now the norm for clinical trials. (What is double blind? What is placebo controlled?)

In the USA, 6000 children died of polio in 1916 and over 25000 were paralyzed. In 1952, the number of recorded cases exceeded 50,000. But after the vaccine was brought in use, the number of cases came down by over 85% in two years. A phenomenal achievement!

Asthma is a chronic disease in which a patient experiences breathing difficulties, shortness of breath etc. Prevalence of this disease is high. A common medication for relief is a bronchodilator. Strong medicines, especially when in use for a long period of time, can have significant adverse side effects. About half the patients take up some form of alternative treatment. Such treatments also have to be examined with clinical trials. One treatment that has shown some benefit is the so called Buteyko method. This method devised by a Russian doctor involved regular and extended practice of breathing exercises. These are similar to Pranayam in yogic practices. Homeopathic treatment is also found to have mild benefit. On

the other hand acupuncture, body massage, dust reduction etc are treatments that have generally failed to demonstrate effect. So you can see how clinical trials can help us fight chronic diseases just as much as they help with acute diseases. (Learn more about these terms using internet).

3.2.4 Penicillin and serendipity: Alexander Fleming discovered antibiotics. It happened partly by accident. (Such accidental discovery is sometimes called serendipity. Read more about this word. It has an interesting connection with Sri Lanka). He had some glass plates for culturing bacteria (what is culturing?), which were unused for a while and some fungus had grown on the plates. Fleming noticed that there was a zone around the fungus infected area. This zone was free of the pathogenic bacteria under study. It turned out that the fungus produced an agent (later named penicillin because the fungus was of the genus penicillium), which had anti-bacterial effect. This was in 1928. It took another 12 years for all the difficulties to be worked out and in 1940 the material was available in a stable form. This is when clinical trials began on safety and efficacy of the antibiotics. [Some other examples of serendipitous discoveries:

- Role of pancreas in glucose metabolism, by Oskar Minkowski. Dogs that had their pancreas removed for an unrelated physiological investigation urinated profusely; the urine also attracted flies, signaling its high glucose content.
- Chemical synthesis of urea, by Friedrich Woehler. He was attempting to produce ammonium cyanate by mixing potassium cyanate and ammonium chloride and got urea, the first organic chemical to be synthesized, often called the 'Last Nail' in the coffin of the vitalist theory (as opposed to mechanist theory). What are these two theories? Please look up on the internet.)]

3.2.5 Exercises

E3.2.1. Web search: Search the website www.clinicaltrials.gov and find out the number of trials reported for each of Ayurved, Unani, Siddha, Homeopathy and Chinese traditional medicine

E3.2.2. Smallpox and cowpox: "Edward Jenner, a country doctor in England, discovered in 1796 a way to protect his patients from small pox. He observed that among the villagers, those who worked with cows and got cow-pox (a much milder disease) never got small pox." Try to guess what the data looked like. Write a null hypothesis expressing the idea. Use your imagination and cook up a plausible data set. Notice the word 'never' in the quote. Could it be replaced with 'rarely'? When will the data fail to convince you about the presumed relationship between cowpox and small pox? Perhaps this question is too vague. If so, that is not by mistake. Real life situations faced by medical statisticians are rarely cut and dry. They are very often unclear and a statistician has to sort them out. (Hint: Fisher's exact test)

3.3 Principles of Design of Experiments: At this point we remind students of things they have learnt in a course on design of experiments (and a clinical trial is an experiment mainly for comparing treatments). R. A. Fisher, the founding father of modern statistics, laid down the principles of DOE in 1930s. They are replication, randomization and local control. Let us recall these ideas briefly.

3.3.1 Replication: There should be enough observations in an experiment. Replicates are independent observations taken under identical conditions. If observations are not independent, they are not replicates (could be called **pseudo-replicates**). Suppose in a study of growth rate of infants, we take repeat measurements on the same subject every day for a week. Do I have 7 replicates? Not really! Growth in a day or two is not even measurable. So, we end up measuring the same number. These 7 measurements may at best tell us something about accuracy of our measurement. To study growth we should measure height of the same child at time points that are adequately separated (perhaps one year). Difference between initial and final measurement will give us growth. We should measure growth of different infants. That will be replication. Genuine replication serves two purposes. One is that power of any test based on the data, goes up, second is that good estimate of variability of the system (σ^2) can be obtained. We will spend some time later in the course, discussing calculation of sample size needed.

3.3.2 Local control: This ensures that experimental units in a group are similar and treatments are compared on a 'level playing field'. This is easily done by dividing the experimental units into blocks. Typically, in clinical trials, blocking is done by gender, ethnicity etc. This is because response to a medicine is likely to be affected by these features of a subject. A good experimental design will ensure that different treatments are tried on similar subjects so that any difference in response is directly attributed to difference in treatments and not to differences among subjects.

3.3.3 Randomization: Homogeneity among subjects is mainly achieved through blocking. This is good enough for class differences (gender, profession, ethnic group etc). There are two situations where this kind of local control is not adequate. One is a situation in which differences among subjects are in a continuum. Parameters such as age, blood pressure, blood sugar are examples of such differences among subjects. Here it is nearly impossible to have blocks that are homogeneous. In such a case, we adjust for differences by using ANCOVA. Second difficulty which cannot be taken care of using local control alone, is one in which we do not know about differences among subjects. If some unknown feature is likely to bias our results, how can we protect against it? This is where randomization helps. Within a block, we allocate treatments to subjects in a random manner. This ensures that any unknown factor gets 'averaged out'. (Let us put it formally. We need to compare two treatments. There are 2n subjects in a block and they are randomly divided into two groups of size n each. Consider an unobservable random variable X. The 2n subjects have values X_i, i=1,2...,2n. Randomization will ensure that expected value of X will be the same for each of the two groups.

Incidentally, regulatory authorities make it mandatory to allocate treatments to subjects in a random manner. So, it is important not only to use randomization but also to

save evidence of it. Such evidence is in the form of a set of random numbers that were generated and used for treatment allocation.

To repeat, when treatment allocation is randomly determined, any variable that might get mixed up with treatment will, on an average, be balanced and thus its effect will be canceled out. Please remember the caveat about the protection against confounding being only on average. If the sample size is large, then the same may hold true even for a single randomization exercise. If on the other hand the sample size is small, there may be some difficulty. Strongest point in favor of randomization is that it provides some protection against unsuspected confounding effects.

In practice, one common situation is that the experimenter has data on some covariate and wants to ensure balance with respect to it. In case of small samples one efficient way is to divide 2n subjects into n pairs such that covariate values are similar within a pair. Last step is to randomly assign one member of the pair to each treatment.

This discussion may give the impression that randomization is the norm and universally accepted method of treatment allocation. This is not quite true. There is a long-standing and often bitter controversy about randomized trials. Is it feasible? Is it ethically right? Opponents of randomization are mainly surgeons and those who practice alternative unconventional therapies. Supporters of randomized studies are practitioners of internal medicine. Of course biostatisticians are supporters of randomization as are the regulatory agencies. In 1978 Rimm and Bortin [41] remarked that the randomized trial is not only a ritual but has all elements of a religion - they called it TRIALISM -, with gods, devils and 10 commandments the first of which is: *Thou shall randomize.*

Consider the following statement: If treatment assignment is randomized, then the background attributes for the different treatment groups will be roughly equivalent and therefore any differences observed between treatments can be linked to the treatment effect and is not a characteristic of the individuals in the group. The logic behind this statement is very strong. We will work on the assumption that a trial is always suitably randomized. Let us now review briefly the mechanics of randomization.

3.3.4 Fixed allocation randomization: In this scheme, probability of allocating a patient to a particular treatment group is pre assigned, usually equal. The method is simple and easy to implement. Consider a situation where participants are to be allocated to one of the three groups, Treatment 1, Treatment 2 and a control. We can draw a random number (R) from uniform distribution over (0, 1). If this number is less than $1/3^{rd}$, the subject will be allocated to Treatment 1. If R is between $1/3^{rd}$ and $2/3^{rd}$, then Treatment 2 is selected for the subject. Otherwise subject will be assigned to control group. Following table shows an illustration of such schedule for 15 patients. Second column gives the random numbers generated and third column gives the treatment selected using the above rule. Please check if all entries in the last column are correct.

Thus 5 patients are allotted to each treatment. In other words exactly five random numbers came out below one third and five above two third.

Table 3.3.1 Random Allocation of Treatments					
SUB. NO.	R	TREATMENT	SUB. NO.	.R	TREATMENT
1	0.188	T1	9	0.308	T1
2	0.502	T2	10	0.262	T1
3	0.466	T2	11	0.159	T1
4	0.587	T2	12	0.349	T2
5	0.896	C	13	0.742	C
6	0.89	C	14	0.681	C
7	0.498	T2	15	0.816	C
8	0.255	T1			

However, this may not happen every time. So the method is not flawless. A simple modification of the allocation rule can take care of this shortcoming. Identify 5 smallest random numbers and associated individuals get treatment 1. Analogously, 5 largest random numbers send the subjects to control. Remaining 5 cases go to Treatment 2. Are you convinced that it is easy to modify the allocation rule to ensure equal allocation? But that is not the only problem in randomization. You will notice one more peculiar feature in this particular schedule. You can often see the same treatment assigned to successive cases. It may make an observer feel uncomfortable. Could this really be a random arrangement? You wonder. How come the last three cases all go to control? But remember, randomness is not the property of the outcome. It is the property of the process. One has to accept this as a random allocation if the process of generating those numbers is random. If it is desired that allocation of the same treatment to multiple subjects in succession should be avoided, then we should put more restrictions on how we do the exercise. Here a scheme of permuted block randomization is more appropriate.

3.3.5 Permuted Block Randomization: In this method, randomization is within a block. Suppose four treatments are under study. Then each group of 4 patients can be considered as a block and treatments are allocated randomly within this block. In this scheme, the rank within set of four successive random numbers decides the treatment. Such a scheme ensures that all four treatments are allotted equally frequently, at least for blocks that are complete. Repeated successive occurrence of the same treatment is eliminated. [But there is one negative feature. Treatment of last person in each block is predictable. To avoid this predictability, blocking factor can be varied. So blocking can be randomly made in groups of size 4, 8, 12 etc.] The table below shows an example of block randomization. There are 16 subjects to be allocated to 4 treatments. So we have a column of 16 random numbers from U(0,1). Each set of 4 numbers is considered at a time. In the first such set, largest number is 0.9895 and the associated subject gets 4^{th} treatment namely C. Smallest random number is 0.0816. It gets rank 1 and the associated subject gets the first treatment T1 etc. Please check if all entries in the last column are correct.

In this example proportion of cases getting placebo is the same as that of any active treatment. This is not always so. It seems unethical to deny some people a good treatment and to give them a sugar pill instead. This cannot be eliminated altogether. But the proportion can be reduced. So, often a 3:1 ratio is practiced. In other words, 6 subjects get active treatment while 2 get a placebo. (What is a placebo?)

Table 3.3.2: Block randomization with 4 treatments							
Sr.	R	Rank	Tr.	Sr.No.	R	Rank	Tr
1	0.0816	1	T1	9	0.5158	3	T3
2	0.9895	4	C	10	0.9895	4	C
3	0.8786	3	T3	11	0.2152	2	T2
4	0.7392	2	T2	12	0.1525	1	T1
5	0.4873	2	T3	13	0.9592	4	C
6	0.5202	3	T2	14	0.8511	3	T3
7	0.6294	4	C	15	0.2796	1	T1
8	0.1106	1	T1	16	0.701	2	T2

3.3.6 Stratified and adaptive randomization: If the subject population can be divided into groups that are known to be similar (in terms of proneness to disease etc) then, there should be randomization within each such group. Thus gender is a possible grouping factor. [Ethnic background or profession may also be relevant.] Treatments should be randomized within each gender separately. Or in a multi center trial, each center can be considered as a block and allocation is then randomized within each center.

Sometimes, planners of clinical trials wish to take into account success or failure of a treatment while deciding allocation at a point. A treatment that seems to be more successful is given to more subjects. This is the so-called 'play the winner' rule. This is a scheme in which allocation probability changes as the study progresses. Why would that be? Because it increases the chance of a patient getting the better treatment! One way of implementing such a scheme is using urn models. If we have to compare two treatments A (control) and B (active), we identify treatment A with white balls and treatment B with red balls. We start with one white and one red ball in the urn. When a new patient arrives, we draw one ball from the urn. If it is white, the patient gets treatment A etc. So, initially both treatments are equally probable. But now, (assuming that the situation is suitable for such assessment) we check if the treatment helped the patient. If treatment A was given and patient felt better, we add one white ball to the urn increasing the count from one to two. (If on the other hand, treatment A did not help the patient, we add a red ball to the urn).

Table 3.3.3: Response Adaptive Randomization							
Patient Number	R	P(W)	P(RED)	UL on R for allocating control (White)	Treatment allocated	Outcome	Ball added
1	0.6142	0.5	0.5	0.5	Active	No change	White
2	0.2391	11/21	10/21	11/21	Control	No change	Red
3	0.7418	11/22	11/22	0.5	Active	improvement	Red
4	0.4100	11/23	12/23	11/23	Control	No change	Red
5	0.6285	11/24	13/24	11/24	Active	improvement	Red
6	0.7379	11/25	14/25	11/25	Active	No change	White
7	0.0936	12/26	14/26	12/26	Control	No change	Red
8	0.5237	12/27	15/27	12/27	Active	improvement	Red
9	0.2092	12/28	16/28	12/28	Control	No change	Red
10	0.3089	12/29	17/29	12/29	Control	improvement	white
11	0.6832	13/30	17/30	13/30	Active	No change	White
12	0.6064	14/31	17/31	14/31	control	No change	White

When the next patient comes, we draw a ball randomly. Here probability of getting a white ball has gone up to 2/3. If one treatment succeeds repeatedly, then it accumulates more balls in the urn. Suppose there are 10 balls of each color to begin with. We have to allocate a total of 12 subjects to two treatments randomly. To simulate this exercise, instead of actually drawing balls from an urn, we will use a set of 12 random numbers from U (0, 1). We will carry out a thought experiment of treatment allocation. Details of this exercise are given in the table below (for first 12 patients).

Draw 1: Probability of drawing a ball of either color is the same. Hence if the random number selected is less than 0.5 we presume that a White ball is drawn (allocate the subject to Control). Otherwise it is presumed that a RED ball is drawn (allocate the subject to ACTIVE treatment). In the present case the first number is 0.6142. That is equivalent to drawing a red ball. So patient will be allocated to active treatment group. The ball is replaced. Now, if the patient gets better, a red ball is added (increasing the count of red balls) and a white ball is added if the active treatment fails to show improvement. In case of failure of active treatment, a white ball is added and consequently probability of drawing a red ball at the next draw is 10/21 and that of drawing white ball is 11/21, slightly more than 1/2.

Draw 2: Second random number is 0.2391. It is less than 11/21, which means a white ball is drawn and hence second patient will be allocated to control group. White ball is replaced and a red ball is added because the treatment given did not lead to improvement.. Now probability of drawing a white ball is half again.

It would be useful to work out entries in the table. It is easy, but one can miss a step. Please verify the table.

As an exercise, you have to rework this table. You are given a set of random numbers to decide the outcome in each row of the above table. If the random number is above 0.5, call it 'improvement', otherwise call it 'no change'. Now work through all the rows and generate a new table.

Table 3.3.4: Response Adaptive Randomization (Exercise)					
0.55167	0.77564	0.19249	0.09276	0.31587	0.56429
0.26265	0.74405	0.33993	0.58285	0.96962	0.34766

How were these random numbers generated? There are many ways. One is to use a table of random numbers. That is the old way. Now we can use software. In particular EXCEL generates random numbers from U(0,1) (function RAND). Alternatively, you can search for a generator (which is nothing but a computer program that uses a mathematical function) on the internet. One website that does it for you is http://www.random.org.

Note that it gives you one random integer between user specified limits. Thus, you can get a number between 0 and 100 (say) and you can put a decimal point before the number to get a random number from U(0, 1). There are websites that help in random assignment of treatments to subjects e.g. http://www.randomizer.org. This site has been used over 7.5 million times. Another one is- http://www.graphpad.com/quickcalcs/randomize1.cfm

3.3.7 Blinding: In addition to these classical ideas on how to do comparative experiments, we need to know some concepts that are especially relevant to clinical trials. In any randomized trial the comparison of treatments may get distorted if the patient and those responsible for evaluation know which treatment is being used. An open label trial is one in which patient and doctor; both know which drug is being given. A *blind trial* is one where the patient does not know whether (s)he is receiving the active drug or a placebo. A *double blind trial* is one where neither patient nor clinician knows which treatment is being given. Blinding is introduced to eliminate any bias arising from knowledge about treatment. Sometimes blinding is not possible. One example is massage and such other physical therapy, use of infra red light etc. Here it is not possible to hide the treatment from the patient or the researcher. Another example would be particular diet. Suppose we study effect of different sources of protein in diet on some response. One source is meat, the other is eggs and the third is pulses. Here too, the participant will instantly know what he/she is subjected to. In such a case, blinding is not possible. Blinding is very important especially when the response is opinion of the subject. [Perhaps this word should not be used when explaining the trial to patients with eye problems!]

In 1784, King of France appointed a Royal Commission headed by Benjamin Franklin and Antoine Lavoisier for checking claims of Anton Mesmer about his ability to cure sickness. Mesmer claimed that he could hypnotize patients and then cure them of many diseases. The commission designed a series of ingenious experiments to check the claims. In these experiments, patients were blindfolded so that they did not see what Mesmer was doing, nor was there any eye contact. (Later this became a model for a controlled clinical trial.) The results showed that mesmerism was ineffective when participants were blind to the treatment condition.

As a result of this work Mesmer became notorious. No one believed anything he said. In fact 'hypnosis' was not all fraud and later in the 19^{th} century, an English surgeon showed that he could do many operations on patients without anesthesia when he used hypnosis. Unfortunately no one would believe him.

A control is a treatment that is useful as a standard for comparison. Why is there a need for such a treatment? An example can bring out the point easily. A man taking a walk in forest with a friend recites a mantra. His friend asks him the purpose of this action. The man explains that reciting the mantra keeps the tiger away. The friend is impressed. He watches out for a tiger and none appears on the scene. The claim that 'reciting a mantra keeps the tiger away' seems to be true. But then, a doubt creeps in his mind. So, he takes a walk in the forest alone and (since he does not know the mantra) skips the mantra. No tiger this time either! So, is the mantra doing anything? Perhaps not!

A king standing on a beach commands the sea to go back and it does. Then he orders the sea to return and waves return. Courtiers are impressed by this remarkable show of obedience by the sea. The skeptic murmurs, "The Sea does the same even when the King is silent or even absent from the sea shore".

We hope the message is clear. It is better not to believe in a cause-effect claim unless there is comparative data. In clinical trials different kinds of controls are used. A historical control is a set of observations from an earlier study for similar subjects and for similar purpose. Another is called an active control. In this case an alternative medication is used for comparison. Yet another is placebo control. Here the treatment is very similar to the one under study, except that the active drug is excluded. It is just a sugar pill. Think of an aspirin tablet for headache. It is round, white, and the size of a paisa. Placebo will also be a white round tablet that looks like the true aspirin tablet, smells like one etc. But it has no aspirin in it. Why should we take the trouble to create a placebo like this? The reason has to be understood clearly.

Sometimes a patient gets relief when treated, but the supposedly active compound is not responsible for the relief. Just the fact of getting attention and rest, being under loving and tender care of the family is itself enough for the person to feel better. If such is the case, we will wrongly attribute merit to the drug. This has to be avoided. Getting relief even when the pill contains no active compound is called a placebo effect. What would be the placebo for an antibiotic injection? It will have to be an injection, though not of an antibiotic but just distilled water. Some patients are very keen to get an injection and may insist on one even when the doctor thinks that it is not called for. If he gives a vitamin injection and patient reports good effect, we know it is placebo effect. In short, placebo is included as a treatment in a clinical trial to make sure that we do not give undeserved credit to a drug or treatment.

3.3.8 Exercises

E3.3.1 Use of replication: Replication is useful for estimation of error variance.

a) Consider an experiment with 4 observations each on 2 treatments. We assume that all 8 experimental units are homogeneous. Write down a formula for estimate of σ^2, the common unknown variance. Show that the estimate is unbiased.
b) Now extend this to one-way ANOVA. We have 3 treatments and there are 4 observations on each. Obtain an estimator for the common unknown variance and show that it is unbiased. This has to be done without reference to any linear model.

c) In the next extension, we have 3 treatments and 4 blocks. So there are altogether 12 observations. But there is no replication. Two observations on the same treatment belong to two different blocks and hence do not have the same distribution. Now, if you assume a linear model, you can still obtain an unbiased estimator of the common variance. Derive the estimator and show that it is unbiased.

Remember, this question is nothing but some theory you have learnt, but it is posed in a slightly different way.

d) It is clear that in case of one-way ANOVA, we can easily write down an unbiased estimator of variance based on any one sample and then all such estimators can be averaged to get one overall estimator. But there seem to be two possibilities. We can take each unbiased estimator and take arithmetic mean of them. If we have k samples, it will be mean of k estimates. The second possibility is the so-called pooled estimate. Here we add up

corrected sums of squares from different samples and then divide by n-k where n is the pooled sample size. Show that the two approaches give identical result if all k samples are of equal size. But the two estimators are different if sample sizes are not all equal. In that case, which estimator is better? Since this question is about two estimators, both unbiased, we should compare their variances. Since we all use the pooled estimator, it should be better than the other one. Compare the variances of the two estimators of variance.

E3.3.2. Randomization:

a) Effect of randomization

Prove the following assertion about the effect of randomization: We need to compare two treatments. There are 2n subjects in a block and they are randomly divided into two groups of size n each. Consider an unobservable random variable X. The 2n subjects have values X_i, i=1,2...,2n. Randomization will ensure that expected value of X will be the same for each of the two groups.

b) Simulation to study the effect of randomization:

Purpose of randomization is to eliminate bias due to unknown causes. Effect of randomization should be to make different groups [getting different treatments] similar with respect to unknown features. So, here is a small simulation exercise to illustrate this aspect. We have to compare two treatments for a digestive disorder. There are 20 subjects to be randomized to two groups. Subjects are otherwise similar. You have to generate a set of 20 random numbers from U(0,1) and associate them with the twenty subjects. Now rank the random numbers and the smallest 10 random numbers get the first treatment. The key question is whether the two groups so constituted are balanced with respect to any unknown trait.

Table 3.3.5: Randomization and Checking Balance

Ser.No.	Sys. BP	Ser.No.	Sys. BP	Ser.No.	Sys. BP	Ser.No.	Sys. BP
1	107	6	111	11	123	16	114
2	110	7	107	12	109	17	119
3	123	8	112	13	112	18	112
4	129	9	136	14	102	19	110
5	112	10	102	15	98	20	106

We have given values of systolic blood pressure (in mmHg). When you divide the 20 subjects into 2 groups, you can calculate mean bp for each group and difference between two groups. Now repeat the exercise 100 times and store values of group averages and their differences. Summarize these and examine whether two groups turn out to be similar on average.

E 3.3.3 Study of a randomization rule. We have 20 subjects and they are to be randomized to two treatments. Here is our rule of randomization: Draw 20 numbers from U (0,1). If a number is below 0.5, the corresponding subject gets assigned to group 1, otherwise to group 2. Estimate through simulations the probability that fewer than 10 subjects get assigned to group 1. Calculate the theoretical probability.

E 3.3.4 Block randomization This refers to example of permuted block randomization (section 3.3.5). There are 16 subjects and 4 treatments (C, T1, T2, T3). Hence blocks of size 4 were used. Now for illustration, assume that there are 5 treatments and block size has to be 5. Obtain 16 random numbers from U (0, 1) and carry out permuted block randomization. What is the impact on the last block?

E3.3.5 Urn models for clinical trials

a) Urn problem 1. We have two urns. One contains 10 white balls and 5 black balls. Second urn contains 4 white and 8 black balls. From each urn one ball is drawn. From this pair of balls, one is selected randomly. What is the probability that the selected ball is white?
(Hint- add probabilities of three mutually exclusive events)

b) Urn problem 2- An urn has 1 white and 1 black ball. A ball is randomly selected, its color observed. If it is white, we replace the ball and also add one more white ball. If it is black, we stop. Find the probability that we see n white balls in succession.

c) Urn problem 3- An urn contains 1 white and 1 black ball (representing 2 treatments). We draw one ball randomly, observe it, replace it and then add one ball of the other color (increasing the probability of the other treatment being assigned to the next patient). Find the probability that two balls drawn in succession are of two different colors.

E3.3.6. You have 5 treatments to compare. Use the following 15 random numbers and decide allocation for 15 subjects. Check if the allocation is balanced. If not adopt a block randomization approach and rework the allocation. Verify that the new set up is indeed balanced.

Table 3.3.6 Randomization		
0.77824	0.77378	0.52939
0.45346	0.13768	0.03829
0.403	0.78649	0.99025
0.61497	0.55466	0.01376
0.70558	0.91665	0.12456

3.4 Phases of clinical trials: We hope a student does not get the impression that there is just one experiment on a drug and then a decision about its usefulness is made. That is far from reality. In practice, many experiments (trials) are needed before a final judgment can be passed on a drug. Three or four dozen experiments is a realistic number. And these are carried out only after a promising drug is identified, which itself is a major task. The table below gives an outline of different stages through which the drug research process goes. Scientists begin with a large class of candidate compounds, which are regarded as potentially useful for a certain disease. These compounds are first tested on animals (mice, rabbits, dogs, monkeys etc) to check safety. Healthy human volunteers come next. If the drug is found to be safe for healthy humans, then it is tried tentatively on patients and suitable dose levels are

ascertained. After this, usefulness of the drug is checked on a large number of patients. If this testing also comes out well, then the entire work is submitted to the regulatory authority for approval. If approved, the drug goes on sale. Even after approval by regulator and commencement of market sale, monitoring continues. Reports from users and doctors are archived regularly. If reports about serious adverse effects (AE) of the drug come up, a fresh inquiry may be instituted. All in all, the process is very long and very expensive. Cost of the process, when carried out in USA, has been estimated at around US $ 800 million.

Table 3.4.1: Stages in Drug Development Process

Stage of development	Duration in years	Number of compounds
Discovery and preclinical testing	6.5	5000
Phase I , safety, PK /PD /Tolerability	1.5	10
Phase II, proof of concept in patients, safety, PK-PD, tolerability, Dose selection	2	5
Phase III, efficacy and safety in patients	3.5	2
FDA review	1.5	1
Phase IV	indefinite	1

In order to persuade companies to invest such resources, there has to be a pot of gold at the end of the effort. It is the chance of huge sales and profits without competition. American and other governments grant patent rights to the companies whose drug gets approved. It means that no one can produce and sell that drug for a specified number of years (say about 20). Of course patent application for a drug is filed before trials start. In effect, trial time is cut out of the total protection period. Hence a drug is under patent protection approximately for 7 to 12 years. So, companies hope to recover all their cost and earn good profit during this period. Once this period is over, the drug is up for grabs. In other words, anyone can (with approval from regulators) produce and sell the drug. When other companies produce 'copies' of this drug that was under patent protection earlier, the products are called generic drugs. These have to have the same active ingredients as original drug. Further they have to be 'bio-equivalent' to the original drug. (We shall study this aspect later in this course.) Producers of generic drugs do not incur the heavy burden of cost of discovery and early phase trials. They do have to prove bio-equivalence, but that is much cheaper. Further, people already know the patented drug and hence marketing cost is also lower. So, once generic drugs come into the market, price / cost of the drug goes down significantly for patient. [Some Indian companies have achieved considerable success in the field of generic drugs.]

If a company is interested in getting approval for its drug, it has to register with the regulator all information about the clinical trials. Each trial (which is nothing but an experiment) has to be registered with the regulator before it is begun. Why should this be so? What is the need for registering each experiment? Why bother about failed experiments? Why not just record successes and give approval? The reason is very statistical in nature. Consider the following scinnerio: *a company that has a drug known to be useless. Suppose the company starts trying it on patients. Then it tests significance of the difference between the test drug and a control. The test fails to reject the null hypothesis of no difference. So,*

what does the company do? Does it just discard the drug? No. it conducts one more experiment. If the drug fails to show improvement over control, the test is repeated. This goes on till in one experiment the drug shows statistically significant improvement over placebo. So, happily, the company applies to the regulator with results of this last trial as evidence.

Remember that a statistical test of the null hypothesis that the drug is no better than control and using a significance level of say 5% will give a false rejection of the null hypothesis in about 5% of the times. So, if the company shows patience, sooner or later it will be rewarded with a statistically significant result. Please remember that such an experiment is of little scientific value unless it is put in the right context of all the previous experiments in which the null hypothesis could not be rejected.

So, what have we learnt? The regulator must know about ALL trials, not just the successful ones! You can see lists of such registered trials in progress in different parts of the world. The website http://clinicaltrials.gov/ is one such place for registering clinical trials. It is from USA. Similar registry in India has the website
http://www.ctri.in:8080/Clinicaltrials/trials_jsp/index.jsp
Please visit these websites and discuss what you find there.

We have seen many examples of diseases studied and drugs discovered for treating/preventing the diseases. Now we can get into the details of how a newly proposed drug may be tried, to check whether it is useful or not.

We saw that Edward Jenner tried the exudates from the blisters of a cow-pox patient on another person. Louis Pasteur gave his treatment to a boy that had been bitten by rabid dog. Is this how we do it in today's world? The answer is 'no'. Today's clinical trial goes through several phases (stages).

3.4.1: Pre-clinical study: It all begins with tests on animals. Drug is given to experimental animals (rats, rabbits, guinea pigs, monkeys etc) and reactions are observed. Depending on the drug under trial, scientists may look for different kinds of responses (sometimes called endpoints). Is the blood pressure normal or has it gone up? Is the movement normal or is the animal unable to walk? These are immediate responses. Does the pregnant female treated with the drug, produce normal pups or are they deformed? Of course the ultimate adverse effect is death. A drug that causes major adverse effects is not suitable for use in humans. There may be an attempt to study tolerance. In this case, dose of the drug is increased gradually till a certain adverse reaction is obtained. If the reaction is death, then the dose that caused it is called a lethal dose. In toxicity studies people look for median lethal dose (dose that will kill half the target population) or some other quantile of the dose distribution. That would be relevant if we are studying a pesticide. In drugs we would check if a given dose leads to any unacceptable response. If it does, that dose is not suitable.

Let us suppose our finding from animal study is that a certain range of doses of the drug under study is acceptable. Then we have to extend the study to human beings. This is called a phase 1 study.

3.4.2: Phase I study: Here the aim is to check safety of the drug. At this point we hold back the question of whether the drug in fact cures the disease. That is left to a later study. How do we decide which dose is to be tried for checking safety? Generally, the dose is specified as milligrams per kilogram of body weight. If say 10mg per kg weight was found to be safe in animal studies, and if our human subject weighs 50 kg, then we can give that person 50X10= 500 mg of the drug during trial. In this sense, the finding from animal study can be extended to humans. In this phase participants are healthy human volunteers. Once the dose is given, subjects are kept under observation and a large number of parameters are checked to see if any one of them shifts away from normal range of values and becomes abnormal. That is regarded as a red flag. What are these parameters? Here is a partial list:
Body temperature, pulse rate, blood pressure, respiration rate, (these are called vital signs)
Electro-cardiogram (ECG): This study reveals effect on heart, if any.
Laboratory tests- blood/ urine chemistry: concentrations of many constituents of blood/ urine are checked e.g. white blood cells, red blood cells, blood sugar, cholesterol, creatinin, sodium, potassium etc.

The number of human participants in a phase 1 trial is quite small (a typical number is 8, of which 6 people are given the drug of interest while 2 are given a placebo).

3.4.3: Phase II: In this phase the drug is introduced in selected patients with the disease for which it is intended. This is the big change. Since in phase I all subjects are healthy, there is no possibility of checking whether the drug in fact gives benefit to the person (this benefit is called efficacy). That aspect is looked into in phase II. Of course the issue of safety is always at the back of all experiments, no matter which phase. Generally number of subjects participating in the trial at this stage is much larger. It may be of the order of 100. These trials are considered as pilot studies and are aimed at identifying the optimal dose (magnitude, frequency and duration, to be tested in Phase III trials). The aim is to identify drugs that have poor prospects and should not make it to the next phase of testing. It allows researchers to build on what they learned in Phase I. Because more subjects are involved, investigators may discover here some uncommon side effects. This phase is sometimes divided into two sub phases. In Phase II (a), the main objective is to evaluate biologic activity. In Phase II (b) interest is in estimating rate of adverse events. At this point the idea of optimal dose can be introduced. Here optimality refers to benefit risk ratio. Efficacy i.e. improvement in the condition of the patient is obviously the benefit. Often it goes up with dose. Adverse impact, side effect etc can be considered as risk. This risk also tends to go up with dose. In other words, prospect of a safe usage declines. So, a good drug yields much improvement together with high assurance of safety i.e. without causing undue harm. If the curve of benefit rises much faster than that for adverse impact then the drug is attractive.

Concept of optimal dose hints at a good choice of trade off. There may be no universal solution to this problem. When disease condition is severe and prognosis is bad, it may seem reasonable to take greater risks. In treatment of cancer, chemotherapy is known to have serious side effects. And yet, there is very limited choice. On the other hand, one would not use a drug with any chance of toxicity, to treat common cold. It would be better to let

nature take its course. Generally, benefits from higher doses are higher too but the curve rises at a falling rate. In other words, incremental gain is small once you reach high dose level. Using language of calculus we can say that the efficacy function has a positive first derivative and a negative second derivative. Story of the safety curve is just the opposite. At low dose, there is only an occasional adverse reaction. But as dose level goes up, chance of avoiding an adverse event falls. In fact at very high doses, it is difficult to avoid adverse effects.

Here is an example of a study of safety of a chemical. "The safety of daily application of N, N-diethyl-m-toluamide (DEET) (1.7 g of DEET/day) in the second and third trimesters of pregnancy was assessed as part of a double-blind, randomized, therapeutic trial of insect repellents for the prevention of malaria in pregnancy (n = 897). No adverse neurological, gastrointestinal or dermatologic effects were observed for women who applied a median total dose of 214.2 g of DEET per pregnancy (range = 0-345.1 g). DEET crossed the placenta and was detected in 8% (95% confidence interval = 2.6-18.2) of cord blood samples from a randomly selected subgroup of DEET users (n = 50). No adverse effects on survival, growth, or development at birth, or at one year, were found. This is the first study to document the safety of DEET applied regularly in the second and third trimesters of pregnancy. The results suggest that the risk of DEET accumulating in the fetus is low and that DEET is safe to use in later pregnancy."

Please try to understand the various terms used and the statistical tools used for judging safety.

3.4.4: Phase III: These studies, which take several years, can involve thousands of patients at multiple trial centers. They are aimed at definitively determining the drug's effectiveness and its side effect profiles. These studies are typically double- blinded and often have an active control group to compare with. (Do you remember what an active control group is?) Investigators try to find out if the new treatment works better than, or is the same as, or is worse than the standard treatment. Participants may range from newly diagnosed patients to people with advanced disease. These are expanded clinical trials to gather additional evidence of effectiveness and to better understand safety and drug related adverse effects.
If the phase III trial results indicate safety and efficacy of the drug at satisfactory level, investigators now have to seek approval for the drug.

Application is made to the regulator of the market in which the drug is to be sold. If approved, the company can manufacture and market the drug. Who are the regulators?
In India it is the Central Drug Control Authority. http://cdsco.nic.in. In the Europe it is European Medicine Agency. http://www.emea.europa.eu. In USA it is Food and Drug Administration. www.fda.gov. In Japan it is Pharmaceutical and Medical Safety Bureau. www.mhlw.go.jp/english/org

Regulatory bodies in different countries were established at different periods. Rules and regulations for approval of drugs developed simultaneously with the progress of clinical research. There were differences in technical requirements of regulatory agencies for drug approval. This created problems for drug produced in one country to be marketed in another.

Getting approval in another country necessitated repetition of research work. This led to higher cost and also to delay in making the drug available to people in another country. Hence it became essential to bring uniformity in norms and regulations for drug approval process. With this objective an 'International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use' was held. All this eventually led to establishment of a new co-coordinating body ICH. In the last ten years or so, ICH has formulated 50 different guidelines. To give a couple of examples, technical documents required for approval needed to be similar. So, ICH has a guideline about the structure of a Clinical Study Report (CSR). In the field of safety, ICH has laid down how to examine risk of cancer due to use of a drug. (The coordinating body ICH has the website. www.ich.org).

What is a CSR? This document contains all relevant information from the trial, so organized as to enable an efficient review process by the regulator. Following website gives details of the guideline developed by ICH: www.fda.gov/CDER/GUIDANCE/iche3.pdf

Main body of this report contains information required by three regulating agencies (USA, Europe and Japan). Any particular item required by one body is provided in the appendices.

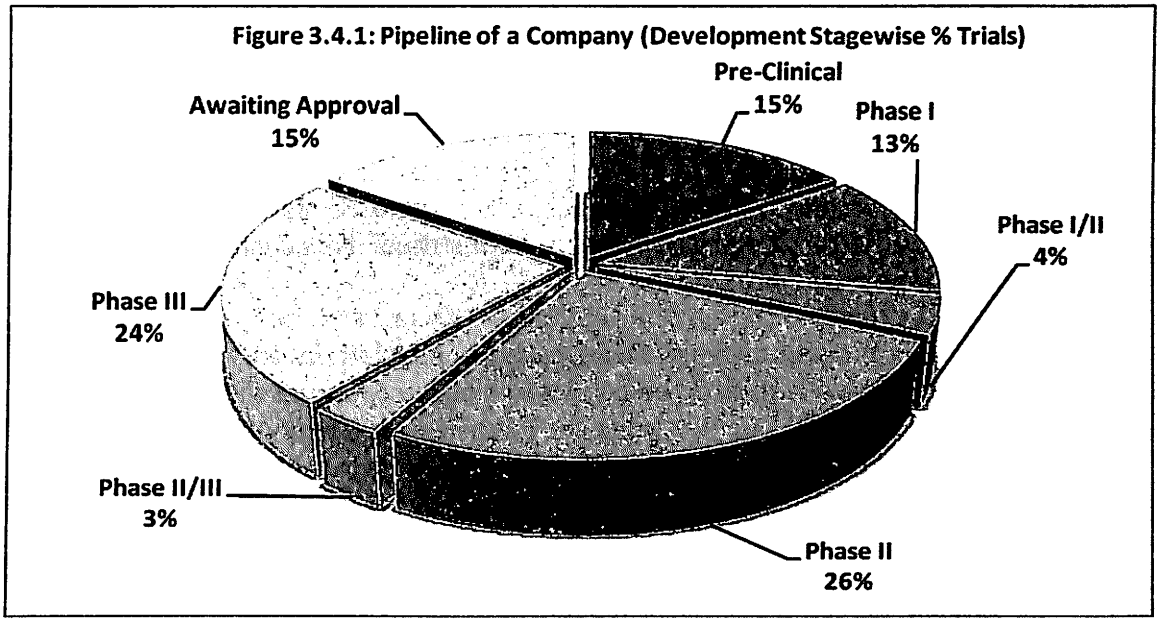
3.4.5 Phase IV: Even after regulatory bodies approve the drug for marketing, research does not end. There are some issues yet to be checked. These include exploration for additional or special benefits. In this phase, after a drug has been launched, pharmaceutical companies conduct further studies to examine long- term safety and to see efficacy of the treatment in certain population groups. If dangerous side effects are found, the drug or treatment is taken off the market. Following table gives some illustrations of such withdrawal.

Table 3.4.2: Some drugs withdrawn from market

Drug	Indication	Adverse effect	Year	Manufacturer
Pondimin	Obesity	Risk of heart valve abnormalities	1973	Wyeth
Rezulin	Type 2 diabetes	Risk of severe liver toxicity	1997	Parke-Davis
Baycol	Cholesterol drug	Risk of severe damage to muscle that is sometimes fatal,	1997	Bayer Pharmaceutical

Studies in this phase may have just a single group, i.e. there may not be a control treatment to compare with. The study population may be different than studied previously (different ethnic group, age group etc.).

Typically a pharmaceutical company has many drugs under investigation simultaneously. They are often in different stages of testing. Following pie chart gives some idea about proportion of drugs in different phases of development in one company.



In view of the long duration of the research process and the limited time for which the drug is under patent protection, this pipeline may decide the long term success of a firm.

Of late productivity of the research pipeline of many companies has fallen. Also, a number of drugs will come out of patent protection/market exclusivity in the second decade of the 21st century. Hence companies, which depended on such drugs, may enter the field of generic drugs. So in all this business of drug development, what exactly is the role of a statistician? What is expected of her/him?

3.4.6 Role of statistician: Statistics plays a crucial role in clinical trial design and analysis. 'Trial Statistician' has the responsibility of ensuring that statistical principles are applied appropriately in clinical trials supporting drug development. This has to be done in collaboration with other clinical trial professionals (such as doctors, pharmacists etc). The statistician has to participate in all stages of a trial right from developing the protocol to submission of final CSR. A very important job is to control bias and increase precision. Bias is inclination in one direction. If the trial has a bias in favor of the test treatment, efficacy of the treatment is overestimated. One way in which such bias can get introduced is by assigning the preferred treatment to a specific group of patients. If more serious patients get the test treatment, perhaps improvement is quantitatively more than if marginally sick patients get that treatment. Here strict adherence to randomization can eliminate the bias. Blinding and randomization are helpful in controlling bias.

Some other sources of bias include faulty design, improper conduct (protocol violations), improper statistical analysis (exclusion of subjects upon knowledge of outcome) etc. Statistician should pay attention to identify all possible sources of bias in a trial.

In addition to controlling bias it is important to ensure that the conclusions drawn are robust. What does it mean? Statistical tests used are based on assumptions. In reality these assumptions may not be satisfied or at best, satisfied approximately. How serious is the

impact of such deviation on the conclusions? Statistician should test robustness of conclusions by reanalysis using different methods or using different assumptions etc. Results should be stable.

What is a protocol? It is a document that spells out all the important details of trial design; its conduct and principal features of its proposed statistical analysis. Any deviation from steps specified in the protocol has to be justified. This ensures that no convenient post-hoc explanations are given.

An overall plan, which needs to be conducted in an orderly manner, should be developed. In this plan, specific objectives at each step, appropriate decision points should be defined and there should be flexibility for modification as knowledge accumulates. If several trials are involved, a meta- analysis may be informative. Meta analysis means combining results from several trials. This can be done effectively if broad aspects of the protocols of these trials are similar.

Thus drug development from concept to market is a long process. Even after a drug is in the market, trials and their evaluation continue. Statistician has to play an important role at each step.

3.4.7 Exercises

E3.4.1 Insecticide efficacy A particular mosquito repellent named DEET has been studied thoroughly. 50% concentration of DEET provides about 4 hours of protection against the mosquito species Aedes egypti. Increasing the concentration further to 100% gives only 1 extra hour of protection. Use these numbers to draw a concentration-protection time curve. Assume a logistic form. Try K = 24 hrs.

E3.4.2 Insecticide efficacy Here is a study of DEET for protection against Aedes albopictus. At 12.5% concentration we get 6 hours of protection. Doubling concentration to 25% increases the protection time to 8 hours. Now draw a curve depicting use of DEET against this mosquito species.

3.5 Some designs commonly used in clinical trials: So now we know that once a drug is thought of as a possible cure for a disease, it is sent through a long drawn out process of testing. This process can take as much as ten years and can cost hundreds of crores of rupees. So, it is very important to ensure that the tests are conducted in the best possible and most efficient way. One of the critical inputs in this research is statistical design of the experiment. We will take a brief review of designs commonly used in clinical trials.

3.5.1 Parallel groups design: Let us begin in a naïve way. We have a few patients; let us say of high blood pressure. (We recommend that students read some material about hypertension. They should have some idea of systolic and diastolic bp, and also of different positions in which the person is asked to be while bp is measured. [recumbent, semi-recumbent, standing, supine etc.]). We have a proposed medication. So how do we test it? Well, we measure bp first, and then give a dose of the drug and after suitable time period, measure bp again. We

have bi-variate data i.e. two observations on each patient; one before and second after treatment. Our null hypothesis is that population mean of difference is zero. If this null hypothesis is rejected in favor of the alternative that bp after treatment is lower than bp before, then we have confirmed efficacy. Let us call this 'Plan 1'.

Now here are a couple of questions about this plan. Firstly, what test will you use for this hypothesis? Secondly, do you find any weakness in the plan?

At this point you should close the book and think about these questions. Discussing them among friends is also a good idea. Please write down your answers in your note book before you continue to read. The reason for this request is that it is the only way we can put you in the situation that an applied statistician is in, every day.

Here are the answers to the questions raised above. Test: We calculate difference (before- after). It is often called 'change from base line and abbreviated as CFB. We can apply a one sample t test to CFB if the data are from a normal distribution. If not, a Wilcoxon signed rank test will be more suitable. Plan: The plan of this trial has one weakness. There is no 'control'. Suppose 'bp after' is significantly lower than 'bp before'. How do we know that it is due to the drug? It could be a **placebo effect**.

If the test drug group shows reduction in bp significantly greater than reduction in the placebo group, then we can safely claim support for the test drug. A design in which there are two independent groups (or possibly more than two) getting two different treatments is called a parallel design. In parallel designs, we have two or more independent samples. Here subjects are randomized to one of two (or more) arms. Each arm is allocated a different treatment. This set of treatments includes the test treatment(s) at one or more doses and also control(s)/ placebo. This is a design that is simple to understand, simple to implement and simple to analyze. The basic analysis is one way ANOVA. In view of these comments we propose that there should be two groups. One group receives aspirin (a common supplement for hypertension patients) and the other receives placebo. In each case bp is measured before and after treatment, differences are taken and a two sample t-test is applied to two sets of CFBs. The null hypothesis is equality of means. The alternative is that the mean is greater for aspirin. Let us call this 'Plan 2'.

Now you can comment on plan 2. Do you find it satisfactory? Can you punch any holes? Here is a hint. The new plan is better than the earlier one. But it still has a weakness. Can you see it? Perhaps discussing the plan with someone who takes medication for high bp may give you some hints. Again, give yourself some time to search for a weakness and then return to the book.

Please recall that placebo treatment involves a pill without an active medication in it. If you are a patient of bp, you would not like to be on a placebo. Ethics of conducting trials demands that we should not subject people to unnecessary risk. So, a group on placebo is not a good idea. Instead, we should compare the test drug with some other drug already present in the market. Such a drug will be called an active control.

In contrast, if the same person receives two treatments, (placebo in period one and test drug in period two) it will be called a fixed sequence design. The idea here is that all individuals receive two treatments in the same sequence. If the sample is spilt into two groups and for the second group the sequence is flipped then we get a cross over design.

Here is description of a trial related to blood pressure. It is mainly illustrative and for discussion.

3.5.2 A case study- Yoga and Blood Pressure

It has been suggested that yoga exercise can reduce blood pressure. Twenty men participated in an experiment to test this hypothesis. Ten of the men took a yoga course for 12 weeks while the others did not. Response was blood pressure of each subject before and after the 12-week period.

Is blinding possible in this experiment? Answer is in the negative. Think of reasons for it. Further, what can we say about the ethical issue of exposing patients to unnecessary risk? Are the people in the control group subject to greater risk? Well, we shall assume that each patient was taking usual medication for the hypertensive condition. In this sense there is no extra risk for the control group. But that raises another question. Is it possible that result of comparison of yoga with control may depend on the medication in use? It is such doubts that necessitate increasingly complex designs for RCT. (This is one common short form for randomized controlled trials.) What about the treatment group? There should be some prior research on safety in yoga. If literature gives many examples of adverse events (pain, injury, dizziness etc) then the treatment may not be worth testing. It better be dropped on safety grounds. What would you recommend if it is felt that effect of yoga may depend on the medication that participants are on?

Table 3.5.1: Yoga and Blood Pressure

Treatment	SBPBegin	SBPEnd	Decrease	Treatment	SBPBegin	SBPEnd	Decrease
yoga	107	100	7	control	123	124	-1
yoga	110	106	4	control	109	97	12
yoga	123	105	18	control	112	113	-1
yoga	129	112	17	control	102	105	-3
yoga	112	115	-3	control	98	95	3
yoga	111	116	-5	control	114	119	-5
yoga	107	106	1	control	119	114	5
yoga	112	102	10	control	112	114	-2
yoga	136	125	11	control	110	121	-11
yoga	102	104	-2	control	117	118	-1

Treatments: (a) yoga exercise and (b) control, Begin: seated diastolic blood pressure before treatment (units- mmHg), End: seated diastolic blood pressure after treatment, Decrease: Decrease in blood pressure (Begin - End)

Coming back to the trial, what is the null hypothesis of interest? What is the appropriate test? Are distributional assumptions satisfied? Table 3.5.1gives data collected during the trial.

The analysis is to be done using some software. We will illustrate use of EXCEL here. We will first carry out a two sample t test. The variable of interest is 'difference' and the two samples are yoga and control. The output is as follows:

You have to learn how to get this output using EXCEL. Also, since this is a simple problem, it would be useful to verify the entries by doing the analysis using calculators (and not using computer programs). In any case we have to understand the following output.

Table 3.5.2: EXCEL Output for Two Sample t-test

	Variable 1	Variable 2
Mean	5.8	-0.4
Variance	66.84444	37.6
Observations	10	10
Pooled Variance	52.22222	
Hypothesized Mean Difference	0	
df	18	
t Stat	1.918444	
P(T<=t) one-tail	0.035529	
t Critical one-tail	1.734064	
P(T<=t) two-tail	0.071057	
t Critical two-tail	2.100922	

Note that 5.8 is the mean of the 'decrease' for yoga group while -0.4 is the mean for the control group. Null hypothesis is of no difference between two group means. Alternative of interest is one sided, that yoga has a mean decrease that is larger than control population. [Do you see why this is the alternative of interest? We expect yoga to reduce blood pressure more than control. Since we are taking the difference (Begin-end), larger difference means a better treatment. If we were to take the difference (End-Begin), it would be opposite. This fact has to be understood clearly. Perhaps it is simple. But we encounter too many students who miss it.] In fact the sample means are consistent with the alternative. In other words, average sample reduction in blood pressure with yoga is 5.8 mmHg while the control group experiences a rise in the average. Question is whether there is statistical significance. That is indicated by the one tail probability, which is 0.0355. This is small. So, we can suggest that the trial resulted in supporting the claim that yoga exercises help in reducing blood pressure.

Wait! Do not rush to decide. In science it helps to be skeptical. Usefulness of a treatment has to be proved beyond reasonable doubt and statisticians have to help doctors in raising reasonable doubts. So, in the present case, we will raise a doubt. Can you see that the two cases with largest initial value of blood pressure are in the yoga group? Perhaps the yoga group has overall higher initial blood pressure and such individuals are more prone to benefit from any treatment. So, the apparent superiority of yoga may be due to high initial values.

This may or may not be true. It is just a reasonable doubt. We can use the technique of analysis of co-variance (ANCOVA) to respond to such a doubt. The strategy to resolve the doubt is as follows: We argue that the model appropriate in the present case is

$$y = C_0 + C_1 X_T + C_2 X_B + \varepsilon$$

Here y is the response. C_0 is the general mean. X_T is the variable indicating which treatment the subject gets. It equals 1 if the subject practices yoga and equals zero otherwise. Coefficient of X_T namely C_1 can be interpreted as the mean difference between yoga and control. (Of course this is informal. Formally speaking, it is the difference in the population means of the two groups.) The novelty in the model is in the next term. Here X_B is the initial blood pressure. It is a variable that may contain some information about y and in this sense it is a covariate of y. Its inclusion in the model reflects our suspicion that initial blood pressure may affect response. If that conjecture is true, regression coefficient C_2 should be nonzero. If it turns out to be essentially equal to zero, we can conclude that our concern was wrong. So, the null hypothesis $C_2 = 0$ is of interest to us. If we reject the null hypothesis then comparison of treatments has to be carried out after correcting for/ eliminating the effect of the covariate. How do we do this? The null hypothesis of no treatment effect is nothing other than $C_1 = 0$. If the covariate is included in the model and we use least squares approach, then this hypothesis about treatments is tested with due correction for covariate effect. [Here we are assuming that effect of the covariate is linear, if any.]

So, now we will carry out comparison of treatments and we will use regression analysis for this purpose. You will wonder why we use regression analysis when we have two groups to compare and we should use a two sample t test. The reason is that regression method can be easily extended to problems with more than two treatments and in fact to very complicated situations. How does one use regression analysis? We have to replace treatments by suitable indicator variables. We should create a column with 'yoga' replaced by 1 and 'control' replaced by 0. This is what we have already done. Name for such a variable is 'an indicator variable'. Another name is 'dummy variable'.

Table 3.5.3: EXCEL output for ANOCOVA

SUMMARY OUTPUT Regression Statistics						
Multiple R	0.578999					
R Square	0.33524					
Adjusted R Square	0.257033					
Standard Error	6.653798					
Observations	20					
ANOVA	**df**	**SS**	**MS**	**F**	**Significance F**	
Regression	2	379.5584	189.7792	4.286565	0.031092	
Residual	17	752.6416	44.27303			
Total	19	1132.2				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-39.0349	18.89824	-2.06553	0.054465	-78.9067	0.836912
X Variable 1*	5.057571	3.027047	1.670793	0.113069	-1.32894	11.44408
X Variable 2**	0.346191	0.168286	2.057155	0.055349	-0.00886	0.701244

* Baseline covariate, ** Treatment indicator variable (0/1)

One point of detail concerns the number of digits after the decimal point. This table has values with 5 digits after the decimal point. We have reproduced the table without editing. It is better to retain only a couple of digits while reporting. Sometimes there is a convention about this. If not, we should minimize the digits consistent with the message to be conveyed. [What does that mean? It means that context decides what the right answer is. In the year 2009, one US $ is roughly equal to fifty Indian rupees. This crude answer is good enough for general purpose. On the other hand if you are a foreign exchange dealer, then the rate on February 26 at 5pm was 49.9251123 Indian Rupees(INR) for one US $. Why so many digits after the decimal places? Because this value is to be used in buying/selling millions of dollars and crude answer can cause financial loss.]

Now notice that the p-value associated with 'X variable 1' (i.e. with test of $C_1 = 0$) is 0.113069. This is rather large and it would not be right to claim significant difference between the two treatments. So, it seems, our doubt may have something to it. Or are we making a mistake? Have you realized that the test is two sided? [Please use the value of the test statistic and by referring to a t distribution, check using EXCEL what the one sided p-value should be.] So, in view of the fact that we are interested in a one sided alternative, we should really look at half of the value given. It is slightly more than 0.056. Shall we continue to say that the null hypothesis cannot be rejected? Some people will be firm in this. If we work at 5% level and p-value exceeds the level of significance chosen, we must stay with the null hypothesis. This is one view. Some people will be somewhat hesitant. We hope students understand that level of significance is not a very sacrosanct number. Students may complain that things are too shaky here. They would prefer a firm rule. We appreciate any such feelings. A sharp rule with no ambiguity is so much easier to use. But in real life things are quite different. Realizing that decisions in real life are not as cut and dry as mathematical theorems is a sign of maturity as an applied statistician. In any case once we take into account the covariate, p- value changes from 0.0355 to 0.056. That is an indication of the role of covariate. We can say that initial value does impact change in blood pressure.

3.5.3 Factorial designs: When a patient is treated with only one drug it is called mono-therapy. Instead when the treatment consists of more than one drug it is called combination therapy or multi-drug therapy. A factorial design may be relevant while studying multi-drug therapy.

The important concept of a factorial design was introduced by R. A. Fisher. Let us understand what is so special about it. Prior to his time, the norm of experimental design in science was 'change one thing at a time'. You wish to study effect of 3 factors on taste of tea. Factors are (i) quantity of sugar (ii) quantity of milk and (iii) duration of boiling of tea leaves in water. Does taste improve if we add more sugar? To answer this question (which in technical terms would be called main effect of sugar), we keep other factors constant at chosen levels. Why? Obvious! If we change quantity of sugar as well as quantity of milk at the same time, and taste of tea improves, what is it due to? Is it due to sugar or due to milk? Situation is much worse if several factors are changed at once. Fisher overhauled this approach completely. He showed how to make sense of observations obtained after changing many factors at the same time. Not only that, but he went further and showed that changing

multiple factors at the same time was the only way by which one could bring out the phenomenon of interaction between factors. In other words, it is only with a factorial experiment that we can study interdependence among factors. If effect of one factor changes as level of another factor is changed, that will be brought out only by a factorial design.

A simple 2^2 factorial design would have one group of subjects testing therapy A alone, another testing therapy B alone, a third group testing A and B combined, and a control group receiving neither A nor B. Factorial designs are considered an efficient way to test medicines in combination, but their results are not always easy to interpret. The main interest in such a trial is whether a combination of treatments A and B is better than either of them alone (or in other words, whether the two drugs interact). One of the most useful things to do as part of analysis of a factorial experiment is to plot cell means. On X axis we take levels of one factor. On Y axis we take response. When a cell mean is plotted, we write the level of the second factor beside the point. Such a plot can show presence of interaction.

Let us consider one concrete example. In treating hypertension, mono-therapy is successful in only about half the cases. It does not work well in the remaining half. This may be because multiple mechanisms are involved in causing hypertension, and a single drug class is inadequate to counter all the causes.

The web site http://www.merck.co.za/home.asp?pid=945 gives some information on combination drug therapy for hypertension. Diuretics and ß-blockers (ask doctors about these terms. Also check on the web) are excellent agents to combine in treating hypertension because together they can lower blood pressure more effectively than when either drug is used alone. Now to find 'best' dose level (combination) of the component drugs, factorial design may be helpful. The table below is a summary given on this webpage.

Table 3.5.4: Factorial Design for BP study

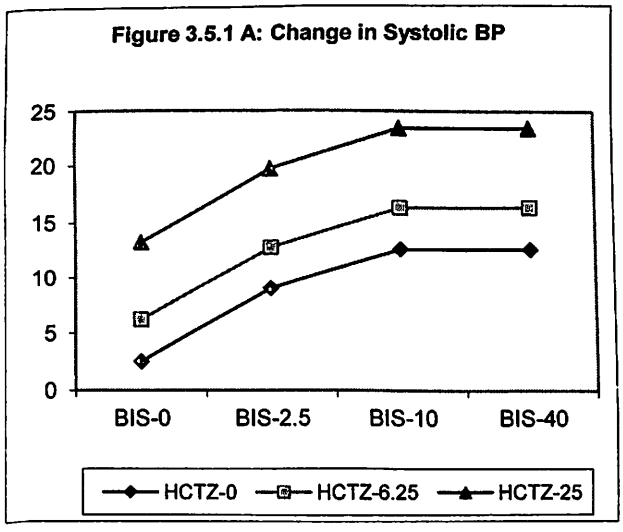
Level (mg / dl) of		# Patients.	DBP-CFB	SBP-CFB
HCTZ	Bisoprolol			
0	0	56	3.8 ± 0.7	2.5 ± 1.2
	2.5	59	8.4 ± 0.7	9.0 ± 1.2
	10	62	10.9 ± 0.7	12.6 ± 1.2
	40	59	12.6 ± 0.7	12.6 ± 1.2
6.25	0	23	6.4 ± 0.9	6.2 ± 1.5
	2.5	28	10.8 ± 0.8	12.8 ± 1.4
	10	25	13.4 ± 0.8	16.4 ± 1.5
	40	29	15.2 ± 0.8	16.4 ± 1.4
25	0	33	8.4 ± 0.8	13.3 ± 1.4
	2.5	30	12.9 ± 0.8	19.9 ± 1.4
	10	30	15.4 ± 0.8	23.5 ± 1.4
	40	31	17.2 ± 0.8	23.5 ± 1.4

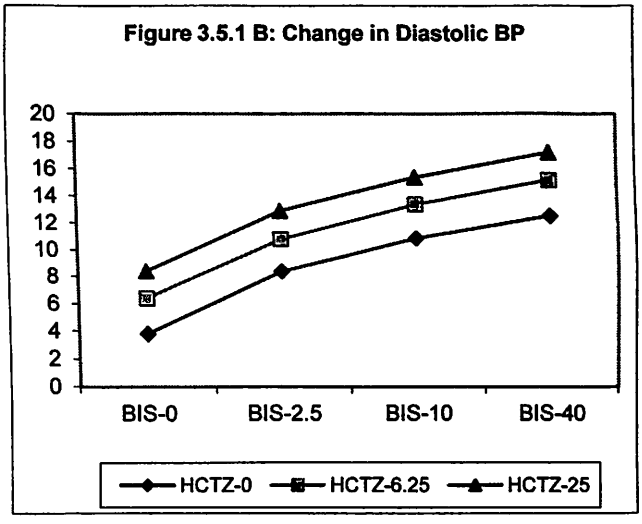
Change in mean BP due to change in HCTZ (at different levels of BIS)

Table 3.5.5 A: Mean Change in SBP over Dose levels		
Level of BIS	HCTZ changes from 0 to 6.25	HCTZ changes from 6.25 to 25
0	6.2-2.5= 3.7	13.3-6.2=7.1
2.5	12.8-9.0=3.8	19.9-12.8=7.1
10	16.4-12.6= 3.8	23.5-16.4=7.1
40	16.4-12.6= 3.8	23.5-16.4=7.1

Table 3.5.5 B: Mean Change in DBP over Dose levels		
Level of BIS	HCTZ changes from 0 to 6.25	HCTZ changes from 6.25 to 25
0	6.4-3.8=2.6	8.4-6.4=2
2.5	10.8-8.4=2.4	12.9-10.8=2.1
10	13.4-10.9=2.5	15.4-13.4=2
40	15.2-12.6=2.6	17.2-15.2=2

The above tables show the effect of change of HCTZ at different levels of BIS. The differences remain remarkably constant. Of course there may be differences in the second decimal place. But nevertheless, this is a case of absence of interaction.





Look at the graph for systolic BP.

a) Each line represents one level of HCTZ.
b) Whatever the level of HCTZ, the response (change in BP) increases with level of BIS. The increase tapers off after BIS = 10.

Now look at the second graph.

c) Here also, the response increases with level of BIS at each level of HCTZ.
d) For both the drugs effect increases with dose, at a decreasing rate.
e) The fact that we have parallel lines shows that there is no interaction. Effect of changing level of HCTZ remains the same whatever is the level of BIS.
f) Now you have to do similar work with roles of two treatments reversed. Thus you should draw graphs with level of HCTZ on x axis and there should be one line graph for each level of BIS

3.5.4 A two group parallel design for drug interaction: While factorial design described above is a fine way to study two-drug interaction, it is not the only way. In fact there is a reason for avoiding the factorial design. It is not a statistical reason but an ethical one. In the

factorial design, one group gets treatment A, second group gets treatment B, third group gets both A and B and fourth group gets only placebo. In some situations, placebo as one treatment may not be acceptable. It may be hazardous to have a group of patients with no medication at all. Hence some alternative has to be found. In the alternative design, there are only two groups of patients. First group gets treatment A in the beginning. This is the so called period I. Then in period II the group gets a combination therapy i.e. two drugs A and B together. The second group gets drug B in period I, and combination A+B in period II. Another way of describing the design is to say that there are two parallel groups. Each group takes a sequence of two treatments in two successive periods. The two sequences are (i) A followed by A+B and (ii) B followed by A+B. Main interest is comparing efficacy of A in presence of B with that of A alone and similarly efficacy of B in presence of A with that of B alone.

A fixed sequence two group parallel design		
Group/Period	Period I	Period II
Group I	A	A+B
Group II	B	A+B

We can compare results of group I in two periods. Differences if any will be attributed to treatment B (given that treatment A is being administered also). Results in the first period for group II tell us about effect of B in the absence of A. If the two measures are similar, we can say that effect of B is the same regardless of whether A is given or not. Analogous comparison will tell us if effect of A remains the same in the presence or absence of B. So this is another way to study interaction (and eliminating the placebo group). We hope you see that the job of a trial statistician is to provide a design, which satisfies all the constraints of a clinical trial and still obtains valid estimates of effects of interest.

3.5.5 Crossover Design

Another modification of the randomized controlled trial is the crossover design. This is particularly useful when outcome is measured by reports of subjective symptoms, but it can only be applied when the effects of treatment are short lived (for example, pain relief from an analgesic).

In a crossover study, eligible patients who have consented to participate receive each treatment sequentially, often with a "wash out" period between treatments to eliminate any carry over effects. However, the order in which treatments are given is randomized so that different patients receive them in different sequence. Outcome is monitored during each period of treatment, and in this way each patient can serve as one's own control.

Let us first understand the main motivation behind this design. We have already discussed the importance of local control in comparing treatments. In simple terms, it means we have to ensure a level playing field for all treatments under comparison. If the experimental units are homogeneous, then each treatment gets similar subjects and there is no difficulty with validity of comparisons. But that is an uncommon situation. By and large, subjects do differ among themselves in ways that are important for safety, efficacy etc. So, to ensure valid comparisons, we can create relatively more homogenous subgroups of subjects and then compare treatments within each such subgroup. But in fact, individuals are rarely

similar in their genetic make up. (The only exception being twins and even that statement is true only if they are not just any twins but identical twins).

Perhaps one way out of the difficulty here is to make each subject his/her own control. But how can a person take two medications at the same time? Of course not! So, we give the two medications to the same person on two different days (periods). But then the effect of the drug given earlier may remain and get mixed up with effect of the second drug. To take care of this hurdle, we can leave a sufficient gap between the time points of taking the two drugs. This gap has to be long enough so that any residual effect of earlier medication is eliminated from the body. Also, clearly, the medication has to be a short term remedy and the sickness has to be of a low intensity and long term nature (such as asthma). When all these conditions are satisfied, we can eliminate any objections about comparability of subjects.

One last problem is the possible dependence of response on the 'sequence' in which medications are given. Perhaps the treatment given first shows better result. If that is the case, validity of comparison is in jeopardy again. To take care of such a possibility, another trick is used. A second group of subjects is given the two drugs in the reverse sequence. Taking one treatment first and then switching to the second treatment is sometimes called a cross-over, hence the name of the design. This is how we get the two period two sequence cross over design (2X2).

Suppose we name the two treatments as A and B. We have two periods. Hence the design may be described in terms of sequences. Sequence AB is implemented if treatment A is given in period 1 and treatment B is given in period 2. Sequence BA is interpreted similarly. Each subject is assigned to one of these two sequences using randomization. This simple maneuver reduces the number of subjects and the number of assessments needed to achieve a specific power. The time gap in two successive treatment administrations is termed 'a washout period'.

Crossover designs have a number of problems that can invalidate their results. The chief difficulty concerns carryover, that is, the residual influence of treatments in subsequent periods. In an additive model the effect of unequal carryover will be to bias direct treatment comparisons. If two treatments have similar carry over effect then treatment comparison is not affected.

When crossover design is used it is important to avoid carryover. This is best done by selective and careful use of the design on the basis of adequate knowledge of both the disease area and the new medication. The disease under study should be chronic and stable. The relevant effects of the medication should develop fully within the treatment period. The washout periods should be sufficiently long for complete reversibility of drug effect. The fact that these conditions are likely to be met should be established in advance of the trial by means of prior information and data. There are additional problems that need careful attention in crossover trials. The most notable of these are the complications of analysis and interpretation arising from the loss of subjects. Also, the potential for carryover leads to difficulties in assigning adverse events, which occur in later treatment periods to the

appropriate treatment. These, and other issues, are described in ICH E4. The crossover design should generally be restricted to situations where losses of subjects from the trial are expected to be small.

A common, and generally satisfactory, use of the 2×2 crossover design is to demonstrate the bioequivalence of two formulations of the same medication. In this particular application in healthy volunteers, carryover effects on the relevant pharmacokinetic variable are most unlikely to occur if the wash-out time between the two periods is sufficiently long. However it is still important to check this assumption during analysis on the basis of the data obtained, for example by demonstrating that no drug is detectable at the start of each period.

Theory and analysis of cross over designs are complicated. They are subjects of full length books. (see Byron Jones and Michael G. Kenward (1990) Design and analysis of cross over trials, Chapman and Hall, Stephen Senn (2002) Cross over trials in clinical research, Wiley).

A general cross over design is complicated hence we will restrict our discussion to the simplest case viz. 2 X 2. Instead of using treatment labels A and B we use T (test) and R(reference). Following matrix explains the nature of a 2X2 design in a simple way.

Layout of cross over design: Treatment administration by sequence and period		
Sequence	**Period 1**	**Period 2**
RT	R	T
TR	T	R

Bioequivalence implies similar bioavailability. Therefore the aim is to see if bioavailabilities of two treatments are comparable. Two nuisance factors (carry over and period effect) arise because of peculiar nature of cross over design. They have to be tackled first. So analysis of cross over design is carried out in three steps.

Step 1: Test for carry over effect
It is possible that response to drug administered in second period is partly attributable to the drug administered in period 1 (see table above). This effect is called a 'carry over effect'. Hence first we test for carry over effect. If it is significant then we cannot use data from both periods for treatment comparison. We will have to use only data from period one and the procedure will be similar to the one for parallel design.

Step 2: Test for period effect
Each subject gets two treatments in different periods. Hence we have to account for possible differences in response due to periods. That is why we test the hypothesis of equality of period effects.

Step 3: Comparison of two treatments

This is the main objective. We will now provide illustration of above procedures with one data set. Source of data: Bradstreet, T.E. (1992) "Favorite Data Sets from Early Phases of Drug Research - Part 2." *Proceedings of the Section on Statistical Education of the American Statistical Association}.*

Eight healthy male volunteers were allocated randomly to a two-period crossover design. Each subject followed one of two eating sequences [Fasted (-) then Fed (+), or Fed (+) then Fasted (-)]. During each study period the subjects received a single dose of a antihypertensive therapy. A five to seven day washout period separated the two study periods. The pharmacokinetic variables namely area under the plasma concentration by time curve (AUC), maximum plasma level (Cmax), and the time to maximum plasma level (Tmax) were calculated for each subject from plasma concentrations assayed. The pharmacokinetic parameters were estimated both for the parent compound (P) and the metabolite (M). Question of interest is 'Does food intake affect pharmacokinetics of the parent compound/ metabolite?' We will take you through analysis steps for data on parent compound.

Table 3.5.6: Food Effect PK Study for Antihypertensive Therapy (Parent Compound)

Sub	Seq	Period	AUC (ng*hr/ml)	Trt	Cmax (ng/ml)	lnAUC	lnCmax	Tmax (hours)
3	FaFe	1	901.11	Fa	475.5	6.80363	6.16437	1.250
4	FaFe	1	1146.96	Fa	688.2	7.04487	6.53408	0.667
6	FaFe	1	745.51	Fa	469.9	6.61407	6.15252	0.667
8	FaFe	1	852.86	Fa	908.5	6.74860	6.81179	0.667
1	FeFa	1	809.44	Fe	155.8	6.69634	5.04857	1.750
2	FeFa	1	428.00	Fe	139.1	6.05912	4.93519	4.000
5	FeFa	1	712.24	Fe	326.5	6.56841	5.78843	3.000
7	FeFa	1	511.84	Fe	117.7	6.23801	4.76814	5
3	FaFe	2	757.71	Fe	633	6.6303	6.45047	0.833
4	FaFe	2	906.83	Fe	258.5	6.80996	5.5549	3
6	FaFe	2	561.77	Fe	212.6	6.33109	5.35941	5
8	FaFe	2	756.6	Fe	294.6	6.62883	5.68562	1
1	FeFa	2	967.82	Fa	440.6	6.87505	6.08814	1.517
2	FeFa	2	746.45	Fa	493.3	6.61533	6.20112	0.833
5	FeFa	2	678.16	Fa	366.6	6.51938	5.90427	1.667
7	FeFa	2	568.98	Fa	269.4	6.34385	5.5962	3

The three effects viz. carry over; period and treatment are functions of means corresponding to four cells shown in the 'Layout Table' above. Hence we first calculate these cell means. In the table 3.5.7, the average response (logAUC) of 6.803 corresponds to sequence 1 (FaFe) and period 1. This average is denoted by ($\overline{Y}_{P1S1}$). Since the sequence FaFe means fasted condition will occur in first period and fed condition in the second period, the response 6.803 is under fasted condition. Similarly we can interpret other entries in the table.

Table 3.5.7: Sequence X Period Cell Means			
	Seq1(FaFe)	Seq 2(FeFa)	All
Period 1	6.803	6.390	6.597
	$\overline{Y}_{P1S1}$	$\overline{Y}_{P1S2}$	
Period 2	6.600	6.588	6.594
	$\overline{Y}_{P2S1}$	$\overline{Y}_{P2S2}$	
All	6.701	6.489	6.595

Testing for carry over effect:

In terms of cell means the formula for estimating carry over effect is given below.

$$\hat{C}=(\overline{Y}_{P1S1}+\overline{Y}_{P2S1})-(\overline{Y}_{P1S2}+\overline{Y}_{P2S2})$$

Now we substitute relevant values and get $\hat{C}$ = (6.803+6.600) - (6.390+6.588) = 0.425

Variance of the above estimate is given by the following formula.

$$\hat{\sigma}_C^2=\frac{1}{2(n-1)}\sum_{k=1}^{2}\sum_{i=1}^{n}[(Y_{iP1Sk}+Y_{iP2Sk})-(\overline{Y}_{P1Sk}+\overline{Y}_{P2Sk})]^2 .$$

Here n is the number of subjects in each sequence (4 in the present illustration). [For simplicity we are assuming equal number of subjects per sequence. In general the number of subjects may be different in two sequences. In that case the formula becomes slightly more complicated.] In the above expression, the first bracket ( ) inside square bracket contains sum of responses in two periods for the same subject; second bracket is the mean of these sums over subjects in a sequence. These terms are calculated for each subject, their difference taken, squared and summed over all subjects from a sequence. We have to calculate these for two sequences and add. Table 3.5.8 shows these calculations. The variable under study is logarithm of AUC.

Table 3.5.8: Calculations for Carryover effect and Its Variance						
Sub	Sequence	lnAUC (P1)	lnAUC (P2)	Sum over 2 periods	Sum-Mean	Square
(1)	(2)	(3)	(4)	(5)=(3)+(4)	(6) =(4) -13.40 (or 12.98)	$(6)^2$
3	FaFe	6.80363	6.6303	13.43393	0.0310925	0.000967
4	FaFe	7.04487	6.80996	13.85483	0.4519925	0.204297
6	FaFe	6.61407	6.33109	12.94516	-0.4576775	0.209469
8	FaFe	6.7486	6.62883	13.37743	-0.0254075	0.000646
1	FeFa	6.69634	6.87505	13.57139	0.5925175	0.351077
2	FeFa	6.05912	6.61533	12.67445	-0.3044225	0.092673
5	FeFa	6.56841	6.51938	13.08779	0.1089175	0.011863
7	FeFa	6.23801	6.34385	12.58186	-0.3970125	0.157619
			Mean FaFe	13.40284	Sum	1.02861
			Mean FeFa	12.97887	SigmaSq (C)	0.171435

The t-statistic for testing the hypothesis 'carry over effect =0' is given by

$T_C = \dfrac{\hat{C}}{\hat{\sigma}_C\sqrt{\dfrac{2}{n}}} = \dfrac{0.425}{\sqrt{0.171435(\frac{2}{4})}} = 0.425/0.2928 = 1.45$. This will have a t- distribution with 2(n-1) =6 degrees of freedom. This will be a two-sided test and corresponding p-value is 0.197. (Why two sided? Presence of carry over implies that treatment in the first period has an impact on treatment effect in second period. The impact can be either inflating or deflating the response. Either way it distorts measurement of response to treatment in second period). The p-value is large and carry over effect is not statistically significant. Hence data from second period can also be used for treatment comparison.

Testing for Period effect:

Now we check significance of second nuisance factor viz. period effect. The estimate of period effect is given by $\hat{P}=\frac{1}{2}[(\bar{Y}_{P1S1}-\bar{Y}_{P2S1})+(\bar{Y}_{P1S2}-\bar{Y}_{P2S2})]$ .The two brackets represent difference (period 1mean- period 2 mean) for sequence S1 and S2 respectively. The terms for two sequences are then added to get estimate of period effect. In our example we get on substitution $\hat{P}=\frac{1}{2}[(6.803-6.6)+(6.39-6.588)]=(0.203+(-0.198))/2=0.0025$. Variance of this estimate is given by, $\hat{\sigma}^2{}_P=\dfrac{1}{2(n-1)}\sum_{k=1}^{2}\sum_{i=1}^{n}[(Y_{iP1Sk}-Y_{iP2Sk})-(\bar{Y}_{.P1Sk}-\bar{Y}_{.P2Sk})]^2$

In the above expression, first bracket () inside square bracket gives period difference for individual subject and second bracket gives the difference of period averages. These terms have to be calculated separately for each sequence, squared and then added. These calculations are shown in Table 3.5.9.

Table 3.5.9: Calculations for Period effect and its variance

Sub	Sequence	lnAUC (Period1)	lnAUC (Period2)	PeriodDiff	PriodDiff- MeanDiff	Square
3	FaFe	6.80363	6.6303	0.17333	-0.0294175	0.000865
4	FaFe	7.04487	6.80996	0.23491	0.0321625	0.001034
6	FaFe	6.61407	6.33109	0.28298	0.0802325	0.006437
8	FaFe	6.7486	6.62883	0.11977	-0.0829775	0.006885
1	FeFa	6.69634	6.87505	-0.17871	0.0192225	0.00037
2	FeFa	6.05912	6.61533	-0.55621	-0.3582775	0.128363
5	FeFa	6.56841	6.51938	0.04903	0.2469625	0.06099
7	FeFa	6.23801	6.34385	-0.10584	0.0920925	0.008481
			Mean FaFe	0.202748	Sum	0.213426
			Mean FeFa	-0.19793	SigmaSq (C)	0.035571

The t-statistic for testing the hypothesis 'period effect =0' is given by

$T_p = \dfrac{\hat{P}}{\sqrt{\hat{\sigma}_P^2 \dfrac{1}{2n}}} = \dfrac{0.0025}{\sqrt{0.035571(\frac{1}{8})}} = 0.0025/0.0667 = 0.035$. This has a t- distribution with 6 degrees of freedom. Here also we have a two-sided test. The corresponding p-value is 0.973. Period effect is also not statistically significant.

Comparison of treatments:

Lastly we compare AUC under T (Fed) and R (Fasted) conditions. The treatment effect (generally denoted by F for formulation) is given by

$$\hat{F} = \frac{1}{2}\left[\left(\overline{Y}_{P1S1} - \overline{Y}_{P2S1}\right) - \left(\overline{Y}_{.P1S2} - \overline{Y}_{P2S2}\right)\right].$$

The two brackets inside square bracket are same as those for period effect but here we subtract the terms for two sequences. Therefore, $\hat{F} = (0.203 - (-0.198))/2 = 0.2005$. The estimate of variance for $\hat{F}$ is the same as that for $\hat{P}$. Using this $\hat{F}$ and corresponding [SE($\hat{F}$)] we get the 90% confidence interval for the difference as [$\hat{F} \pm t_{6,\,0.95}$* SE ($\hat{F}$)], where $t_{6,\,0.95}$ is the upper 0.95 value of the t distribution with 6 degrees of freedom. On substitution we get LL = [0.2005 – 1.943*0.0667] and UL= [0.2005 + 1.943*0.0667]. The resulting confidence interval is (0.0709, 0.3301). Since this CI excludes zero, the difference between bioavailability of treatment under fast and fed condition is different.

We will end this section with brief discussion of two points. (1) comparison of parallel and cross over designs and (2) comments on limitations of the analysis presented above.

Comparison of parallel and cross over designs:

In parallel group design each subject gets only one treatment, in cross over design each subject gets multiple treatments. In parallel group designs if inter subject variability is large, treatment difference may get masked. In cross over design inter subject difference does not play a big role since each subject is exposed to multiple treatments and treatment comparison is within subject. Large within - subject variability dampens the advantage of cross over design. The problem of carry over effect is nonexistent in parallel design. This problem has to be managed in cross over design. In a 2X2 design, if carry over effect is significant, treatment effects can be compared only after ignoring data in period 2. In a higher order design (more periods and/or more sequences) even if carry over effect is significant the problem can be handled without discarding any data. We will not go into the details of this analysis because of complexity involved. In case of drugs with long wash out period, cross over design increases study period unreasonably. Parallel design may be preferred in such situations. In some cases it may not be possible to administer two treatments to same subjects, e.g. surgery. Cross over design cannot be used here.

Comments on limitations of the analysis presented above:

The analysis of cross over design described above is limited to the case of 2 X 2. It does not work in higher order design. A general method which works in all cross over designs is based on 'mixed model' ANOVA. Sophisticated soft wares package such as SAS have the necessary facility.

3.5.6 Exercises

E 3.5.1 Factorial design (blood pressure) Refer to the data on a factorial experiment on blood pressure with two drugs namely HCTZ and Bisoprolol. Our aim is to understand whether the two drugs interact. We have to do it by plotting cell means. You have to do it separately for systolic and diastolic bp. In each case response is to be on y axis and level of HCTZ is to be on x axis. Draw a line graph for each level of BIS. Also prepare tables showing effect of change in level of BIS at each level of HCTZ. Comment on the tables. Here is one question about the numbers involved. Find out what typical values of systolic and diastolic bp are. Check whether the data in table 3.5.1 (Yoga and BP) match with your information. If not, find out why.

E 3.5.2 Factorial design (bone density in women) Women experience noticeable changes in their bodies during menopause. One is loss of bone density. Hormone replacement therapy (HRT) is one treatment for this condition. In a 2X2 experiment reported in the Journal of American Medical Association (2003) combination of HRT with alendronate (ALN) was studied. Response was annual percentage change in bone mineral density. This was measured for several bones. The following table gives cell means for 4 bones. Prepare suitable plots of these cell means and comment on them.

Table 3.5.10: Annual percentage change in bone mineral density

Bone	Treatment			
	Placebo	HRT	ALN	HRT+ALN
Total hip	-0.81	0.48	0.97	1.43
Lumbar spine	-0.12	1.61	1.76	2.58
Femoral neck	-0.68	0.27	0.59	1.00
Ultradistal radius	-1.55	-0.53	-0.58	-0.16

3.6 Pharmacokinetics

3.6.1 What is Pharmacokinetics: Now we will introduce two important terms in drug study. One is pharmaco-kinetics (PK) and the other is pharmaco-dynamics (PD). Both terms come from the Greek word pharmacon, which means a drug. PK is the study (including mathematical modeling) of the time course of Absorption, Distribution, Metabolism and Excretion (ADME) of drugs in the body. A popular way of paraphrasing it is to say that PK is what body does to the drug (and in contrast, PD is what drug does to the body). If this sounds mysterious, we hope that at the end of this unit you will recognize that it is not.

In this section we will take a few baby steps to learn about PK and PD. An important issue of interest is the path the drug takes after administration. Pharmacokinetics is the study of this path; how drug moves around the body and how quickly this movement occurs etc. Fate of the drug depends on how it is taken. If a drug is taken orally (swallowed), it goes into stomach, some of it gets broken down, some is excreted through feces, urine etc, and one part goes to the liver. From liver some part is discarded through bile and finally the remaining portion goes into bloodstream. If drug is given as an intravenous injection, it goes straight into blood. Of course there are other ways of taking a drug (e.g. inhalation or through skin etc).

A focal point of PK study is the concentration of the drug in blood. This is the variable that is relatively easy to measure. A sample of blood is collected from the subject at different time points and drug concentration is estimated. There is an implicit assumption that effect of the drug is related closely to its concentration in blood. Value of this variable is zero before the drug is given. Then as processing of the drug within the body progresses, value of concentration keeps rising. Finally, entire drug is processed and concentration reaches a peak. After that there is a continuous decline due to elimination from blood stream. (This description is over simplified. In fact absorption and elimination occur together in the early phase. Later, there is nothing left to absorb but elimination continues.) This progression is depicted in a time-concentration graph. In this graph, time (in minutes or hours since administration of drug) is on X axis and concentration of the drug in blood at that time point is on Y-axis. As the argument here suggests, this curve is expected to rise first and then decline (unless the drug is given intravenously, in which case concentration reaches a peak right after the injection).

As an illustration of how PK properties of a drug are viewed, consider the following description about pharmacokinetics of 'Lipitor' a very popular drug for control of cholesterol. Lipitor lowers cholesterol by inhibiting its synthesis in the body. (Incidentally, Lipitor is a brand name or commercial name. The active compound in it is called 'Atorvastatin'. You should check out the website for Lipitor.)

Atorvastatin is rapidly absorbed after oral administration; maximum plasma concentrations occur within 1 to 2 hours. Extent of absorption increases in proportion to atorvastatin dose. The absolute bioavailability of atorvastatin is approximately 14%, however, Food reduces the rate and extent of atorvastatin absorption. Administration of atorvastatin with food produces a 25% reduction in Cmax and a 9% reduction in AUC (extent of absorption). However, food does not affect the plasma LDL-C lowering efficacy of atorvastatin. Evening atorvastatin dose administration is known to reduce the Cmax and AUC (extent of absorption) by 30% each. However, time of administration does not affect the plasma LDL-C lowering efficacy of atorvastatin.

It is primarily eliminated via hepatic bile excretion with less than 2% of atorvastatin recovered in the urine. Atorvastatin has an approximate elimination half-life of 14 hours. Plasma drug concentrations are significantly affected by concurrent liver disease. Patients with A stage liver disease show a 4-fold increase in both Cmax and AUC. Patients with B stage liver disease show an 16-fold increase in Cmax and an 11-fold increase in AUC.

Geriatric patients (>65 years old) show altered pharmacokinetics of atorvastatin compared to young adults. The mean AUC and Cmax values are higher (40% and 30%, respectively) for geriatric patients. Additionally, healthy elderly patients show a greater pharmacodynamic response to atorvastatin at any dose therefore, this population may have lower effective doses.

3.6.2 PK parameters: There are several points (related to the time concentration curve) mentioned in the above description. There is interest in $\mathbf{C_{max}}$ the maximum concentration reached. This is because in many cases a drug can do very little for the body unless its concentration reaches a certain threshold level. So if C_{max} is not high enough, the drug is

ineffective. Furthermore, assuming that C_{max} exceeds the threshold level required, greater the gap between C_{max} and the threshold level required, greater the benefit. Another aspect of interest is the time taken to reach C_{max}. This time is denoted by $\mathbf{T_{max}}$. Faster the approach of concentration to the maximum level, faster is the relief to the patient. This is important in acute and emergency situations. Yet another feature is plasma half-life (alternatively called elimination half life), often denoted by $t_{1/2}$. It is the time taken to lower the concentration from C_{max} to half of that level. If this parameter is large it means that the drug lingers longer in the system. It may have implications in terms of how frequent the dose should be or whether another drug can be given soon after a dose of this drug. So, in view of these aspects, it should be clear that study of PK is rather crucial.

Following graph shows a typical time concentration curve. This curve is very important in drug study and several parameters mentioned are measured using this time concentration curve. If the concentration level is below some critical threshold, the drug may not have any effect. So, a second dose may have to be scheduled before the time concentration curve falls below the threshold. As shown in the figure, T_{max} is the time interval needed for the drug to reach its highest concentration in blood. Smaller T_{max} implies faster absorption. AUC is nothing but area under the curve. It measures the total exposure of the body to the drug. Intuitively speaking, higher value of AUC is preferred. Sometimes there may be interest in comparing various formulations of the drug (tablet, syrup, intravenous injection, intramuscular injection etc) and then AUC values can be compared. Since AUC can be calculated from data on concentration, it involves no models. This is sometimes called a non-compartmental approach in contrast with methods that use the so called compartmental models. It is common to declare the range of x values over which the AUC is calculated by use of suitable suffixes such as $AUC_{(0,t)}$.

Figure 3.6.1: Pharmacokinetic curve showing parameters of interests

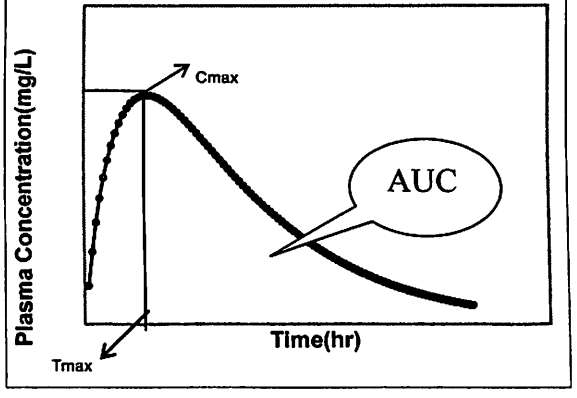
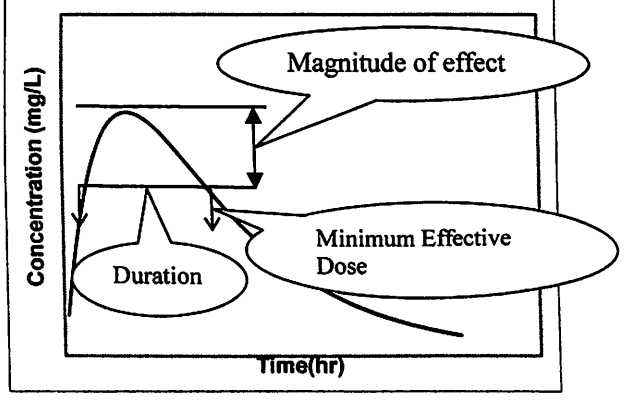


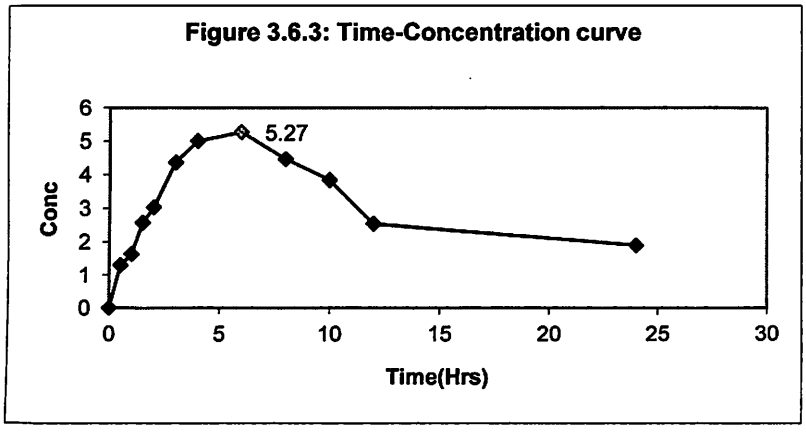
Figure 3.6.2: Pharmacokinetic curve



3.6.3 Measurement of AUC. The most common method of estimating AUC is the so called trapezoidal rule. The trapezoidal rule works by approximating the region under the graph of the function f(x) by a trapezium and calculating its area. Here the x range is divided into suitable parts. For each part (with ends a and b) the approximate area is (b-a)*[f(b)+f(a)]/2. Such areas are then summed. We will see the estimation procedure using a numerical illustration.

Table 3.6.1: Time- Drug Concentration Profile												
'i'	1	2	3	4	5	6	7	8	9	10	11	12
Time(Ti)(hr)	0	0.5	1	1.5	2	3	4	6	8	10	12	24
Conc.(Ci) (mcg/ml)	0	1.3	1.64	2.58	3.04	4.37	5.01	5.27	4.47	3.85	2.56	1.91

This time concentration profile is shown graphically in Figure 3.6.3.



For the above data set, $C_{max} = 5.27$ mcg.hr/ml and $T_{max} = 6.0$ hr. Calculation of AUC for the x range (0, 24), using trapezoidal rule will be

$$AUC_{(0\text{-}24)} = \sum_{i=1}^{12}\left[\frac{C_{i-1}+C_i}{2}\right](t_i - t_{i-1}).$$

Verify that the answer is 73.485 (mcg.hr/ml), C_{max}.

3.6.4 Measurement of $t_{1/2}$: Now we turn to the parameter called half life denoted by $t_{1/2}$. It is the time taken for concentration to decline to half of its highest level. This can be approximately found by graphical method. Draw a smooth curve through the plotted concentration points. Locate the value C_{max} on the Y axis and then locate half of it. Now move horizontally till you meet the concentration curve. The co-ordinates of this point will be ($t_{1/2}$, $0.5*C_{max}$). Another method of estimating $t_{1/2}$ is based on an exponential model for the rate of decline in concentration as time progresses beyond T_{max}. In this exercise, origin is shifted to T_{max}. In other words, time is counted after the maximum concentration is reached. The model is $C_t = C_{max} \exp(-\lambda t)$. Here t is to be understood as time since reaching the maximum concentration. Parameter λ is unknown and we will have to estimate it. At $t = t_{1/2}$, we know the value of the concentration $C_{1/2} = 0.5* C_{max}$. Hence $t_{1/2} = \ln(0.5)/ \lambda$(estimate).

Table 3.6.2: Elimination Phase		
Time	Concentration	*ln*(concentration)
0	5.27	1.662
2	4.47	1.497
4	3.85	1.348
6	2.56	0.940
18	1.91	0.647

Sometimes notation k_e is used in place of λ. The idea is that we have an elimination constant and hence the subscript e. On log scale the model becomes linear and the k_e is obtained as regression coefficient when log C_t is regressed on time t. Now let us apply this method to the data set above. Note that there are altogether 12 data points. Maximum concentration is achieved at the 8th point. There are 5 points in the elimination phase (including the peak). Those 5 data points are to be used for fitting a linear regression on a log scale. The regression output from EXCEL is given below.

Table 3.6.3: Estimating Elimination Rate

Regression Statistics						
Multiple R	0.923677					
R Square	0.853179					
Adjusted R Square	0.804239					
Standard Error	0.184473					
Observations	5					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	0.593248	0.593248	17.43304	0.02502	
Residual	3	0.10209	0.03403			
Total	4	0.695339				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.5457	0.113716	13.59257	0.000861	1.183803	1.907596
X Variable 1	-0.05446	0.013044	-4.17529	0.02502	-0.09598	-0.01295

The fitted line is $ln(C_t) = 1.55 - 0.05\ t$. So value of $k_e = 0.05$ /hr. Please verify these calculations. It is of interest to estimate AUC $_{(0,\ t)}$ for different values of t. If the value of t is in the observed range of values, then trapezoidal rule can give us the estimate. However, the trick does not work for values of t beyond the observed range (in the numerical example above, the concentration is observed over the range of time from 0 to 24 hours).

To estimate AUC $_{(0-\infty)}$, consider the model C (t) = constant* exp (- k_e.t). If t_k denotes time of last observation and C_k is the corresponding concentration then area under the curve beyond last observation will be given by, $\int_{t_k}^{\infty} C(t)dt = C_k / k_e$. To verify this, substitute the model, carry out the integration and substitute the two limits. At infinity, the value is zero. At t_k, the value is C_k (in view of the model). In the present illustration this turns out to be 1.91/0.05 = 38.2. This is the area under the curve beyond t= 24. To that we add the area under the curve for the time range (0, 24). Hence the overall estimate is

$$AUC_{(0-\infty)} = 73.485+38.2 = 111.685.$$

Since $t_{1/2} = -\ ln(1/2)\ /k_e$ we get $t_{1/2} = 0.69315 / 0.05 = 13.86$ hr. Importance of this parameter namely elimination half-life is that in practice time needed for almost all of the drug to be eliminated from the body is roughly 5 times $t_{1/2}$. This information is useful in cross over trials for deciding the 'washout period' i.e. how long to wait before giving the other drug. With present data set, 5 times $t_{1/2}$ will be approximately 69 hours. Hence a wash out period of about 70 hours (roughly 3 days) will be regarded as adequate.

Planning a PK study:

PK studies are valuable for understanding the nature of a drug and also for comparison of drugs. Here are a few general points to be kept in mind. For comparison of two drugs, a two period two sequence cross-over design is the design of choice with wash out period of at least five times the half life of the relevant drug. If half life is too long a parallel design is acceptable. Sample size should be decided keeping in mind (a) error variance (b) significance level, usually 0.05, (c) difference to be detected (d) power (at least 0.80). Minimum number of subjects should be 16 unless ethical considerations demand otherwise. Blood sampling should continue for at least 3 elimination half lives. Sampling should be continued for a sufficient period to ensure that the area extrapolated from the time of last measured concentration to infinite time is only a small percentage (normally less than 20%) of the total AUC. The use of a truncated AUC is undesirable.

There should be at least three sampling points during the absorption phase, three to four at the projected Tmax and four points during the elimination phase. Interval between successive sampling points during the elimination phase should not be longer than half life of the study drug. The following website gives extensive material on PK modeling.
http://www.scribd.com/doc/504982/BASIC-PHARMACOKINETICS

3.6.5 Exercises

E3.6.1 Calculation of AUC: Consider the following data on time and concentration for a hypothetical drug. Estimate C_{max}, T_{max}. Also calculate AUC $_{(0\text{-}12)}$.

Table 3.6.4: Estimation of AUC (0-t) for IV Application of Drug										
Time(hr)	0.25	0.50	0.75	1.00	2.00	3.00	4.00	6.00	9.00	12.00
Conc (mg/L)	15	12.5	10.5	9	5.9	4.5	3.6	2.4	1.3	0.7

E3.6.2. A patient of high blood pressure is given an intravenous injection of 160mg. of a beta-blocker. Blood samples are taken for 8 hours and concentration values are recorded. Results are given below. Plot the data on raw scale and semi-log scale. Estimate C_{max}, T_{max}. Calculate AUC $_{(0,\,480)}$.

Table 3.6.5: Estimation of AUC (0-t) for IV Application of Drug							
Time (min)	30	60	120	150	240	360	480
Concentration (ng/ml)	700	620	400	300	150	50	25

E 3.6.3 Given below are caffeine concentration values after taking a dose. Estimate C_{max}, T_{max}. Calculate AUC $_{(0,180)}$.

Table 3.6.6: Estimation of AUC (0-t) for Caffeine						
Time in (minutes) since dose	10	30	60	90	120	180
Concentration (microgram/mL)	4	3	1	0.75	0.55	0.3

E 3.6.4. Computing half life and concentration at five times the half life: Assume that λ, the elimination rate (alternative notation is k_e) is unity. Show that half life is $ln(0.5)$. Further show that when t equals five times $t_{1/2}$ the concentration reduces approximately to 3%.

E 3.6.5. Show that the above calculations are invariant for any value of λ.

E 3.6.6 Show that after 4 half lives, elimination is 94% complete. (You may run into different ways of expressing the same idea) You should not get confused by the apparent differences. That is the purpose of this exercise.

3.7 bio-availability and bioequivalence, non-inferiority

3.7.1 Bioavailability and relative bioavailability:

Bioavailability (denoted by letter *F*), sometimes termed absolute bioavailability, is the fraction of an administered dose (of a drug) reaching blood. Intravenous administration (bolus i.e. one shot, not a drip) of medication is considered to have 100% bioavailability (F=1). It does not pass through the stomach or any other intermediary but goes directly into blood. However when a medication is administered via other routes (such as orally or by inhalation etc), its bioavailability decreases due to incomplete absorption and such other reasons. Relative bioavailability is a term used to compare different formulations of the same medication.

In order to determine absolute bioavailability of a drug, a pharmacokinetic study must be done to obtain a *plasma drug concentration vs time* plot for the drug after both intravenous (IV) and non-intravenous administration. Here it is assumed that dose is the same for both administrations. If doses are different, then there has to be a correction. The correction is just to divide AUC by dose. The absolute bioavailability is the dose-corrected area under curve (AUC) non-intravenous divided by AUC intravenous. For example, the formula for calculating *F* for a drug administered by the oral route (po) is:

$$F = [AUC_{po}/Dose_{po}]/[AUC_{IV}/Dose_{IV}].$$

This is a useful tool for comparison of formulations. Some studies have found that generic preparations are not necessarily equivalent in bioavailability to brand name versions of medications. Bioavailability is also useful when calculating dosages for non-intravenous routes of administration. If F is less than 1, we may have to find a dose for which dose corrected relative bioavailability is close to 1. For this, AUC would have to increase sharply with rising dose and toxicity would have to remain acceptably low.

Majority of drugs are taken orally. Bioavailability of drugs depends on their formulation, which determines the rate at which they dissolve in the gastrointestinal tract. Issue of bioavailability of vitamin, mineral, and herbal supplements follows the same principles, for example, calcium (bound to an organic acid such as citrate) is more easily absorbed by the gastrointestinal tract than calcium carbonate. Here is one interesting news item: New York (MedscapeWire) Nov 21, 2000 — An important follow-up study that reaffirms calcium citrate's superior bioavailability when compared with calcium carbonate also provides new evidence of calcium citrate's role in protecting against bone loss.

The study, published in the November issue of the *Journal of Clinical Pharmacology,* used 3 measures to determine calcium bioavailability — serum calcium, urinary calcium, and serum parathyroid hormone (PTH). This randomized crossover study compared the single-dose bioavailability and effects on PTH of commercial calcium citrate 250 mg (Citracal,

Mission Pharma.) and calcium carbonate 500 mg (Os-Cal, Smith-Kline Beecham) supplements in postmenopausal women.

Calcium is an important nutrient for maintenance of bone strength. Women tend to experience weakened bones due to loss of calcium. Hence it is important to ensure replacement of lost calcium. In this context, different forms of calcium supplement are compared. It turns out that some forms are better than others. In particular, calcium citrate is superior to calcium carbonate. This is because of better bioavailability.

3.7.2 Bioequivalence: For a drug, one feature of major interest in phase I study is its equivalence with another drug. Bioequivalence can be of interest for several reasons. The drug under study may need to be compared with a competitor's formulation. If drug under study is a generic one then it may be compared with innovator's drug. Or it could be just two forms of the same drug like tablet versus suspension or injectable versus oral etc. Conventionally the two drugs are labeled as a test drug (denoted by T) and a reference drug (denoted by R). The equivalence can be either of efficacy variable, or of safety variable or of benefit/risk ratio.

In the strict sense of the term, bioequivalence implies equivalence of bioavailability of two formulations. This is assessed through plasma concentration-time profiles of two formulations in terms of AUC, C_{max} etc., discussed earlier. Then there is a fundamental assumption of bioequivalence, which states that 'when two drug products are equivalent in the rate and extent to which the active drug ingredient is absorbed and becomes available at the site of drug action, it is assumed that they will produce equivalent therapeutic effect'.

This assumption is considered reasonable in the following sense. The bioavailability is assessed through the concentration of drug in the blood, two formulations containing equivalent amounts of the same drug are expected to produce similar plasma concentration-time profiles (bioequivalence) and in turn equivalent clinical responses (therapeutic equivalence).

From a statistical angle, the Bioequivalence hypothesis in terms of bioavailability may be a point hypothesis i.e. 'mean bioavailability of two formulations is equal' against the alternative that it is not equal (H_0: $AUC_R = AUC_T$, H_1: $AUC_R \neq AUC_T$). If this null hypothesis is rejected it means there is statistically significant difference between mean bioavailability of two drugs. However, this difference may not be of clinical importance e.g. a difference in CFB of systolic blood pressure due to two drugs is 5 mmHg. This may turn out to be statistically significant if either sample size is too large or groups are homogeneous so that standard error is small. This difference of 5 mmHg may not be clinically interesting. So do we say that the two drugs are equivalent or that they are different? This anomalous situation can be avoided by defining bioequivalence in terms of range of difference in AUC (which may correspond to similar therapeutic effect). If the difference between bioavailability of two products is within these limits, the two products are considered to be therapeutically equivalent. These limits generally depend on nature of drug, patient population, clinical end points etc. They have to be specified by clinicians and have to be taken by statisticians as given.

With this fundamental assumption, two formulations are claimed to be bioequivalent if they provide same therapeutic effect.

Population Bioequivalence: Here we refer to the concept of population pharmacokinetics. From statistical viewpoint, two drugs are bioequivalent if marginal distributions of the pharmacokinetic parameters of interest for the two formulations are similar. This concept is referred to as population bioequivalence. Under the assumption of normality of pharmacokinetic parameters of interest, this equivalence can be assessed through first two moments i.e. by comparison of (i) mean bioavailability and (ii) variance of bioavailability. Currently FDA regulations require establishing bioequivalence only in mean bioavailability. Issue of bioavailability/bioequivalence of generic drugs is very important. As we know, checking safety and efficacy of a drug is very time consuming and expensive. The company that goes through it and proves efficacy is rewarded with exclusive rights for marketing the product for a limited period. Once this period is over, other companies can bring in their own versions of the same drug. The share of such generic drugs doubled in US in 10 years after 'Waxman-Hatch Act' (1984), which eliminated the need for separate research to prove efficacy and safety of generic drugs. Now it is enough to demonstrate its bioequivalence to the branded (innovator's) product whereas earlier a much more laborious proof of safety and efficacy had to be given.

Cost of generic drugs is a fraction of the cost of the original (innovator's) drug. As an example in 2007, Government of Thailand decided to allow import from India (Dr Reddy's Laboratory) of a generic version of 'Plavix' a blood thinning drug for heart attacks, originally developed by Sanofi-Aventis. Thai government explained that the original drug costs about US $2 per dose while the generic version from India costs only 3 cents (almost a 99% discount).

3.7.3 Statistical Criteria for Assessment of bioequivalence:

Conventionally some rules are in use for establishing bioequivalence of two formulations.

The 75/75 Rule: This rule considers ratio of individual bioavailability of test formulation with that of reference formulation. The rule says that 'the two formulations can be considered as bioequivalent if at least 75% of the individual subject ratios are within the range 0.75 to 1.25. The main advantage of this rule is that it is easy to apply. Also it considers relative bioavailability within each subject. Thus the effect of heterogeneity due to inter subject variability is taken care of. Each subject acts as his own control.

Suppose there are n subjects and bioavailability of test drug is measured and found to be $BA_T(i)$ for subject i. Bioavailability of reference drug is $BA_R(i)$, and i= 1,2,...n. For each subject we can compute $[BA_T(i) / BA_R(i)]$. This is called subject ratio. It is our random variable of interest. Let us denote it by X_i. If this falls within the interval (0.75, 1.25) then we have a success. If there are r successes in the n trials, proportion of successes is r/n. If it is at least 0.75 then this rule declares bioequivalence. What is the probability of success in case of an individual if the null hypothesis is true? It depends on distribution of X_i. We do not know the distribution of this ratio. Generally ratio of two variables is a difficult statistic to handle.

It is reasonable to assume that numerator as well as denominator of X_i is log-normally distributed because AUC is a positively skewed curve and experience shows that logarithm of AUC follows a normal distribution approximately. So, $ln(X_i)$ equals $ln\,[BA_T(i)]$- $ln\,[BA_R(i)]$. This is a difference between two normal variables and will also have a normal distribution. Under the null hypothesis, it has a zero mean because the two AUC curves are expected to be similar. For simplicity we can assume that the variance is unity. Now you can find the distribution of r the number of cases in which the subject ratio falls in the interval (0.75, 1.25).

It is generally assumed that logarithm of the subject ratio follows a normal distribution. Assuming bioequivalence of two drugs, the log (ratio) has a zero mean. However, the rule is sensitive for drugs that show large inter/intra subject variability. This may result in rejecting the equivalence too often. Hence FDA does not consider it very favorably.

The 80/20 Rule: This is based on average bioavailability of two groups. To conclude bioequivalence there are two requirements. (i) Average bioavailability of Test formulation is statistically not different than that of reference formulation and (ii) The statistical test procedure has at least 80% power for detection of a difference between two formulations, when the difference is 20% of the average bioavailability for reference formulation. Thus if we denote by μ_R and μ_T the average bioavailability of reference and test formulations respectively, then the rule requires that (i) the hypothesis H_0: $\mu_R = \mu_T$ be accepted and (ii) the probability P (rejecting H_0 when $|\mu_T - \mu_R| > 0.2\,\mu_R) > 0.8$. The second part of requirement is often used to decide the sample size.

The ± 20% Rule: By this rule, bioequivalence is accepted if bioavailability of the test formulation is within ± 20% of that of the reference formulation with a certain assurance. This is the most commonly used rule.

Can we apply this rule to compare two new drugs with a reference drug? Suppose drug A is the reference. Test drug B shows efficacy, which is 120 % of reference drug efficacy. Test drug C shows efficacy, which is 80 % of reference drug efficacy. So by this rule we will conclude A ≡ B, and A ≡C. Can we then also infer that B and C may be substitutes of each other? In fact the efficacy of drug B is 50% more as compared to efficacy of drug C. In general one drug showing efficacy within limits but on lower side of reference drug and another showing efficacy within limits but on higher side of reference drug may not be equivalent to each other. To avoid such a situation, it is suggested that bioequivalence needs to be established between two drugs only through direct comparison.

The 90% Confidence Interval rule: Recently, the criteria for approval of generic drugs are redefined. Now it requires application of confidence interval for mean (say of AUC) based on two one-sided tests. Currently, this rule is considered to be the best available method for evaluation of bioequivalence. (See: Guidance for Industry, Statistical Approaches to Establishing Bioequivalence, U.S. Department of Health and Human Services
Food and Drug Administration, Center for Drug Evaluation and Research (CDER), January 2001).

Assessing bioequivalence: Let us think about the problem from first principles. We want to test the idea that the new drug has bioavailability comparable to the innovator's drug. We wish to use test of hypothesis for this purpose. So the first question is 'what should be the null hypothesis?' It is tempting to say that the null hypothesis should be that 'there is NO DIFFERENCE'. This is how most null hypotheses in statistics are. However, you must remember that the idea of null hypothesis is that the responsibility of proof is on the party making the claim. The claim is admitted only if there is strong evidence in favor of the claim. Default action is denying the claim. Here, the producer of the generic drug wants to make a claim that it is bioequivalent to the innovator's drug. Hence default action is rejecting bioequivalence. So, null hypothesis has to be 'Test drug is NOT bioequivalent to Reference drug'.

This is a very important point and deserves careful attention. So, how do we set up the null hypothesis? The bioequivalence between average bioavailability of two formulations can be demonstrated by two approaches (i) using tests for interval hypothesis (ii) using a confidence interval approach.

3.7.4 Interval Hypothesis testing: In previous section we considered the point hypothesis H_0: $\mu_R = \mu_T$. In practice no two formulations will show exactly identical bioavailability and it may not be necessary either. Ultimate interest is in therapeutic equivalence. Two drugs can show marginally different bioavailability but still have equivalent therapeutic effect. Hence instead of a point hypothesis an interval hypothesis approach may be more meaningful. This was the argument of Schuirmann (1981) and now this is accepted widely. In this approach, if the two formulations differ in their bioavailability by less than a clinically meaningful limit then they may be considered bioequivalent. These limits are also called tolerance limits. If μ_T is the mean bioavailability for test drug and μ_R is the value for the reference drug, then the two are regarded as bioequivalent if $\mu_T - \mu_R$ is smaller than θ in absolute value. So this has to be the alternative. This is an interval alternative: $-\theta < \mu_T - \mu_R < \theta$. In many cases, it is not possible to lay down an absolute number θ and it may be more realistic to specify a relative value, say by specifying that the test drug has a mean that differs from the reference drug by no more than 20% of the mean for reference drug. In other words, the interval alternative is:

$$-0.2(\mu_R) < \mu_T - \mu_R < 0.2(\mu_R).$$

According to ± 20% rule, tolerance limits for difference between average bioavailability of two formulations will be (θ_L, θ_U) where, $\theta_L = -0.2\mu_R$ and $\theta_U = 0.2\ \mu_R$ and those for the ratio μ_T/μ_R will be (δ_L, δ_U) where $\delta_L = 0.80$ (or sometimes expressed as 80%) and $\delta_U = 1.25$ (or equivalently 125%). One technical problem is that μ_R is unknown and hence the limits cannot be calculated. This is resolved pragmatically by replacing the unknown quantity μ_R in the limit by corresponding sample mean.

Here the null hypothesis is not a single statement but comprises of two statements of one-sided hypotheses. These are H_{01}: $\mu_T - \mu_R \leq \theta_L$ and H_{02}: $\mu_T - \mu_R \geq \theta_U$ and the alternative is H_1: $\theta_L < \mu_T - \mu_R < \theta_U$. We can also split the alternative hypothesis as H_{11}: $\theta_L < \mu_T - \mu_R$ and H_{12}: $\mu_T - \mu_R < \theta_U$. The idea is to show bioequivalence by rejecting the null hypothesis of bio-

nonequivalence. Some times original efficacy variable (e.g. AUC, C_{max}) does not follow a normal distribution. Instead a lognormal distribution is more suitable. Then the hypothesis regarding difference of log means of two formulations will be equivalent to hypothesis about ratio of two mean bio-availabilities. Thus, H_{01}: $\mu_T/\mu_R \leq \delta_L$ and H_{02}: $\mu_T / \mu_R \geq \delta_U$ and the alternative is H_1: $\delta_L < \mu_T/\mu_R < \delta_U$ where, $\delta_L = \exp(\theta_L)$ and $\delta_U = \exp(\theta_U)$.

How do we carry out the test(s)? In fact it is done in two steps. The first step is to verify that bioavailability of the test product is not too low as compared to that of reference product. Rejection of this will imply that efficacy is adequate. The second part of null hypothesis is to check that the bioavailability is not too high. Any intervention is not completely free of toxic effects. Higher efficacy may be associated with more toxicity. Hence in case of higher bioavailability safety issue may be of concern. Rejection of this part will imply safety of test product is at least as much as the reference. Hence if both null hypotheses are rejected, implication is that the test product is bioequivalent to reference product. Depending on the objective of study a non-inferiority hypothesis will be same as that of bioequivalence.

Schuirmann's test: Under the normality assumption, the two sets of hypothesis can be tested with usual one-sided t-test. Thus $T_L = \dfrac{(\bar{Y}_T - \bar{Y}_R) - \theta_L}{\hat{\sigma}_F \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ and $T_U = \dfrac{(\bar{Y}_T - \bar{Y}_R) - \theta_U}{\hat{\sigma}_F \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$. The test procedure is to reject H_{01} if $T_L > t(\alpha, n_1 + n_2 - 2)$ and reject H_{02} if $T_U < -t(\alpha, n_1 + n_2 - 2)$. If both are rejected, conclude bioequivalence.

When the assumption of normality of response is doubtful, then we can try either for transformation like log or square root and check for the normality of transformed data. Or we can use non-parametric tests.

3.7.5 Exercises

E 3.7.1. Bioequivalence using the 75:75 rule: It is assumed that under the null hypothesis of equivalence, $\ln(AUC_T / AUC_R)$ follows a normal distribution with zero mean. Now assume for simplicity that the variance is 1. Obtain the probability that we commit type I error.

3.8 Determining sample size

Here we consider an issue that is of utmost importance to the organization that sponsors a clinical trial. If the sample size is not adequate, statistical tests used may lack power and we may fail to confirm efficacy. Regulators also may object if the sample size is too small. On the other hand, cost of the study shoots up as sample size is increased. So, decision has to be taken with due care. Lastly, statisticians play a central role in deciding sample size. We shall work on a few problems of this type.

3.8.1 One sample (known variance) : This is the simplest possible case for deciding sample size. It lacks realism. However, it is a good first step. Suppose X is the response variable,

known to have a normal distribution with unknown mean μ and variance 1. We have a random sample of size n. The null hypothesis can be written as H_0: $\mu = 0$, which is to be tested against the alternative, H_1: $\mu > 0$. We need to know the sample size to ensure a power of 0.80 when level of significance is 0.05 and the alternative is $\mu = \delta$ specified.

Since the alternative is right sided, we will reject the null hypothesis if the sample mean exceeds a suitably chosen value C. So, first we have to find C. We want that under null hypothesis P(sample mean > C) should equal 0.05. In other words
P($\sqrt{n}$(sample mean) > $\sqrt{n}$ C) = 0.05. Note that ($\sqrt{n}$ (sample mean)) follows a standard normal distribution when the null hypothesis is true (can you show this?). Hence $\sqrt{n}$ C = 1.64. Now we turn to power at the specified alternative. It is the same as P($\sqrt{n}$(sample mean) > 1.64) when the specified alternative is true. Let us take δ= .5. Now ($\sqrt{n}$(sample mean -0.5)) has a standard normal distribution. So we want to find n such that the following equation is satisfied: P(Z>1.64-($\sqrt{n}$)/2) = 0.80.

Referring to the table of normal probability distribution, we get 1.64-($\sqrt{n}$)/2 = -0.8416. So, the sample size is 25. Please verify this.

3.8.2 Two independent samples (known variance): Let X be the primary response variable, assumed to have a normal distribution with mean μ and variance σ^2. Suppose we are interested in comparing mean response of two formulations, control/placebo (C) and test (T). Let μ_C and μ_T denote the population means of responses to two formulations (unknown). Such situation is typical of a parallel group design. To begin with we assume σ^2 to be known.

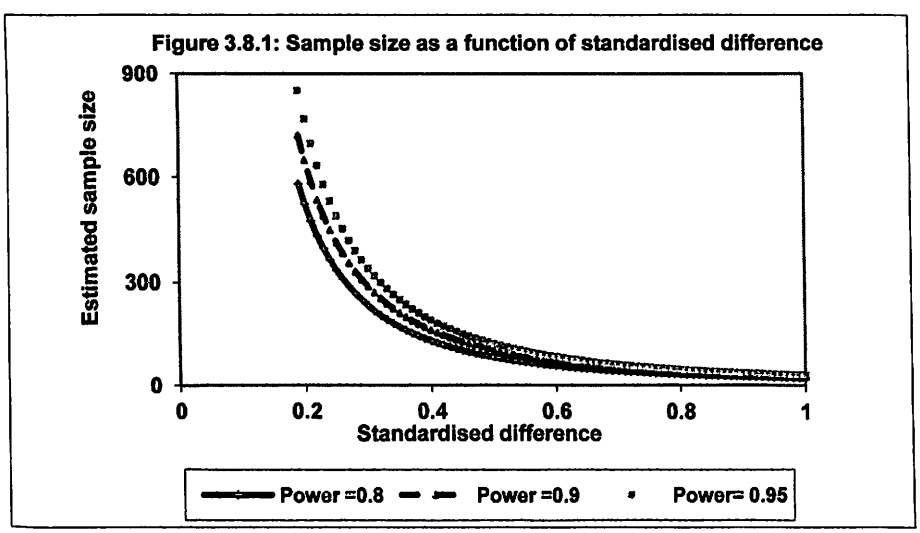
The null hypothesis can be written as H_0: $\delta = \mu_T - \mu_C = 0$, which is to be tested against the alternative, H_1: $\delta = \mu_T - \mu_C > 0$. Suppose we have observations $Y_{C1}, Y_{C2}, \ldots, Y_{CN_C}$ from Control group and $Y_{T1}, Y_{T2}, \ldots, , \ldots, Y_{TN_T}$ from two groups. Let $\bar{Y}_C$ and $\bar{Y}_T$ denote the sample means for control and test group respectively. Since we are assuming variance to be known, the test statistic will be $Z = \dfrac{(\bar{Y}_T - \bar{Y}_C)}{\sigma\sqrt{\dfrac{1}{N_C} + \dfrac{1}{N_T}}}$. This will have a standard normal distribution. The rejection region will be defined by $Z > Z_{1-\alpha}$. Suppose we wish to have a power $(1-\beta)$ to detect a true difference $\delta = \mu_T - \mu_C$ (δ specified). Then sample size calculations are as follows:

$$\text{Power} = 1-\beta = P[\text{reject } H_0 \mid \mu_T - \mu_C = \delta]$$
$$= P[Z > Z_{1-\alpha} \mid \mu_T - \mu_C = \delta]$$
$$= P\left[\frac{(\bar{Y}_T - \bar{Y}_C)}{\sigma\sqrt{\frac{1}{N_C} + \frac{1}{N_T}}} > Z_{1-\alpha} \mid \mu_T - \mu_C = \delta\right]$$

$$=P\left[\frac{(\bar{Y}_T-\bar{Y}_C)-(\mu_T-\mu_C)}{\sigma\sqrt{\frac{1}{N_C}+\frac{1}{N_T}}} > Z_{1-\alpha}-\frac{(\mu_T-\mu_C)}{\sigma\sqrt{\frac{1}{N_C}+\frac{1}{N_T}}}\right] =P\left[Z > Z_{1-\alpha}-\frac{(\mu_T-\mu_C)}{\sigma\sqrt{\frac{1}{N_C}+\frac{1}{N_T}}}\right]$$

Hence $Z_{1-\alpha}-\frac{\delta}{\sigma\sqrt{\frac{1}{N_C}+\frac{1}{N_T}}} = Z_\beta$. The recommended sample size will be same for both groups, though later it may change due to dropouts etc. So this expression simplifies to, $N=\frac{2(Z_{1-\alpha}-Z_\beta)^2\sigma^2}{\delta^2}$. This clearly shows how power, variance and difference to be detected are related to sample size. N denotes sample size from each group. Hence the total sample size will be 2N.

Following graphs show the story pictorially. Here we will consider change in sample size with respect to δ/σ. This is called standardized difference. Thus if the difference to be detected is 20% of the standard deviation, a sample of about 600 will ensure only 80% power. To increase the power to 90%, the increase in sample size is more than 100. A sample of about 850 will be needed to get a power 95%.



Suppose we keep power fixed at 80%. Then to detect a standardized difference of 40%, a sample of < 150 will suffice and a sample of just 85 will be enough if standardized difference is 50%. This standardized difference is directly related to absolute difference δ between means of two populations, but inversely related to SD.

3.8.3 Paired data: In many situations response variable is change from baseline value e.g. in a trial to test an iron supplementation, change in hemoglobin (Hb) level from baseline indicates efficacy of the test product. Other examples would be change in cholesterol level or change in systolic blood pressure. In such situations, treatment effect can be judged by a

paired t-test. Let d_i denote the difference in response between before and after treatment. Let $\bar{d}$ be the mean difference for N subjects, S_d be the corresponding sample standard deviation. A simple statistic is given by, $Z = \frac{\bar{d}}{\sigma_d \sqrt{\frac{1}{N}}}$. Hence with logic similar to the case of two independent samples (section 3.8.2), required sample size is given by, $N_d = \frac{(Z_{1-\alpha} - Z_\beta)^2 \sigma_d^2}{\delta_d^2}$, where $\delta_d > 0$ is the clinically meaningful difference to be detected. Here σ^2_d is unknown. Its estimate can be substituted to get an approximate solution.

3.8.4 Two independent samples, variance unknown: This is a more commonly occurring situation. If the variance is unknown we use a t-test instead of Z test. So the test statistics will be $t_{n^*} = \frac{(\bar{Y}_T - \bar{Y}_C)}{S\sqrt{\frac{1}{N_C} + \frac{1}{N_T}}}$ where $n^* = N_C + N_T - 2$ and S is the pooled mean square given by,

$$S^2 = \frac{\left(\sum_{i=1}^{N_C} (Y_{Ci} - \bar{Y}_C)^2 + \sum_{i=1}^{N_T} (Y_{Ti} - \bar{Y}_T)^2 \right)}{(N_C + N_T - 2)}.$$

Suppose $N_C = N_T = N$. Then n^* will be 2(N-1). Accordingly, the sample size N for each group will be given by, $N = \frac{2(t_{n^*1-\alpha} - t_{n^*\beta})^2 S^2}{\delta^2}$. The problem here is that unless we know N, we cannot get percentile points of t distribution. This is an implicit expression. Hence we have to go for an iterative procedure. We begin by assuming S^2 as true variance and hence test statistic as Z and not t. This will give us some value of N. Then we use that value of N to get degrees of freedom for t distribution and recalculate N. The procedure is iterated till convergence.

Consider the following illustration. For the sake of simplicity, suppose $S/\delta = 1$. Let $\alpha = 0.05$ and $\beta = 0.2$. Then our first approximation is $N = 2 (Z_{0.95} - Z_{0.20})^2 = 2(1.6449 + 0.8416)^2 = 12.36$. So we take N=13. Therefore $n^* = 24$. Now $t_{24, 0.95} = 1.7109$ and $t_{24, 0.20} = -0.8569$. This will lead to N in the next iteration to be 13.18. So we can take N=14 and go for second next iteration. Now $n^* = 26$. So, $t_{26, 0.95} = 1.7056$ and $t_{24, 0.20} = -0.8557$. Hence revised value of N will be 13.12 which means N=14. Hence we stop iterating further.

3.8.5 Exercises

E 3.8.1 For the one sample case with known variance, calculate the sample size assuming level of significance to be 0.01 instead of 0.05.

E 3.8.2 In the problem of two samples with known variance, it is stated that "if the difference to be detected is 20% of the standard deviation, a sample of about 600 will ensure only 80% power. To increase the power to 90%, the increase in sample size is more than 100. A sample of about 850 will be needed to get a power 95%. Suppose we keep power fixed at 80%. Then to detect a standardized difference of 40%, a sample of < 150 will suffice and a sample of just 85 will be enough if standardized difference is 50%."
Prove this assertion.

E 3.8.3. In case of paired data show that when test statistic is $Z = \dfrac{\bar{d}}{\sigma_d \sqrt{\dfrac{1}{N}}}$ required sample size is given by, $N_d = \dfrac{(Z_{1-\alpha} - Z_\beta)^2 \sigma_d^2}{\delta_d^2}$, where $\delta_d > 0$ is the clinically meaningful difference to be detected.

E 3.8.4 Bioequivalence using the 80:20 rule: Instead of individual ratios, if we take average bioavailability of the whole sample, then it is better approximated by a normal distribution. We have to set up a test with 5% level of significance. Again for simplicity, assume that individual bioavailability has unit variance. It means the sample mean bioavailability will have a variance of 1/n. Hence the difference between two sample averages will have a variance of 2/n (ignoring covariance). Now you can set up a Z test and find the rejection region. Here n is unknown. We need power to be 0.80 when the alternative is that the population mean difference in bioavailability is 0.1. Now you have to find n to satisfy this condition.

Epilogue

It is time to wrap up our discussions. Let us recapitulate what we planned to do and see how far we have been able to do it.

Broad purpose of the book was to introduce students of statistics to the world of health research in general and clinical trials in particular. Prior to that we covered some ground in terms of models in population dynamics. Modeling is a powerful tool for understanding nature and statisticians are privileged because they can handle deterministic as well as stochastic models. Simulation is a device that helps us bring out behavior of a model that is too complicated for a 'head on' analytical attack. We learnt how to summarize data on life span. These summaries are very helpful to understand health situation of a country.

We devoted considerable part of the book to epidemiology, the science of discovering how disease takes root and spreads. We went over several historical situations in order to learn how experts examined data at hand to figure out the process of epidemic. These case studies are very valuable and we recommend that students should read more material along these lines. We have learnt how epidemiological data can be analyzed and how it is possible to guess how a disease is spreading even if one is not a medical expert. The logical basis of

such thinking is the most important aspect to learn. Statisticians are often valued highly for their logical ability to see patterns in data. Hence students should revisit such discussions and also try to think through the problem on their own.

Talking about logical connections, we have learnt one important distinction namely, prospective versus retrospective studies. In retrospective studies you trace back the history of patients to locate a common cause. In prospective study subjects (not patients) are identified and followed up to see who develops signs of disease. Another equally important distinction is between surveys and experimental studies.

Surveys are observational. Investigators study people as they find them. Thus, subjects exposed to a risk factor often differ from those who are unexposed in other ways, which independently influence their risk of disease. If such confounding influences are identified in advance then allowing for them in the design and analysis of the study may be possible. There is still, however, a chance of un-recognized confounders. Example: Suppose we survey beedi smokers and check incidence of cancer among them. It may turn out that subjects are all from low income groups. So, is cancer due to beedi smoking or due to poverty? This is confounding.

Experimental studies are less susceptible to confounding because the investigator determines who is exposed and who is unexposed. In particular, if exposure is allocated randomly and the number of groups or individuals randomized is large then even unrecognized confounding effects become statistically unlikely.

There are, of course, ethical constraints on experimental research in humans, and it is not acceptable to expose subjects deliberately to potentially serious hazards. This limits the application of experimental methods in the investigation of disease etiology, although it may be possible to evaluate preventive strategies experimentally. For example, factories participating in a coronary heart disease prevention project were assigned to two groups, one receiving a program of screening for coronary risk factors and health education, and the other being left alone. Subsequent disease incidence was then compared between the two groups. The main application of experimental studies, however, is in evaluating therapeutic interventions by randomized controlled trials.

We have seen how principles of designs of experiments are applied in clinical trials. We hope these applications help in clarifying the basic statistical concepts.

***

About the authors

Dr Anil P Gore

Presently Vice-President, Bio-statistical Services, CYTEL Statistical Software and Services Pvt. Ltd. Pune, Professor and Head, Department of Statistics, University of Pune (retired), Fellow Indian Academy of Science, Elected Member International Statistical Institute, Currently Chairman, ISI-Outreach Committee for South Asia, Published over 80 research papers, Twice winner of Bose-Nandi Award of the Calcutta Statistical Association for the best paper in Applied Statistics, Taught at several universities in Malaysia, Canada and USA, Co-author of the following books -A Course on Mathematical and Statistical Ecology, Statistical Analysis of Non-normal Data, Statistics for Everyone, Medical Statistics, Author of many newspaper articles in Marathi and books titled 'Naracha Narayan' and 'Makadcheshta', Formerly Associate editor, 'Biometrics', Member, Editorial Board, 'Resonance', Member National Tiger Conservation Authority (Govt. of India), Member National Statistical Commission (Govt. of India), President Indian Society for Probability and Statistics

Dr. Mrs. Sharayu A Paranjpe

Currently working as Principal Statistician, CYTEL Statistical Software and Services Pvt. Ltd. Pune, Professor, Department of Statistics, University of Pune (retired), Published over 36 research papers, Worked as Consultant to many research groups, Winner of Bose-Nandi Award of the Calcutta Statistical Association for the best paper in Applied Statistics, Co-author of the books: A Course on Mathematical and Statistical Ecology, Statistics for Everyone, Medical Statistics, The Human Face of Clergy, Awarded- the Fellowship of the Association of Commonwealth Universities

-

Dr. Madhav B Kulkarni

Presently Head and Associate Professor in the Department of Mathematics and Statistics, B. Y. K. (Sinnar) College of Commerce, Nashik, Worked in G. E. Society's RNC Arts, JDB Com and NSC Science. College, HPT Arts and RYK Science College. Formerly, lecturer department of Statistics University of Pune, Visiting Lecturer in the Interdisciplinary School of Health Sciences University of Pune, Visiting professor at School of Biological Sciences, IISER, Thiruanantpuram, Kerala, Elected member of International Statistical Institute, Netherlands, Conducts training programs in "Advanced Statistical Computing". Co-author of books – Common Statistical Tests, Introduction to Discrete Probability and Probability Distributions, Introduction to Statistical Ecology, ASANKHYA, Statistics for Everyone, Medical Statistics . Author of "Quantitative Techniques" a book written for Yahswantrao Chavan Maharashtra Open University, Nashik. Guided one M. Phil student and has published eighteen research papers. Prepared audio-visual study material for YCM Open University, Nashik. Written a series of articles "Sankhyashastra Sarvansathi" in local newspapers, Delivered lectures on All India Radio on applied statistics. .

Dr. Anil P. Gore

Presently Vice-president, Vislation Inc Pune. Professor and Head, Department of Statistics, Pune University (Retired). Recipient of (i) "Distinguished Statistical Ecologist" award of International Association of Ecologists in Italy in 1998, (ii) first "Bose-Nandi Award" of the 'Calcutta Statistical Association' (1999) for the best paper in Applied Statistics. Taught in several universities in India, USA, Malaysia and Canada. Consultant to research institutes and commercial organizations. Author of many newspaper articles in Marathi and Marathi books - "Naracha Narayan" and "Makadcheshta". Co-author of "Statistical Analysis of Non-normal data". Published about 80 research papers in international journals of repute.

Member- National Tiger Conservation Authority, Govt. of India.,
Ex-member, National Statistics Commission, New Delhi

Dr. Mrs. Sharayu A. Paranjpe

Currently working as a Principal Statistician in Vislation Inc. Pune. Consultant to research institutes and commercial organizations. Ex-Professor, Department of Statistics, University of Pune. Awarded "Fellowship of Association of Common-Wealth Universities" (1997). Co-author of two books (i) The Human Face of Clergy (1991) and (ii) "A Course in Mathematical and Statistical Ecology". Published 36 research papers..

Dr. Madhav B. Kulkarni

Presently working as Head, and Associate Professor in the Department of Mathematics and Statistics, B. Y. K. (Sinnar) College of Commerce, Nashik, worked in G. E. Society's RNC Arts, JDB Com and NSC Science. College, HPT Arts and RYK Science College.. Formerly, lecturer department of Statistics University of Pune, Visiting Lecturer in the Interdisciplinary School of Health Sciences, University of Pune. Conducts training programs in "Advanced Statistical Computing". Co-author of three books – Common Statistical Tests, Introduction to Discrete Probability and Probability Distributions, Introduction to Statistical Ecology, . Author of "Quantitative Techniques" a book written for Yahswantrao Chavan Maharashtra Open University, Nashik. Guided one M. Phil student and has published ten research papers. Prepared audio-visual study material for YCM Open University, Nashik, Written a series of articles "Sankhyashastra Sarvansathi" in local newspapers.